IBM SPSS Modeler 18.1.1 - Noeuds source, de processus et de sortie

Remarque

Avant d'utiliser le présent document et le produit associé, prenez connaissance des informations générales figurant à la section «Remarques», à la page 417.

Informations produit

LE PRESENT DOCUMENT EST LIVRE EN L'ETAT SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DECLINE NOTAMMENT TOUTE RESPONSABILITE RELATIVE A CES INFORMATIONS EN CAS DE CONTREFACON AINSI QU'EN CAS DE DEFAUT D'APTITUDE A L'EXECUTION D'UN TRAVAIL DONNE.

Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. Les informations qui y sont fournies sont susceptibles d'être modifiées avant que les produits décrits ne deviennent eux-mêmes disponibles. En outre, il peut contenir des informations ou des références concernant certains produits, logiciels ou services non annoncés dans ce pays. Cela ne signifie cependant pas qu'ils y seront annoncés.

Pour plus de détails, pour toute demande d'ordre technique, ou pour obtenir des exemplaires de documents IBM, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial.

Vous pouvez également consulter les serveurs Internet suivants :

- http://www.fr.ibm.com (serveur IBM en France)
- http://www.ibm.com/ca/fr (serveur IBM au Canada)
- http://www.ibm.com (serveur IBM aux Etats-Unis)

Compagnie IBM France
Direction Qualité
17, avenue de l'Europe
92275 Bois-Colombes Cedex

Cette édition s'applique à la version 18.1.1 d'IBM SPSS Modeler et à toutes les éditions et modifications ultérieures, sauf indication contraire dans les nouvelles éditions.

Table des matières

Chapitre 4. Noeuds d'opérations sur les champs 127

Avis aux lecteurs canadiens

Le présent document a été traduit en France. Voici les principales différences et particularités dont vous devez tenir compte.

Illustrations

Les illustrations sont fournies à titre d'exemple. Certaines peuvent contenir des données propres à la France.

Terminologie

La terminologie des titres IBM peut différer d'un pays à l'autre. Reportez-vous au tableau ci-dessous, au besoin.

IBM France	**IBM Canada**
ingénieur commercial	représentant
agence commerciale	succursale
ingénieur technico-commercial	informaticien
inspecteur	technicien du matériel

Claviers

Les lettres sont disposées différemment : le clavier français est de type AZERTY, et le clavier français-canadien de type QWERTY.

OS/2 et Windows - Paramètres canadiens

Au Canada, on utilise :

- les pages de codes 850 (multilingue) et 863 (français-canadien),
- le code pays 002,
- le code clavier CF.

Nomenclature

Les touches présentées dans le tableau d'équivalence suivant sont libellées différemment selon qu'il s'agit du clavier de la France, du clavier du Canada ou du clavier des États-Unis. Reportez-vous à ce tableau pour faire correspondre les touches françaises figurant dans le présent document aux touches de votre clavier.

France	Canada	Etats-Unis
(Pos1)		Home
Fin	Fin	End
(PgAr)		PgUp
(PgAv)		PgDn
Inser	Inser	Ins
Suppr	Suppr	Del
Echap	Echap	Esc
Attn	Intrp	Break
Impr écran	ImpEc	PrtSc
Verr num	Num	Num Lock
Arrêt défil	Défil	Scroll Lock
(Verr maj)	FixMaj	Caps Lock
AltGr	AltCar	Alt (à droite)

Brevets

Il est possible qu'IBM détienne des brevets ou qu'elle ait déposé des demandes de brevets portant sur certains sujets abordés dans ce document. Le fait qu'IBM vous fournisse le présent document ne signifie pas qu'elle vous accorde un permis d'utilisation de ces brevets. Vous pouvez envoyer, par écrit, vos demandes de renseignements relatives aux permis d'utilisation au directeur général des relations commerciales d'IBM, 3600 Steeles Avenue East, Markham, Ontario, L3R 9Z7.

Assistance téléphonique

Si vous avez besoin d'assistance ou si vous voulez commander du matériel, des logiciels et des publications IBM, contactez IBM direct au 1 800 465-1234.

Préface

IBM® SPSS Modeler est le puissant utilitaire d'exploration de données de IBM Corp.. SPSS Modeler aide les entreprises et les organismes à améliorer leurs relations avec les clients et les citoyens grâce à une compréhension approfondie des données. A l'aide des connaissances plus précises obtenues par le biais de SPSS Modeler, les entreprises et les organismes peuvent conserver les clients rentables, identifier les opportunités de vente croisée, attirer de nouveaux clients, détecter les éventuelles fraudes, réduire les risques et améliorer les services gouvernementaux.

L'interface visuelle de SPSS Modeler met à contribution les compétences professionnelles de l'utilisateur, ce qui permet d'obtenir des modèles prédictifs plus efficaces et de trouver des solutions plus rapidement. SPSS Modeler dispose de nombreuses techniques de modélisation, telles que les algorithmes de prévision, de classification, de segmentation et de détection d'association. Une fois les modèles créés, l'utilisateur peut utiliser IBM SPSS Modeler Solution Publisher pour les remettre aux responsables, où qu'ils se trouvent dans l'entreprise, ou pour les transférer vers une base de données.

A propos d'IBM Business Analytics

Le logiciel IBM Business Analytics propose des informations complètes, cohérentes et précises auxquelles les preneurs de décisions peuvent se fier pour améliorer les performances de leur entreprise. Un portefeuille étendu de veille économique, d'analyses prédictives, de gestion des performances et de stratégie financières et d'applications analytiques vous offre des informations claires, immédiates et décisionnelles sur les performances actuelles et vous permet de prévoir les résultats futurs. Ce logiciel allie des solutions dédiées à l'industrie, des pratiques ayant fait leur preuve et des services professionnels afin que les organisations de toute taille puissent obtenir la meilleure productivité possible, automatiser leurs décisions en toute confiance et améliorer leurs résultats.

Ce portefeuille intègre le logiciel IBM SPSS Predictive Analytics qui aide les organisations à prévoir les événements à venir et à réagir en fonction des informations afin d'améliorer leurs résultats. Les clients de l'industrie du commerce, de l'éducation et des administrations du monde entier font confiance à la technologie IBM SPSS qui offre un avantage concurrentiel en attirant et fidélisant les clients et en améliorant la base de données de la clientèle tout en diminuant la fraude et en réduisant les risques. En utilisant le logiciel IBM SPSS dans leurs opérations quotidiennes, les organisations deviennent des entreprises prédictives, capables de diriger et d'automatiser les décisions pour répondre aux objectifs commerciaux et obtenir un avantage concurrentiel mesurable. Pour des informations supplémentaires ou pour joindre un représentant, consultez le site http://www.ibm.com/spss.

Assistance technique

L'assistance technique est disponible pour les clients du service de maintenance. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits IBM Corp. ou l'installation dans l'un des environnements matériels pris en charge. Pour contacter l'assistance technique, rendez-vous sur le site Web IBM Corp. à l'adresse http://www.ibm.com/support. Votre nom, celui de votre société, ainsi que votre contrat de prise en charge vous seront demandés.

Chapitre 1. A propos d'IBM SPSS Modeler

IBM SPSS Modeler est un ensemble d'outils d'exploration de données qui vous permet de développer rapidement, grâce à vos compétences professionnelles, des modèles prédictifs et de les déployer dans des applications professionnelles afin de faciliter la prise de décision. Conçu autour d'un modèle confirmé, le modèle CRISP-DM, IBM SPSS Modeler prend en charge l'intégralité du processus d'exploration de données, des données à l'obtention de meilleurs résultats commerciaux.

IBM SPSS Modeler propose différentes méthodes de modélisation issues des domaines de l'apprentissage automatique, de l'intelligence artificielle et des statistiques. Les méthodes disponibles dans la palette Modélisation vous permettent d'extraire de nouvelles informations de vos données et de développer des modèles prédictifs. Chaque méthode possède ses propres avantages et est donc plus adaptée à certains types de problème spécifiques.

Il est possible d'acquérir SPSS Modeler comme produit autonome ou de l'utiliser en tant que client en combinaison avec SPSS Modeler Server. Plusieurs autres options sont également disponibles, telles que décrites dans les sections suivantes. Pour plus d'informations, voir https://www.ibm.com/analytics/us/en/technology/spss/.

Produits IBM SPSS Modeler

La famille des produits IBM SPSS Modeler et les logiciels associés sont composés des éléments suivants.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console (inclus avec IBM SPSS Deployment Manager)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- Adaptateurs IBM SPSS Modeler Server pour IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler est une version complète du produit que vous installez et exécutez sur votre ordinateur personnel. Pour obtenir de meilleures performances lors du traitement de jeux de données volumineux, vous pouvez exécuter SPSS Modeler en mode local, comme produit autonome, ou l'utiliser en mode réparti, en association avec IBM SPSS Modeler Server.

Avec SPSS Modeler, vous pouvez créer des modèles prédictifs précis rapidement et de manière intuitive, sans aucune programmation. L'interface visuelle unique vous permet de visualiser facilement le processus d'exploration de données. Grâce aux analyses avancées intégrées au produit, vous pouvez découvrir des motifs et tendances masqués dans vos données. Vous pouvez modéliser les résultats et comprendre les facteurs qui les influencent, afin d'exploiter les opportunités commerciales et de réduire les risques.

SPSS Modeler est disponible en deux éditions : SPSS Modeler Professional et SPSS Modeler Premium. Pour plus d'informations, voir «Editions d'IBM SPSS Modeler», à la page 2.

IBM SPSS Modeler Server

Grâce à une architecture client/serveur, SPSS Modeler adresse les demandes d'opérations très consommatrices de ressources à un logiciel serveur puissant. Il offre ainsi des performances accrues sur des jeux de données plus volumineux.

SPSS Modeler Server est un produit avec licence distincte qui s'exécute en permanence en mode d'analyse réparti sur un hôte de serveur en combinaison avec une ou plusieurs installations d'IBM SPSS Modeler. Ainsi, SPSS Modeler Server fournit des performances supérieures sur de grands jeux de données car les opérations nécessitant beaucoup de mémoire peuvent être effectuées sur le serveur sans télécharger de données sur l'ordinateur client. IBM SPSS Modeler Server prend également en charge l'optimisation SQL et propose des capacités de modélisation dans la base de données pour des performances et une automatisation améliorées.

IBM SPSS Modeler Administration Console

Modeler Administration Console est une interface graphique permettant de gérer de nombreuses options de SPSS Modeler Server qui peuvent également être configurées au moyen d'un fichier d'options. La console est incluse dans IBM SPSS Deployment Manager et peut être utilisée pour surveiller et configurer vos installations SPSS Modeler Server ; elle est disponible gratuitement pour les clients actuels de SPSS Modeler Server. L'application ne peut être installée que sur des ordinateurs Windows ; en revanche, elle peut administrer un serveur installé sur n'importe quelle plate-forme prise en charge.

IBM SPSS Modeler Batch

Alors que l'exploration de données est généralement un processus interactif, il est également possible d'exécuter SPSS Modeler à partir d'une ligne de commande sans recourir à l'interface utilisateur graphique. Par exemple, vous pouvez avoir des tâches longue durée ou répétitives à exécuter sans intervention de l'utilisateur. SPSS Modeler Batch est une version spécifique du produit qui prend en charge toutes les capacités d'analyse de SPSS Modeler sans avoir besoin d'accéder à l'interface utilisateur standard. SPSS Modeler Server est requis pour utiliser SPSS Modeler Batch.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher est un outil qui permet de créer une version conditionnée d'un flux SPSS Modeler qui peut être exécutée par un moteur d'exécution externe ou intégrée dans une application externe. Ainsi, vous pouvez publier et déployer des flux SPSS Modeler complets dans des environnements où SPSS Modeler n'est pas installé. SPSS Modeler Solution Publisher est fourni avec le service IBM SPSS Collaboration and Deployment Services - Scoring et nécessite une licence distincte. Avec cette licence, vous recevez SPSS Modeler Solution Publisher Runtime qui vous permet d'exécuter les flux publiés.

Pour plus d'informations sur SPSS Modeler Solution Publisher, voir la documentation d'IBM SPSS Collaboration and Deployment Services. Le Knowledge Center d'IBM SPSS Collaboration and Deployment Services contient des sections appelées "IBM SPSS Modeler Solution Publisher" et "IBM SPSS Analytics Toolkit."

Adaptateurs IBM SPSS Modeler Server pour IBM SPSS Collaboration and Deployment Services

Différents adaptateurs pour IBM SPSS Collaboration and Deployment Services sont disponibles et permettent à SPSS Modeler et SPSS Modeler Server d'interagir avec un référentiel IBM SPSS Collaboration and Deployment Services. Ainsi, un flux SPSS Modeler déployé sur le référentiel peut être partagé par différents utilisateurs ou peut être accessible depuis l'application client léger IBM SPSS Modeler Advantage. Installez l'adaptateur sur le système qui héberge le référentiel.

Editions d'IBM SPSS Modeler

SPSS Modeler est disponible dans les éditions suivantes.

SPSS Modeler Professional

SPSS Modeler Professional offre tous les outils nécessaires à l'utilisation de la plupart des types de données structurées, tels que les comportements et interactions suivis dans les systèmes CRM, les

caractéristiques sociodémographiques, les comportements d'achat et les données de vente.

SPSS Modeler Premium

SPSS Modeler Premium est un produit avec licence distincte qui étend le champ d'applications de SPSS Modeler Professional afin de pouvoir traiter des données spécialisées et des données de texte non structurées. SPSS Modeler Premium inclut IBM SPSS Modeler Text Analytics :

IBM SPSS Modeler Text Analytics utilise des technologies linguistiques avancées et le traitement du langage naturel pour traiter rapidement une large variété de données textuelles non structurées, en extraire les concepts clés et les organiser pour les regrouper dans des catégories. Les concepts extraits et les catégories peuvent ensuite être combinés aux données structurées existantes, telles que les données démographiques, et appliqués à la modélisation grâce à la gamme complète d'outils d'exploration de données d'IBM SPSS Modeler, afin de favoriser une prise de décision précise et efficace.

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription propose les mêmes fonctionnalités d'analyse prédictive que le client IBM SPSS Modeler traditionnel. L'édition Subscription vous permet de télécharger régulièrement des mises à jour de produit.

Documentation

Une documentation est disponible dans le menu Aide de SPSS Modeler. Elle ouvre le SPSS Modeler Knowledge Center, qui est disponible au public en dehors du produit.

La documentation complète de chaque produit (y compris les instructions d'installation) au format PDF est également disponible dans un dossier compressé distinct, avec le téléchargement du produit. Les documents PDF peuvent également être téléchargés depuis le Web sur le site http://www.ibm.com/support/docview.wss?uid=swg27046871.

Documentation de SPSS Modeler Professional

La suite de documentation SPSS Modeler Professional (à l'exception des instructions d'installation) est la suivante.

- **IBM SPSS Modeler - Guide d'utilisation.** Introduction générale à SPSS Modeler : création de flux de données, traitement des valeurs manquantes, création d'expressions CLEM,utilisation des projets et des rapports, et regroupement des flux pour le déploiement dans IBM SPSS Collaboration and Deployment Services ou IBM SPSS Modeler Advantage.
- **Noeuds de source, d'exécution et de sortie d'IBM SPSS Modeler.** Descriptions de tous les noeuds utilisés pour lire, traiter et renvoyer les données de sortie dans différents formats. En pratique, cela signifie tous les noeuds autres que les noeuds de modélisation.
- **Noeuds modélisation d'IBM SPSS Modeler.** Descriptions de tous les noeuds utilisés pour créer des modèles d'exploration de données. IBM SPSS Modeler propose différentes méthodes de modélisation issues des domaines de l'apprentissage automatique, de l'intelligence artificielle et des statistiques.
- **Guide des applications IBM SPSS Modeler.** Les exemples de ce guide fournissent des introductions brèves et ciblées aux méthodes et techniques de modélisation. Une version en ligne de ce guide est également disponible dans le menu Aide. Pour plus d'informations, voir la rubrique «Exemples d'application», à la page 4.
- **Guide de génération de scripts Python et d'automatisation IBM SPSS Modeler.** Ce manuel fournit des informations sur l'automatisation du système via des scripts Python, notamment sur les propriétés pouvant être utilisées pour manipuler les noeuds et les flux.
- **Guide de déploiement d'IBM SPSS Modeler.** Informations sur l'exécution des flux IBM SPSS Modeler comme étapes des travaux d'exécution sous IBM SPSS Deployment Manager.

- **Guide du développeur IBM SPSS Modeler CLEF.** CLEF permet d'intégrer des programmes tiers tels que des programmes de traitement de données ou des algorithmes de modélisation en tant que noeuds dans IBM SPSS Modeler.
- **Guide d'exploration de base de données IBM SPSS Modeler.** Informations sur la manière de tirer parti de la puissance de votre base de données pour améliorer les performances et étendre la gamme des capacités d'analyse via des algorithmes tiers.
- **IBM SPSS Modeler Server Guide des performances et d'administration.** Informations sur le mode de configuration et d'administration d'IBM SPSS Modeler Server.
- **IBM SPSS Deployment Manager - Guide d'utilisation.** Informations sur l'utilisation de l'interface utilisateur de la console d'administration incluses dans l'application Deployment Manager pour la surveillance et la configuration d'IBM SPSS Modeler Server.
- **Guide CRISP-DM d'IBM SPSS Modeler.** Guide détaillé sur l'utilisation de la méthodologie CRISP-DM pour l'exploration de données avec SPSS Modeler
- **IBM SPSS Modeler Batch - Guide d'utilisation.** Guide complet sur l'utilisation d'IBM SPSS Modeler en mode de traitement par lots, avec des détails sur l'exécution en mode de traitement par lots et les arguments de ligne de commande. Ce guide est disponible au format PDF uniquement.

Documentation de SPSS Modeler Premium

La suite de documentation SPSS Modeler Premium (à l'exception des instructions d'installation) est la suivante.

- **Guide de l'utilisateur de SPSS Modeler Text Analytics .** Informations sur l'utilisation des analyses de texte avec SPSS Modeler, notamment sur les noeuds Text Mining, l'espace de travail interactif, les modèles et d'autres ressources.

Exemples d'application

Tandis que les outils d'exploration de données de SPSS Modeler peuvent vous aider à résoudre une grande variété de problèmes métier et organisationnels, les exemples d'application fournissent des introductions brèves et ciblées aux méthodes et aux techniques de modélisation. Les jeux de données utilisés ici sont beaucoup plus petits que les énormes entrepôts de données gérés par certains Data Miners, mais les concepts et les méthodes impliqués peuvent être adaptés à des applications réelles.

Pour accéder aux exemples, cliquez sur **Exemples d'application** dans le menu Aide de SPSS Modeler.

Les fichiers de données et les flux d'échantillons sont installés dans le dossier `Demos`, sous le répertoire d'installation du produit. Pour plus d'informations, voir «Dossier Demos».

Exemples de modélisation de base de données. Consultez les exemples dans *IBM SPSS Modeler Guide d'exploration de base de données*.

Exemples de génération de scripts. Consultez les exemples dans *IBM SPSS Modeler Guide de génération de scripts et d'automatisation*.

Dossier Demos

Les fichiers de données et les flux d'échantillons utilisés avec les exemples d'application sont installés dans le dossier `Demos`, sous le répertoire d'installation du produit (par exemple : `C:\Program Files\IBM\SPSS\Modeler\<version>\Demos`). Ce dossier est également accessible à partir du groupe de programmes IBM SPSS Modeler, dans le menu Démarrer de Windows ou en cliquant sur `Demos` dans la liste des répertoires récents de la boîte de dialogue **Fichier** > **Ouvrir un flux**.

Suivi des licences

Lorsque vous utilisez SPSS Modeler, l'utilisation des licences est suivie et consignée à intervalles réguliers. Les métriques de licence consignées sont *AUTHORIZED_USER* et *CONCURRENT_USER* et le type de métrique consigné dépend du type de licence dont vous disposez pour SPSS Modeler.

Les fichiers journaux générés peuvent être traités par IBM License Metric Tool, à partir duquel vous pouvez générer des rapports d'utilisation de licence.

Les fichiers journaux des licences sont créés dans le répertoire dans lequel les fichiers journaux de SPSS Modeler Client sont enregistrés (par défaut, `%ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log`).

Chapitre 2. Noeuds source

Sommaire

Les noeuds source vous permettent d'importer des données stockées dans différents formats : fichiers plats, IBM SPSS Statistics (`.sav`), SAS, Microsoft Excel et bases de données relationnelles compatibles ODBC, entre autres. Vous pouvez également générer des données synthétiques à l'aide du noeud Utilisateur.

La palette Sources contient les noeuds suivants :

La source Analytic Server vous permet d'exécuter un flux sur le système de fichiers HDFS (Hadoop Distributed File System). Les informations qui se trouvent dans une source de données Analytic Server peuvent provenir de divers emplacements, comme des fichiers texte et des bases de données. Pour plus d'informations, voir la rubrique «Noeud source Analytic Server», à la page 13.

Le noeud SGBD peut être utilisé pour importer des données provenant de nombreux autres logiciels utilisant la connectivité ODBC (Open Database Connectivity), tels que Microsoft SQL Server, Db2, Oracle, etc. Pour plus d'informations, voir «noeud source de base de données», à la page 18.

Le noeud Délimité lit les données de fichiers texte de longueur variable, c'est-à-dire les fichiers dont les enregistrements contiennent un nombre fixe de champs et un nombre variable de caractères. Ce noeud est également utile pour les fichiers contenant des textes d'en-tête de longueur fixe et certains types d'annotation. Pour plus d'informations, voir «Noeud Délimité», à la page 27.

Le noeud Fixe permet d'importer les données de fichiers texte de longueur fixe, c'est-à-dire les fichiers dont les champs ne sont pas délimités, mais commencent au même endroit et sont de longueur fixe. Les données générées automatiquement ou héritées sont souvent stockées au format de longueur fixe. Pour plus d'informations, voir «Noeud Fixe», à la page 31.

Le noeud Fichier Statistics lit les données du format de fichier *.sav* ou *.zsav* utilisé par IBM SPSS Statistics, ainsi que des fichiers cache enregistrés dans IBM SPSS Modeler, qui utilisent le même format.

Le noeud Data Collection importe des données d'enquête dans différents formats utilisés par le logiciel d'étude de marché et conformément au modèle de données de Data Collection. Pour pouvoir utiliser ce noeud, vous devez avoir installé auparavant Data Collection Developer Library. Pour plus d'informations, voir «Data Collection Noeud», à la page 34.

Le noeud source IBM Cognos importe des données depuis les bases de données Cognos Analytics.

Le noeud source IBM Cognos TM1 importe des données depuis les bases de données Cognos TM1.

Le noeud Fichier SAS permet d'importer des données SAS dans IBM SPSS Modeler. Pour plus d'informations, voir «Noeud source SAS», à la page 45.

Le noeud Excel permet d'importer des données de Microsoft Excel dans le format de fichier .xlsx. Aucune source de données ODBC n'est requise. Pour plus d'informations, voir «Noeud source Excel», à la page 46.

Le noeud source XML importe des données au format XML dans le flux. Vous pouvez importer un fichier ou tous les fichiers dans un répertoire. Vous pouvez aussi spécifier un fichier de schéma à partir duquel lire la structure XML.

Le noeud Utilisateur représente une façon simple de créer des données synthétiques (à partir de zéro ou en modifiant des données existantes). Ceci est utile, par exemple, si vous souhaitez créer un jeu de données de test pour la modélisation. Pour plus d'informations, voir «Noeud Utilisateur», à la page 49.

Le noeud Génération de simulation permet de générer facilement des données simulées, soit intégralement à l'aide de distributions statistiques spécifiées par l'utilisateur, soit automatiquement à l'aide des distributions obtenues via l'exécution d'un noeud Ajustement de simulation sur des données historiques existantes. Cela s'avère utile si vous voulez évaluer le résultat d'un modèle prédictif en présence d'incertitude dans les entrées du modèle.

Le noeud Vue de données peut être utilisé pour accéder aux sources de données définies dans les vues de données analytiques IBM SPSS Collaboration and Deployment Services. Une vue de données analytiques définit une interface standard d'accès aux données et lui associe plusieurs sources de données physiques. Pour plus d'informations, voir «Noeud Vue de données», à la page 69.

Utilisez le noeud source Géospatial pour ajouter des données de carte ou spatiales dans votre session d'exploration de données. Pour plus d'informations, voir la rubrique «Noeud source Géospatial», à la page 70.

Pour commencer un flux, ajoutez un noeud source à l'espace de travail de flux. Ensuite, double-cliquez sur le noeud pour ouvrir la boîte de dialogue correspondante. Dans les différents onglets de la boîte de dialogue, vous pouvez lire les données, afficher les champs et les valeurs, et définir différentes options, comme les filtres, les types de données, le rôle du champ et la détection des valeurs manquantes.

Définition du stockage et du formatage des champs

Les options de l'onglet Données des noeuds Fixe, Délimité, Source XML et Utilisateur vous permettent d'indiquer le type de stockage des champs à mesure qu'ils sont importés ou créés dans IBM SPSS Modeler. Pour les noeuds Fixe, Délimité et Utilisateur, vous pouvez aussi indiquer le formatage des champs et d'autres métadonnées.

Dans le cas de données lues à partir d'autres sources, le stockage est déterminé automatiquement ; vous pouvez toutefois le modifier par le biais d'une fonction de conversion, telle que `to_integer`, appliquée dans un noeud Remplacer ou Calculer.

Champ Utilisez la colonne **Champ** pour consulter et sélectionner des champs dans le jeu de données en cours.

Remplacer Sélectionnez la case à cocher dans la colonne **Remplacer** pour activer des options dans les colonnes **Stockage** et **Format d'entrée**.

Stockage des données

Le stockage des données décrit la façon dont les données sont stockées dans un champ. Par exemple, un champ comportant les valeurs `1` et `0` stocke des nombres entiers. Il est à différencier du niveau de mesure, qui décrit l'utilisation des données et n'a aucune incidence sur le stockage. Par exemple, vous pouvez définir le niveau de mesure de nombre entier comportant les valeurs `1` et `0` comme étant un champ *indicateur*. En général, `1` correspond à la valeur *True (vrai)* et `0` à la valeur *False (faux)*. Alors que le stockage doit être déterminé au niveau de la source, le niveau de mesure peut être modifié à l'aide d'un noeud Typer en tout point du flux. Pour plus d'informations, voir la rubrique «Niveaux de mesure», à la page 145.

Les types de stockage disponibles sont les suivants :

- **Chaîne** Utilisé pour les champs qui contiennent des données non numériques, aussi appelées données alphanumériques. Une chaîne peut inclure n'importe quelle séquence de caractères, telle que *fred*, *Classe 2* ou *1234*. Notez que les nombres utilisés dans les chaînes ne peuvent pas être inclus dans les calculs.
- **Entier** Champ dont les valeurs sont des entiers.
- **Réel** Les valeurs sont des nombres qui peuvent inclure des décimales (non limitées à des entiers). Le format d'affichage est indiqué dans la boîte de dialogue Propriétés de flux et peut être ignoré pour des champs individuels dans un noeud Typer (onglet Format).
- **Date** Valeurs de date spécifiées au format standard, comme année, moi et jour (par exemple `2007-09-26`). Le format exact est indiqué dans la boîte de dialogue Propriétés de flux.
- **Temps** Temps mesuré en tant que durée. Par exemple, un appel de service ayant duré `1` heure, `26` minutes et `38` secondes peut être représenté sous la forme `01:26:38`, en fonction du format d'heure actuel indiqué dans la boîte de dialogue Propriétés de flux.
- **Horodatage** Valeurs qui incluent un composant de date et un composant d'heure, par exemple `2007-09-26 09:04:00`, selon les formats de date et d'heure définis dans la boîte de dialogue Propriétés du flux. Remarque : il se peut que les valeurs d'horodatage doivent être placées entre guillemets doubles pour être interprétées comme une valeur unique et non comme des valeurs date et heure distinctes. (Cela s'applique, par exemple, lors de la saisie de valeurs dans un noeud Utilisateur).
- **Liste** Introduit dans SPSS Modeler version 17, avec les nouveaux niveaux de mesure Géospatial et Collection, un champ de stockage Liste contient plusieurs valeurs pour un enregistrement unique. Il existe des versions de liste pour tous les autres types de stockage.

Tableau 1. Icônes des types de stockage Liste

Icône	Type de stockage
[A]	Liste de chaînes
[ ]	Liste de nombres entiers
[ ]	Liste de nombres réels
[ ]	Liste d'heures
[ ]	Liste de dates
[ ]	Liste d'horodatages
[[]]	Liste dont la profondeur est supérieure à zéro

De plus, pour une utilisation avec le niveau de mesure Collection, il existe des versions de liste pour les niveaux de mesure ci-dessous.

Tableau 2. Icônes des niveaux de mesure de liste

Icône	Niveau de mesure
[ ]	Liste d'éléments continus
[ ]	Liste d'éléments catégoriels
[ ]	Liste d'indicateurs
[ ]	Liste d'éléments nominaux
[ ]	Liste d'éléments ordinaux

Les listes peuvent être importées dans SPSS Modeler dans l'un des trois noeuds source (Analytic Server, Géospatial ou Délimité) ou créé dans vos flux à l'aide des noeuds d'opération de champ Calculer ou Remplacer.

Pour plus d'informations sur les listes et leur interaction avec les niveaux de mesure Collection et Géospatial, voir «Stockage de liste et niveaux de mesure associés», à la page 12

Conversion de stockages. Vous pouvez également convertir le stockage d'un champ à l'aide de diverses fonctions de conversion, comme `to_string` et `to_integer` dans un noeud Remplacer. Pour plus

d'informations, voir la rubrique «Conversion du stockage à l'aide du noeud Remplacer», à la page 170. Notez que les fonctions de conversion (et toutes les autres fonctions qui nécessitent un type spécifique d'entrée, par exemple une valeur de date ou d'heure) dépendent des formats actuels indiqués dans la boîte de dialogue Propriétés de flux. Par exemple, si vous souhaitez convertir un champ de type chaîne avec des valeurs *Jan 2003*, *Fév 2003*, etc., en stockage de date, sélectionnez **MOIS AAAA** comme format de date par défaut pour le flux. Les fonctions de conversion sont également disponibles depuis le noeud Calculer pour la conversion temporaire lors d'un calcul. Vous pouvez également utiliser le noeud Calculer pour effectuer d'autres manipulations, telles que la modification du codage des champs de type chaîne contenant des valeurs catégorielles. Pour plus d'informations, voir la rubrique «Recodage des valeurs à l'aide du noeud Calculer», à la page 169.

Lecture de données mixtes. Au cours de la lecture des champs de stockage numérique (entier, nombre réel, heure, horodatage ou date), toutes les valeurs non numériques sont définies comme étant nulles ou manquantes dans le système. En effet, contrairement à certaines applications, IBM SPSS Modeler n'autorise pas les types de stockage mixtes au sein d'un champ. Pour éviter ce type de problème, faites en sorte que les champs comportant des données mixtes soient lus en tant que chaînes ; pour cela, modifiez le type de stockage dans le noeud source ou dans l'application externe.

Format d'entrée des champs (noeuds Fixe, Délimité ou Utilisateur uniquement).

Pour tous les types de stockage, à l'exception de Chaîne et Entier, vous pouvez choisir dans la liste déroulante les options de formatage du champ sélectionné. Par exemple, lorsque vous fusionnez les données de plusieurs paramètres régionaux, vous pouvez être amené à utiliser un point (.) comme séparateur décimal pour un champ, alors qu'un autre champ aura une virgule pour séparateur.

Les options d'entrée indiquées dans le noeud source remplacent les options de formatage définies dans la boîte de dialogue des propriétés de flux ; elles n'apparaissent toutefois pas ultérieurement dans le flux. Ces options visent à analyser correctement les entrées fournies en fonction de votre connaissance des données. Les formats spécifiés servent de point de repère à l'analyse de ces données lorsqu'elles sont lues dans IBM SPSS Modeler, et non à déterminer la manière dont elles doivent être formatées après leur lecture dans IBM SPSS Modeler. Pour indiquer le formatage de chaque champ au sein du flux, utilisez l'onglet Format d'un noeud Typer. Pour plus d'informations, voir «Onglet Paramètres du champ», à la page 157.

Les options varient selon le type de stockage utilisé. Par exemple, pour le type de stockage Réel, vous pouvez sélectionner le séparateur décimal **Point (.)** ou **Virgule (,)**. Pour les champs d'horodatage, une autre boîte de dialogue s'ouvre lorsque vous choisissez **Spécifier** dans la liste déroulante. Pour plus d'informations, voir «Définition des options de formatage des champs», à la page 158.

Pour tous les types de stockage, vous pouvez également sélectionner **Flux par défaut** afin d'utiliser les paramètres par défaut du flux pour l'importation. Les paramètres de flux sont répertoriés dans la boîte de dialogue des propriétés du flux.

Options supplémentaires

Vous pouvez utiliser d'autres options à l'aide de l'onglet Données :

- Pour afficher les paramètres de stockage pour les données qui ne sont plus connectées via le noeud en cours (données d'apprentissage, par exemple), sélectionnez **Afficher les paramètres de champ non utilisés**. Vous pouvez effacer les champs hérités en cliquant sur **Effacer**.
- A tout moment lorsque vous travaillez dans cette boîte de dialogue, cliquez sur **Rafraîchir** pour recharger les champs à partir de la source de données. Ceci est utile lorsque vous modifiez des connexions des données au noeud source ou lorsque vous utilisez les différents onglets de la boîte de dialogue.

Stockage de liste et niveaux de mesure associés

Introduit dans SPSS Modeler version 17 pour l'utilisation des nouveaux niveaux de mesure Géospatial et Collection, un champ de stockage Liste contient plusieurs valeurs pour un enregistrement unique. Les listes sont placées entre crochets ([]), par exemple `[1,2,4,16]` et `["abc", "def"]`.

Les listes peuvent être importées dans SPSS Modeler dans l'un des trois noeuds source (Analytic Server, Géospatial ou Délimité) ou créées dans vos flux à l'aide des noeuds d'opération sur les champs Calculer ou Remplacer.

Les listes sont associées à une profondeur ; par exemple, une liste simple comportant des éléments placés entre crochets au format `[1,3]` est enregistrée dans IBM SPSS Modeler avec la profondeur zéro. En plus des listes simples dont la profondeur est zéro, vous pouvez utiliser des listes imbriquées, dans lesquelles chaque valeur de la liste est une liste elle-même.

La profondeur d'une liste imbriquée dépend du niveau de mesure associé. Pour Sans type, la limite de profondeur n'est pas définie ; pour Collection, la profondeur est de zéro ; pour Géospatial, la profondeur doit être comprise entre zéro et deux inclus, selon le nombre d'éléments imbriqués.

Pour les listes de profondeur zéro, vous pouvez définir le niveau de mesure Géospatial ou Collection. Ces deux niveaux sont des niveaux de mesure parent et vous définissez les informations de sous-niveau de mesure dans la boîte de dialogue Valeurs. Le sous-niveau de mesure d'une collection détermine le niveau de mesure des éléments dans cette liste. Tous les niveaux de mesure (sauf Sans type et Géospatial) sont disponibles en tant que sous-niveaux pour les collections. Le niveau de mesure Géospatial possède six sous-niveaux : Point, Chaîne, Polygone, Multipoint, Multichaîne et Multipolygone ; pour plus d'informations, voir «Sous-niveaux de mesure géospatiaux», à la page 147.

Remarque : Le niveau de mesure Collection ne peut être utilisé qu'avec des listes dont la profondeur est zéro ; le niveau de mesure Géospatial ne peut être utilisé qu'avec des listes dont la profondeur maximale est deux ; le niveau de mesure Sans type peut être utilisé avec toutes les profondeurs de liste.

L'exemple suivant illustre la différence entre une liste de profondeur zéro et une liste imbriquée en utilisant la structure des sous-niveaux de mesure Géospatial Point et Chaîne :

- Le sous-niveau de mesure Géospatial Point possède une profondeur de champ de zéro :
  `[1,3]` deux coordonnées
  `[1,3,-1]` trois coordonnées
- Le sous-niveau de mesure Géospatial Chaîne possède une profondeur de champ de un :
  `[ [1,3], [5,0] ]` deux coordonnées
  `[ [1,3,-1], [5,0,8] ]` trois coordonnées

Le champ Point (avec une profondeur de zéro) est une liste normale dans laquelle chaque valeur est composée de deux ou trois coordonnées. Le champ Chaîne (avec une profondeur de un) est une liste de points, dans laquelle chaque point est composé d'une série supplémentaire de valeurs de liste.

Pour plus d'informations sur la création de liste, voir «Calcul d'une liste ou d'un champ géospatial», à la page 166.

Caractères de contrôle non pris en charge

Certains des processus dans SPSS Modeler ne peuvent pas gérer les données incluant plusieurs caractères de contrôle. Si vos données utilisent ces caractères, vous voyez s'afficher un message d'erreur du type suivant :

```
Caractères de contrôle non pris en charge trouvés dans les valeurs de champ{0}
```

Les caractères non pris en charge sont les suivants : de `0x0` à `0x3F` compris, et `0x7F` ; toutefois, la tabulation `0x9(\t)`), le saut de ligne (`0xA(\n)`), et le retour chariot (`0xD(\r)`) ne posent pas problème.

Si vous voyez s'afficher un message d'erreur relatif à des caractères non pris en charge, dans votre flux, après votre noeud Source, utilisez un noeud Remplacer et l'expression CLEM **`stripctrlchars`** pour remplacer les caractères.

Noeud source Analytic Server

La source Analytic Server vous permet d'exécuter un flux sur un système de fichier HDFS (Hadoop Distributed File System). Les informations figurant dans une source de données Analytic Server peuvent provenir de divers emplacements, notamment :

- Fichiers texte sur HDFS
- Bases de données
- HCatalog

Un flux avec une source Analytic Server sera généralement exécuté sur HDFS ; toutefois, si un flux contient un noeud qui n'est pas pris en charge pour une exécution sur HDFS, le flux sera "répercuté" autant que possible dans Analytic Server, puis SPSS Modeler Server tentera de traiter le reste du flux. Vous devrez sous-échantillonner les jeux de données très volumineux ; par exemple, en plaçant un noeud échantillon dans le flux.

Si vous souhaitez utiliser votre propre connexion Analytic Server à la place de la connexion par défaut définie par votre administrateur, désélectionnez **Utiliser le serveur Analytic Server par défaut** et sélectionnez votre connexion. Pour des informations sur la configuration de plusieurs connexions Analytic Server, voir Connexion à Analytic Server.

Source de données. En supposant que votre administrateur SPSS Modeler Server ou vous-même ayez établi une connexion, vous sélectionnez une source de données contenant les données à utiliser. Une source de données contient les fichiers et métadonnées associés à cette source. Cliquez sur **Sélectionner** pour afficher une liste des sources de données disponibles. Pour plus d'informations, reportez-vous à la rubrique «Sélection d'une source de données».

Si vous avez besoin de créer une nouvelle source de données ou d'éditer une source de données existante, cliquez sur **Launch Data Source Editor...**.

Notez que l'utilisation de plusieurs connexions Analytic Server peut s'avérer utile pour contrôler le flux de données. Par exemple, si vous utilisez les noeuds Source et Exportation d'Analytic Server, vous pouvez utiliser des connexions Analytic Server dans différentes branches d'un flux afin que lorsque chaque branche est exécutée, elle utilise son propre serveur Analytic Server, sans extraction de données dans IBM SPSS Modeler Server. Notez que si une branche contient plusieurs connexions Analytic Server, les données seront extraites des serveurs Analytic Server vers IBM SPSS Modeler Server. Pour plus d'informations, notamment sur les restrictions, voir Propriétés du flux Analytic Server.

Sélection d'une source de données

La table Sources de données affiche la liste des sources de données disponibles. Sélectionnez la source que vous souhaitez utiliser et cliquez sur **OK**.

Cliquez sur **Show Owner** pour afficher le propriétaire de la source de données.

Filter by vous permet de filtrer la liste des sources de données à partir d'un **mot-clé** et compare les critères de filtrage avec le nom, la description ou le **propriétaire** de la source de données. Vous pouvez entrer une combinaison de caractères de type chaîne, de type numérique ou génériques décrits ci-dessous comme critères de filtrage. La chaîne de recherche est sensible à la casse. Cliquez sur **Actualiser** pour actualiser la table Sources de données.

_ Un trait de soulignement peut être utilisé pour représenter un caractère unique dans la chaîne de recherche.

% Le symbole du pourcentage peut être utilisé pour représenter une séquence de 0 caractère ou d'un ou de plusieurs caractères dans la chaîne de recherche.

Modification des données d'identification

Si vos données d'identification pour l'accès à Analytic Server sont différentes de celles permettant d'accéder à SPSS Modeler Server, vous devrez entrer les données d'identification Analytic Server lors de l'exécution d'un flux sur Analytic Server. Si vous ne connaissez pas vos données d'identification, contactez votre administrateur de serveur.

Noeuds pris en charge

Un grand nombre de noeuds SPSS Modeler peuvent être exécutés sur HDFS, mais il peut exister des différences dans l'exécution de certains noeuds et quelques-uns ne sont pas pris en charge actuellement. La présente rubrique détaille le niveau de prise en charge actuel.

Général

- Certains caractères qui sont normalement admis dans un nom de champ Modeler entre guillemets ne sont pas acceptés par Analytic Server.
- Pour qu'un flux Modeler soit exécuté dans Analytic Server, il doit commencer par un ou plusieurs noeuds source Analytic Server et se terminer par un noeud de modélisation ou un noeud d'exportation Analytic Server unique.
- Il est recommandé de définir le stockage de cibles continues comme réel plutôt que comme entier. Les modèles d'évaluation écrivent toujours des valeurs réelles dans les fichiers de données de sortie pour les cibles continues, alors que le modèle de données de sortie pour les scores suit le stockage de la cible. Par conséquent, si une cible continue est dotée d'un stockage d'entier, cela provoquera une non-concordance entre les valeurs écrites et le modèle de données pour les scores, qui générera des erreurs lorsque vous tenterez de lire les données évaluées.
- Si une mesure de champ est géospatiale, la fonction de @OFFSET n'est pas prise en charge.

Source

- Un flux ne commençant pas par un noeud source Analytic Server sera exécuté localement.

Opérations sur les lignes

Toutes les opérations d'enregistrement sont prises en charge, sauf les noeuds Streaming TS et Boîtes espace-temps. Vous trouverez ci-après des remarques supplémentaires sur la fonctionnalité des noeuds prise en charge.

Sélectionner

- Prend en charge le même ensemble de fonctions que le noeud dériver.

Echantillon

- L'échantillonnage au niveau des blocs n'est pas pris en charge.
- Les méthodes d'échantillonnage complexes ne sont pas prises en charge.
- L'échantillonnage des `n` premiers avec `"Retirer l'échantillon"` n'est pas pris en charge.
- L'échantillonnage des `n` premiers avec `N>20000` n'est pas pris en charge.
- L'échantillonnage Tous les n'est pas pris en charge si `"Taille maximale de l'échantillon"` n'est pas défini.
- L'échantillonnage Tous les n'est pas pris en charge si `N * "Taille maximale de l'échantillon" > 20000`.
- Le pourcentage d'échantillonnage aléatoire au niveau des blocs n'est pas pris en charge.
- Le pourcentage aléatoire accepte la spécification d'une valeur de départ.

Agréger

- Les clés contiguës ne sont pas prises en charge. Si vous réutilisez un flux existant configuré pour trier les données, puis utilisez ce paramètre dans le noeud agrégé, changez ce flux afin de retirer le noeud Trier.
- Les statistiques d'ordre (Médiane, 1er quartile, 3ème quartile) sont calculées approximativement et prises en charge dans l'onglet Optimisation.

Trier

- L'onglet Optimisation n'est pas pris en charge.

Dans un environnement distribué, seul un nombre limité d'opérations conserve l'ordre des enregistrements établi par le noeud Trier.

- Un tri suivi d'un noeud d'exportation génère une source de données triée.
- Un tri suivi d'un noeud échantillon avec échantillonnage du **Premier** enregistrement renvoie les *N* premiers enregistrements.

En général, vous devez placer un noeud Trier aussi près que possible des opérations nécessitant le tri des enregistrements.

Fusionner

- La fusion par ordre n'est pas prise en charge.
- L'onglet Optimisation n'est pas pris en charge.
- Les opérations de fusion sont relativement lentes. Si vous disposez d'espace disponible dans HDFS, il peut être beaucoup plus rapide de fusionner vos sources de données une fois et d'utiliser la source fusionnée dans les flux suivants que de fusionner les sources de données dans chaque flux.

Transformation R

La syntaxe R dans le noeud doit être composée d'opérations d'enregistrement unique.

Opérations sur les champs

Toutes les opérations sur les champs sont prises en charge, sauf pour les noeuds Anonymiser, Transposer, Intervalles de temps et Historique. Vous trouverez ci-après des remarques supplémentaires sur la fonctionnalité des noeuds prise en charge.

Prép. auto. des données

- La formation du noeud n'est pas prise en charge. L'application à de nouvelles données des transformations figurant dans un noeud Prép. auto. des données formé est prise en charge.

Dériver

- Toutes les fonctions Dériver sont prises en charge, à l'exception des fonctions séquentielles.
- Le calcul d'un nouveau champ en tant que comptage est essentiellement une opération de séquence et n'est donc pas pris en charge.
- Les champs de scission ne peuvent pas être dérivés dans le flux qui les utilise comme scissions. Vous devrez donc créer deux flux : un pour dériver le champ de scission et un autre pour utiliser le champ comme scission.

Remplissage

- Prend en charge le même ensemble de fonctions que le noeud dériver.

Regroupement par casiers

La fonction suivante n'est pas prise en charge.

- Création d'intervalles optimale
- Rangs
- Quantiles -> Quantiles : Somme des valeurs

- Quantiles -> Ex-aequo : Conserver dans l'élément actuel et Attribuer aléatoirement
- Quantiles -> N personnalisé : Valeurs supérieures à 100 et toute valeur N où 100 % de N n'est pas égal à zéro.

Analyse RFM

- L'option Conserver dans l'élément actuel pour le traitement des valeurs ex-aequo n'est pas prise en charge. Les scores RFM (Récence, Fréquence et Monétaire) ne correspondront pas toujours à ceux calculés par Modeler à partir des mêmes données. Les plages de scores seront les mêmes mais les affectations de scores (numéros BIN) peuvent différer d'un point.

Graphes

Tous les noeuds graphiques sont pris en charge.

Modélisation

Les noeuds de modélisation suivants sont pris en charge : Série temporelle, TCM, Isotonic-AS, Modèle d'extension, Tree-AS, Arbre C&R, Quest, CHAID, Linéaire, Linear-AS, Réseau neuronal, GLE, LSVM, TwoStep-AS, Random Trees, STP, Règles d'association, XGBoost-AS, Random Forest et K-Means-AS. La fonction de ces noeuds est décrite en détail ultérieurement.

Linéaire

Lors de la création de modèles basés sur des données volumineuses, vous voudrez généralement modifier l'objectif en Très grands jeux de données, ou spécifier des scissions.

- La formation continue des modèles PSM existants n'est pas prise en charge.
- L'objectif de génération de modèle Standard est uniquement recommandé si les champs de scission sont définis de sorte que le nombre d'enregistrements dans chaque scission ne soit pas trop élevé, la définition du terme "trop élevé" dépendant de la puissance de chaque noeud dans votre cluster Hadoop. Cependant, vous devez également vous assurer que les scissions ne sont pas définies de sorte que le nombre d'enregistrements ne soit pas insuffisant pour créer un modèle.
- L'objectif Boosting n'est pas pris en charge.
- L'objectif Bagging n'est pas pris en charge.
- L'objectif Très grands jeux de données n'est pas recommandé lorsque le nombre d'enregistrements est réduit car, dans la plupart des cas, le modèle ne sera pas généré ou le modèle généré sera dégradé.
- La préparation automatique des données n'est pas prise en charge. Cela risque de générer des problèmes lors de la tentative de création d'un modèle basé sur des données comportant un grand nombre de valeurs manquantes, qui seraient normalement imputées dans le cadre de la préparation automatique des données. Une solution de contournement consisterait à utiliser un modèle d'arbre ou un réseau de neurones avec le paramètre Avancé pour imputer les valeurs manquantes sélectionnées.
- La statistique d'exactitude n'est pas calculée pour les modèles de scission.

Réseau de neurones

Lors de la création de modèles basés sur des données volumineuses, vous voudrez généralement modifier l'objectif en Très grands jeux de données, ou spécifier des scissions.

- La formation continue des modèles standard ou PSM existants n'est pas prise en charge.
- L'objectif de génération de modèle Standard est uniquement recommandé si les champs de scission sont définis de sorte que le nombre d'enregistrements dans chaque scission ne soit pas trop élevé, la définition du terme "trop élevé" dépendant de la puissance de chaque noeud dans votre cluster Hadoop. Cependant, vous devez également vous assurer que les scissions ne sont pas définies de sorte que le nombre d'enregistrements ne soit pas insuffisant pour créer un modèle.

- L'objectif Boosting n'est pas pris en charge.
- L'objectif Bagging n'est pas pris en charge.
- L'objectif Très grands jeux de données n'est pas recommandé lorsque le nombre d'enregistrements est réduit car, dans la plupart des cas, le modèle ne sera pas généré ou le modèle généré sera dégradé.
- Lorsque les données comportent un grand nombre de valeurs manquantes, utilisez le paramètre Avancé pour imputer les valeurs manquantes.
- La statistique d'exactitude n'est pas calculée pour les modèles de scission.

Arbre C&RT, CHAID et Quest

Lors de la création de modèles basés sur des données volumineuses, vous voudrez généralement modifier l'objectif en Très grands jeux de données, ou spécifier des scissions.

- La formation continue des modèles PSM existants n'est pas prise en charge.
- L'objectif de génération de modèle Standard est uniquement recommandé si les champs de scission sont définis de sorte que le nombre d'enregistrements dans chaque scission ne soit pas trop élevé, la définition du terme "trop élevé" dépendant de la puissance de chaque noeud dans votre cluster Hadoop. Cependant, vous devez également vous assurer que les scissions ne sont pas définies de sorte que le nombre d'enregistrements ne soit pas insuffisant pour créer un modèle.
- L'objectif Boosting n'est pas pris en charge.
- L'objectif Bagging n'est pas pris en charge.
- L'objectif Très grands jeux de données n'est pas recommandé lorsque le nombre d'enregistrements est réduit car, dans la plupart des cas, le modèle ne sera pas généré ou le modèle généré sera dégradé.
- Les sessions interactives ne sont pas prises en charge.
- La statistique d'exactitude n'est pas calculée pour les modèles de scission.
- En cas d'existence d'un champ de scission, les modèles d'arbre créés localement dans Modeler sont légèrement différents des modèles d'arbre créés par Analytic Server, et génèrent donc des scores différents. Les algorithmes dans ces deux cas sont valides. Les algorithmes utilisés par Analytic Server sont tout simplement plus récents. Comme les algorithmes d'arbre tendent à comporter de nombreuses règles heuristiques, la différence entre les deux composants est normale.

Evaluation de modèle

Tous les modèles pris en charge pour la modélisation sont également pris en charge pour l'évaluation. De plus, les nuggets de modèle intégrés localement pour les noeuds suivants sont pris en charge pour l'évaluation : C&RT, Quest, CHAID, Linéaire et Réseau neuronal (que le modèle soit standard, boosted bagged ou pour des jeux de données très volumineux), Régression, C5.0, Logistique, Modèles linéaires généralisés, GLMM, Cox, SVM, Réseau Bayésien, TwoStep, KNN, Liste de décision, Discriminant, Auto-formation, Détection des anomalies, Apriori, Carma, K Moyennes, Kohonen, R et Exploration de texte.

- Aucune propension brute ou ajustée ne sera évaluée. Comme solution de contournement, vous pouvez obtenir le même effet en calculant manuellement la propension brute à l'aide d'un noeud dériver avec l'expression suivante : if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif

R La syntaxe R dans le nugget doit être composée d'opérations d'enregistrement unique.

Sortie Les noeuds Matrice, Analyse, Audit des données, Transformation, Valeurs globales, Statistiques, Moyennes et Table sont pris en charge. Vous trouverez ci-après des remarques supplémentaires sur la fonctionnalité des noeuds prise en charge.

Audit données

Le noeud Audit des données ne peut pas générer le mode des champs continus.

Moyennes
Le noeud Moyennes ne peut pas générer une erreur standard ou un intervalle de confiance de 95 %.

Table Le noeud Table est pris en charge via l'écriture d'une source de données Analytic Server temporaire contenant les résultats d'opérations en amont. Il parcourt ensuite le contenu de cette source de données.

Exporter
Un flux peut commencer par un noeud source Analytic Server et se terminer par un noeud d'exportation différent du noeud d'exportation Analytic Server, mais les données seront déplacées de HDFS vers SPSS Modeler Server, puis vers l'emplacement d'exportation.

---

noeud source de base de données

Le noeud source de base de données peut être utilisé pour importer des données provenant de nombreux autres logiciels utilisant la connectivité ODBC (Open Database Connectivity), tels que Microsoft SQL Server, Db2, Oracle, etc.

Pour lire ou écrire sur une base de données, vous devez installer et configurer une source de données ODBC pour la base de données appropriée, avec, le cas échéant, des autorisations en lecture et en écriture. IBM SPSS Data Access Pack contient un ensemble de pilotes ODBC qui peuvent être utilisés dans ce but, et ces pilotes sont disponibles sur le site de téléchargement. Si vous avez des questions sur la création ou la définition d'autorisations pour les sources de données ODBC, contactez l'administrateur de votre base de données.

Pilotes ODBC pris en charge

Pour obtenir les informations les plus récentes sur les bases de données et pilotes ODBC pris en charge et testés pour une utilisation avec IBM SPSS Modeler, consultez les matrices de compatibilité des produits sur le site Web de support technique de l'entreprise (http://www.ibm.com/support).

Où installer les pilotes

Remarque : Vous devez installer et configurer les pilotes ODBC sur chaque ordinateur sur lequel le traitement peut avoir lieu.

- Si IBM SPSS Modeler est exécuté en mode local (autonome), les pilotes doivent être installés sur l'ordinateur local.
- Si vous exécutez IBM SPSS Modeler en mode distribué et IBM SPSS Modeler Server en mode distant, les pilotes ODBC doivent être installés sur le même ordinateur d'installation qu'IBM SPSS Modeler Server. Pour IBM SPSS Modeler Server sur les systèmes UNIX, consultez également "Configuration des pilotes ODBC sur les systèmes UNIX" plus loin dans cette section.
- Si vous devez accéder aux mêmes sources de données provenant d'IBM SPSS Modeler et de IBM SPSS Modeler Server, les pilotes ODBC doivent être installés sur les deux ordinateurs.
- Si IBM SPSS Modeler est exécuté via Terminal Services, les pilotes ODBC doivent être installés sur le même serveur Terminal Services qu'IBM SPSS Modeler.

Accès aux données à partir d'une base de données

Pour accéder à des données dans une base de données, procédez comme suit.

- Installez un pilote ODBC et configurez une source de données pour la base de données à utiliser.
- Dans la boîte de dialogue du noeud Base de données, connectez-vous à une base de données en mode Table ou en mode Requête SQL.
- Sélectionnez une table dans la base de données.

- Dans les onglets de la boîte de dialogue du noeud Base de données, vous pouvez modifier les types d'utilisation et filtrer les champs de données.

Les rubriques de la documentation connexes comportent des détails sur les étapes précédentes.

Remarque : Si vous appelez des procédures mémorisées de base de données depuis SPSS Modeler, un champ de sortie unique appelé `RowsAffected` peut être renvoyé plutôt que la sortie attendue de la procédure mémorisée. Tel est le cas lorsque ODBC ne renvoie pas suffisamment d'informations pour pouvoir déterminer le modèle de données de sortie de la procédure mémorisée. SPSS Modeler ne propose qu'une prise en charge limitée des procédures mémorisées qui renvoient une sortie et il est recommandé, au lieu d'utiliser des procédures mémorisées, d'extraire l'instruction SELECT de la procédure mémorisée et d'utiliser les actions ci-après.

- Créez une vue reposant sur l'instruction SELECT et choisissez-la dans le noeud source Base de données.
- Utilisez l'instruction SELECT directement dans le noeud source Base de données.

Définition des options du noeud SGBD

Vous pouvez utiliser les options de l'onglet Données de la boîte de dialogue du noeud Source de base de données pour accéder à une base de données et lire les données d'une table sélectionnée.

Mode. Sélectionnez **Table** pour vous connecter à une table à l'aide des commandes de la boîte de dialogue.

Sélectionnez **Requête SQL** pour interroger la base de données sélectionnée en utilisant SQL. Pour plus d'informations, voir «Interrogation de la base de données», à la page 26.

Source de données. En mode Table et Requête SQL, vous pouvez entrer un nom dans le champ Source de données ou sélectionner **Ajouter une nouvelle connexion à la base de données** dans la liste déroulante..

Les options suivantes vous permettent de vous connecter à une base de données et de sélectionner une table à l'aide de la boîte de dialogue :

Nom de la table. Si vous connaissez le nom de la table à laquelle vous souhaitez accéder, indiquez-le dans le champ Nom de la table. Dans le cas contraire, cliquez sur le bouton **Sélectionner** pour ouvrir une boîte de dialogue répertoriant les tables disponibles.

Entourer de guillemets les noms des tables et colonnes. Indiquez si vous souhaitez que les noms des tables et des colonnes soient placés entre guillemets lors de l'envoi des requêtes à la base de données (par exemple, s'ils contiennent des espaces ou des signes de ponctuation).

- Si vous sélectionnez **Si nécessaire**, les noms des tables et des champs seront placés entre guillemets *uniquement* s'ils contiennent des caractères non standard. Les caractères non standard sont les caractères non ASCII, l'espace et tous les caractères non alphanumériques autres que le point (.).
- Sélectionnez **Toujours** si vous souhaitez que *tous* les noms de tables et de champs soient mis entre guillemets.
- Sélectionnez **Jamais** si vous ne souhaitez *jamais* mettre les noms des tables et des champs entre guillemets.

Supprimer les espaces de début et de fin. Sélectionnez les options permettant la suppression des espaces situés en début et en fin des chaînes.

Remarque. Les comparaisons entre des chaînes qui utilisent ou nom les conversions SQL peuvent générer des ensembles de résultats différents en présence d'espaces situés en fin des chaînes.

Lecture de chaînes vides issues d'Oracle. Lorsque vous lisez une base de données Oracle ou que vous écrivez dedans, souvenez-vous que, contrairement à IBM SPSS Modeler et à la plupart des autres bases de données, Oracle traite et stocke les valeurs de chaîne vides comme des valeurs nulles. Autrement dit, les mêmes données extraites d'une base de données Oracle, ou d'un fichier ou d'une autre base de données peuvent se comporter différemment, et donc renvoyer des résultats différents.

Ajout d'une connexion de base de données

Pour ouvrir une base de données, sélectionnez d'abord la source de données à laquelle vous voulez vous connecter. Dans l'onglet Données, sélectionnez **Ajouter une nouvelle connexion à la base de données** dans la liste déroulante Source de données.

La boîte de dialogue Connexions de base de données apparaît.

Remarque : Vous pouvez aussi ouvrir cette boîte de dialogue à partir du menu principal en sélectionnant **Outils** > **Bases de données...**

Sources de données. Répertorie les sources de données disponibles. Si la base de données souhaitée n'apparaît pas, faites défiler la liste. Une fois que vous avez sélectionné la source de données et entré les mots de passe, cliquez sur **Connecter**. Cliquez sur **Rafraîchir** pour mettre à jour la liste.

Nom d'utilisateur et mot de passe. Si la source de données est protégée par un mot de passe, entrez votre nom d'utilisateur et le mot de passe associé.

Informations d'identification. Si des données d'identification ont été configurées dans IBM SPSS Collaboration and Deployment Services, vous pouvez sélectionner cette option pour les rechercher dans le référentiel. Le nom d'utilisateur et le mot de passe constituant les données d'identification doivent correspondre au nom d'utilisateur et au mot de passe requis pour accéder à la base de données.

Connexions. Indique les bases de données actuellement connectées.

- **Par défaut.** En option, vous pouvez choisir une connexion par défaut. Cette action prédéfinit cette connexion comme source de données des noeuds source de base de données et d'exportation, mais peut être modifiée au besoin.
- **Sauvegarder.** Vous pouvez également sélectionner une ou plusieurs connexions à afficher de nouveau dans les prochaines sessions.
- **Source de données.** Les chaînes de connexion pour les bases de données actuellement connectées.
- **Prédéfinir.** Indique (à l'aide du caractère *) si des valeurs prédéfinies ont été spécifiées pour la connexion de la base de données. Pour spécifier des valeurs prédéfinies, cliquez sur cette colonne dans la ligne correspondant à la connexion de la base de données et choisissez Spécifier dans la liste. Pour plus d'informations, voir «Spécification de valeurs prédéfinies pour une connexion de la base de données», à la page 22.

Pour supprimer des connexions, sélectionnez-les dans la liste, puis cliquez sur **Supprimer**.

Une fois vos sélections effectuées, cliquez sur **OK**.

Pour lire ou écrire sur une base de données, vous devez installer et configurer une source de données ODBC pour la base de données appropriée, avec, le cas échéant, des autorisations en lecture et en écriture. IBM SPSS Data Access Pack contient un ensemble de pilotes ODBC qui peuvent être utilisés dans ce but, et ces pilotes sont disponibles sur le site de téléchargement. Si vous avez des questions sur la création ou la définition d'autorisations pour les sources de données ODBC, contactez l'administrateur de votre base de données.

Pilotes ODBC pris en charge

Pour obtenir les informations les plus récentes sur les bases de données et pilotes ODBC pris en charge et testés pour une utilisation avec IBM SPSS Modeler, consultez les matrices de compatibilité des produits sur le site Web de support technique de l'entreprise (http://www.ibm.com/support).

Où installer les pilotes

Remarque : Vous devez installer et configurer les pilotes ODBC sur chaque ordinateur sur lequel le traitement peut avoir lieu.

- Si IBM SPSS Modeler est exécuté en mode local (autonome), les pilotes doivent être installés sur l'ordinateur local.
- Si vous exécutez IBM SPSS Modeler en mode distribué et IBM SPSS Modeler Server en mode distant, les pilotes ODBC doivent être installés sur le même ordinateur d'installation qu'IBM SPSS Modeler Server. Pour IBM SPSS Modeler Server sur les systèmes UNIX, consultez également "Configuration des pilotes ODBC sur les systèmes UNIX" plus loin dans cette section.
- Si vous devez accéder aux mêmes sources de données provenant d'IBM SPSS Modeler et de IBM SPSS Modeler Server, les pilotes ODBC doivent être installés sur les deux ordinateurs.
- Si IBM SPSS Modeler est exécuté via Terminal Services, les pilotes ODBC doivent être installés sur le même serveur Terminal Services qu'IBM SPSS Modeler.

Configuration des pilotes ODBC sur les systèmes UNIX

Par défaut, le gestionnaire de pilote DataDirect n'est pas configuré pour IBM SPSS Modeler Server sur les systèmes UNIX. Pour configurer le chargement du gestionnaire de pilote DataDirect sur UNIX, saisissez les commandes suivantes :

```
cd <répertoire_install_modeler_server>/bin
rm -f libspssodbc.so
ln -s libspssodbc_datadirect.so libspssodbc.so
```

Le lien par défaut est alors supprimé et un lien vers le gestionnaire de pilote DataDirect est créé.

Remarque : L'encapsuleur de pilote UTF16 est obligatoire pour utiliser les pilotes SAP HANA ou IBM Db2 CLI de certaines bases de données. DashDB requiert le pilote IBM Db2 CLI. Pour créer un lien vers l'encapsuleur de pilote UTF16, entrez plutôt la commande suivante :

```
rm -f libspssodbc.so
ln -s libspssodbc_datadirect_utf16.so libspssodbc.so
```

Pour configurer SPSS Modeler Server :

1. Configurez le script de démarrage de SPSS Modeler Server `modelersrv.sh` pour sourcer le fichier de l'environnement IBM SPSS Data Access Pack `odbc.sh` en ajoutant la ligne suivante dans `modelersrv.sh` :

   ```
   . /<chemin_install_SDAP>/odbc.sh
   ```

   Où `<chemin_install_SDAP>` est le chemin d'accès complet à votre installation IBM SPSS Data Access Pack.
2. Redémarrez SPSS Modeler Server.

En outre, pour SAP HANA et IBM Db2 uniquement, ajoutez la définition de paramètre suivante au nom de système par défaut dans le fichier `odbc.ini` afin d'éviter les dépassements de mémoire tampon durant la connexion :

```
DriverUnicodeType=1
```

Remarque : L'encapsuleur `libspssodbc_datadirect_utf16.so` est également compatible avec les autres pilotes ODBC pris en charge par SPSS Modeler Server.

Problèmes potentiels de base de données

En fonction de la base de données que vous utilisez, vous devez être conscient de certains problèmes potentiels.

IBM Db2

Lors de la tentative de mise en cache d'un noeud dans un flux qui lit les données d'une base de données Db2, vous pouvez voir le message d'erreur suivant :

```
Il n'existe pas d'espace table par défaut possédant une taille de page d'au moins 4096 que l'ID autorisation TEST peut utiliser
```

Pour configurer Db2 afin de permettre à la mise en mémoire cache dans les bases de données de fonctionner correctement dans SPSS Modeler, l'administrateur de base de données doit créer un espace table "temporaire pour les utilisateurs" et octroyer l'accès à cet espace table aux comptes Db2 appropriés.

Nous vous recommandons d'utiliser une taille de page de 32768 dans le nouvel espace table, car cela augmentera la limite du nombre de champs pouvant être mis en cache.

IBM Db2 for z/OS

- L'évaluation d'un sous-ensemble d'algorithmes, avec les niveaux de fiabilité activés, à l'aide d'un SQL généré peut renvoyer une erreur lors de l'exécution. Le problème est spécifique à Db2 for z/OS ; pour le résoudre, utilisez l'adaptateur de scoring SPSS Modeler Server pour Db2 on z/OS.
- Lors de l'exécution de flux sur Db2 for z/OS, vous pouvez rencontrer des erreurs de base de données si le délai d'attente des connexions de base de données inactives est activé et défini sur une valeur trop faible. Dans Db2 for z/OS version 8, le délai d'attente n'est plus nul, mais égal à deux minutes. La solution consiste à augmenter la valeur du paramètre système Db2 `IDLE THREAD TIMEOUT (IDTHTOIN)` ou de réinitialiser la valeur à 0.

Oracle

Lorsque vous exécutez un flux contenant un noeud Agréger, les valeurs renvoyées pour le premier et le troisième quartiles, lors de la conversion des instructions SQL dans une base de données Oracle, peuvent différer de celles renvoyées en mode natif.

Spécification de valeurs prédéfinies pour une connexion de la base de données

Pour certaines bases de données, il est possible de spécifier des paramètres par défaut pour la connexion de la base de données. Les paramètres concernent tous l'exportation de la base de données.

Les bases de donnés prenant en charge cette fonctionnalité sont les suivantes.

- Editions SQL Server Enterprise et Developer. Pour plus d'informations, voir «Paramètres pour SQL Server», à la page 23.
- Editions Oracle Enterprise ou Personal. Pour plus d'informations, voir «Paramètres pour Oracle», à la page 23.
- IBM Db2 for z/OS et Teradata se connectent à une base de données ou à un schéma de la même façon. Pour plus d'informations, voir «Paramètres d'IBM Db2 for z/OS, IBM Db2 LUW et Teradata», à la page 24.

Si vous êtes connecté à une base de données ou à un schéma qui ne prend pas en charge cette fonctionnalité, le message **Aucun préréglage ne peut être configuré pour cette connexion à la base de données** apparaît.

Paramètres pour SQL Server

Ces paramètres sont affichés pour les éditions SQL Server 2008 et Developer.

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression de. Choisissez le niveau de compression.

- **Ligne.** Active la compression au niveau des lignes (par exemple, l'équivalent de `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW);` dans SQL).
- **Page.** Active la compression au niveau des pages (par exemple `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);` dans SQL).

Paramètres pour Oracle

Paramètres d'Oracle - Option basique

Ces paramètres sont affichés pour Oracle édition Enterprise ou Personal avec l'option Basique.

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression de. Choisissez le niveau de compression.

- **Par défaut.** Active la compression par défaut (par exemple `CREATE TABLE MYTABLE(...) COMPRESS;` dans SQL). Dans ce cas, cela a le même effet que l'option **Basique**.
- **De base.** Active la compression de base (par exemple `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` dans SQL).

Paramètres d'Oracle - Option avancée

Ces paramètres sont affichés pour Oracle édition Enterprise ou Personal avec l'option Avancée.

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression de. Choisissez le niveau de compression.

- **Par défaut.** Active la compression par défaut (par exemple `CREATE TABLE MYTABLE(...) COMPRESS;` dans SQL). Dans ce cas, cela a le même effet que l'option **Basique**.
- **De base.** Active la compression de base (par exemple `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` dans SQL).
- **OLTP.** Active la compression OLTP (par exemple `CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP;` dans SQL).
- **Requête faible/élevée.** (Serveurs Exadata uniquement) Active la compression Exadata Hybrid Columnar Compression pour les requêtes (par exemple `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW;` ou `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH;` dans SQL). La compression des requêtes est utile dans les environnements d'entreposage de données ; `HIGH` fournit un rapport de compression plus grand que `LOW`.
- **Archive faible/élevée.** (Serveurs Exadata uniquement) Active la compression Exadata Hybrid Columnar Compression pour les archives (par exemple `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW;` ou `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH;` dans SQL). La compression des archives est utile pour compresser des données qui seront stockées pendant de longues périodes ; `HIGH` fournit un rapport de compression plus grand que `LOW`.

Paramètres d'IBM Db2 for z/OS, IBM Db2 LUW et Teradata

Lorsque vous spécifiez des préréglages pour IBM Db2 for z/OS, IBM Db2 LUW ou Teradata, vous êtes invité à sélectionner les éléments suivants :

Utiliser la base de données de l'adaptateur Server Scoring ou **Utiliser le schéma de l'adaptateur Serveur Scoring**. Si l'une de ces options est sélectionnée, elle active l'option **Base de données de l'adaptateur Server Scoring** ou **Schéma de l'adaptateur Server Scoring**.

Base de données de l'adaptateur Server Scoring ou **Schéma de l'adaptateur Server Scoring** Dans la liste déroulante, sélectionnez la connexion dont vous avez besoin.

De plus, pour Teradata, vous pouvez aussi définir les détails des bandes de requêtes afin de fournir des métadonnées supplémentaires pour des opérations telles que la gestion des charges de travail, le classement, l'identification et la résolution des requêtes, ainsi que le suivi de l'utilisation de la base de données.

Orthographe des bandes de requêtes. Décidez si les bandes de requêtes doivent être définies une fois pour toute la période de travail avec une connexion de base de données Teradata (**Pour la session**) ou si elles doivent être définies à chaque fois que vous exécutez un flux (**Pour la transaction**).

Remarque : Si vous définissez des bandes de requêtes pour un flux, elles sont perdues si vous copiez le flux sur une autre machine. Pour empêcher cette perte, vous pouvez utiliser un script afin d'exécuter votre flux et utiliser le mot clé *querybanding* dans le script pour appliquer les paramètres de votre choix.

Droits de base de données requis

Pour que les options de la base de données SPSS Modeler fonctionnent correctement, accordez les droits d'accès aux éléments suivants à n'importe quel ID utilisateur :

Db2 LUW

SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABCONSTPKC

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSCAT.TABLESPACES

SYSCAT.SCHEMATA

Db2/z SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABCONSTPKC

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSIBM.SYSDUMMYU

SYSIBM.SYSPACKSTMT

Teradata

DBC.Functions

DBC.USERS

Sélection d'une table de base de données

Une fois connecté à une source de données, vous pouvez importer des champs à partir d'une table ou d'une vue donnée. Dans l'onglet Données de la boîte de dialogue SGBD, vous pouvez entrer le nom d'une table dans le champ Nom de la table ou cliquer sur **Sélectionner** pour ouvrir la boîte de dialogue Sélectionner Table/Vue répertoriant les tables et vues disponibles.

Afficher le propriétaire de la table. Cochez cette case si l'accès à une table d'une source de données requiert que le propriétaire de la table soit identifié. Désélectionnez cette option pour les sources de données qui ne requièrent pas cette identification.

Remarque : Pour les bases de données SAS et Oracle, l'identification du propriétaire est généralement obligatoire.

Tables et vues. Sélectionnez la table ou la vue à importer.

Afficher. Répertorie les colonnes de la source de données à laquelle vous êtes actuellement connecté. Cliquez sur l'une des options suivantes pour personnaliser la vue des tables disponibles :

- Cliquez sur **Tables utilisateur** pour afficher les tables de base de données ordinaires créées par les utilisateurs de la base de données.
- Cliquez sur **Tables système** pour afficher les tables de base de données appartenant au système (par exemple, les tables qui fournissent des informations sur la base de données, comme les détails des

index). Cette option peut être utilisée pour afficher les onglets utilisés dans les bases de données Excel. (Un noeud source Excel distinct est également disponible. Pour plus d'informations, reportez-vous à la rubrique «Noeud source Excel», à la page 46.)

- Cliquez sur **Affichage** pour afficher les tables virtuelles basées sur une requête impliquant des tables ordinaires.
- Cliquez sur **Synonymes** pour afficher les synonymes créés dans la base de données pour toute table existante.

Filtres Nom/Propriétaire. Ces champs vous permettent de filtrer la liste des tables affichées par nom ou propriétaire. Par exemple, saisissez `SYS` pour répertorier les tables de ce propriétaire uniquement. Pour les recherches basées sur des caractères génériques, un caractère de soulignement (_) peut être utilisé pour représenter un caractère ; un caractère pourcentage (%) peut correspondre à une séquence d'au moins zéro caractère.

Définir par défaut. Enregistre les paramètres actuels en tant que paramètres par défaut de l'utilisateur actuel. Ces paramètres seront restaurés ultérieurement lorsqu'un utilisateur ouvrira une nouvelle boîte de dialogue du sélecteur de table *pour les mêmes nom de source de données et connexion utilisateur uniquement*.

Interrogation de la base de données

Une fois connecté à une source de données, vous pouvez choisir d'importer des champs à l'aide de requêtes SQL. Dans la boîte de dialogue principale, sélectionnez **Requête SQL** comme mode de connexion. Cette opération ajoute une fenêtre d'éditeur de requêtes dans la boîte de dialogue. A l'aide de l'éditeur de requêtes, vous pouvez créer ou charger une ou plusieurs requêtes SQL dont les résultats seront lus dans le flux de données.

Si vous spécifiez plusieurs requêtes SQL, séparez-les par des points-virgules (;) et assurez-vous qu'il n'y a pas d'instruction SELECTIONNER multiple.

Pour annuler et fermer la fenêtre de l'éditeur de requêtes, sélectionnez **Table** comme mode de connexion.

Vous pouvez inclure des paramètres de flux SPSS Modeler (un type de variable définie par l'utilisateur) dans les requêtes SQL. Pour plus d'informations, voir «Utilisation des paramètres de flux dans une requête SQL».

Charger la requête. Cliquez sur cette option pour ouvrir le navigateur, que vous pouvez utiliser pour charger une requête précédemment enregistrée.

Enregistrer la requête. Cliquez sur cette option pour ouvrir la boîte de dialogue Enregistrer la requête, que vous pouvez utiliser pour enregistrer la requête en cours.

Importer les éléments par défaut. Cliquez sur cette option pour importer un exemple d'instruction `SQL SELECT` créé automatiquement à l'aide de la table et des colonnes sélectionnées dans la boîte de dialogue.

Effacer. Efface le contenu de la zone de travail. Utilisez cette option pour tout annuler et recommencer.

Scinder le texte. L'option par défaut **Jamais** signifie que la requête sera envoyée à la base de données sous forme globale. Vous pouvez aussi sélectionner **Selon les besoins** qui signifie que SPSS Modeler tente d'analyser la requête et d'identifier si elle contient des instructions SQL qui devraient être envoyées à la base de données les unes après les autres.

Utilisation des paramètres de flux dans une requête SQL

Lors de la rédaction d'une requête SQL pour l'importation de champs, vous pouvez inclure des paramètres de flux SPSS Modeler précédemment définis. Tous les types de paramètre de flux sont pris en charge.

Le tableau suivant présente quelques exemples d'interprétation de paramètres de flux dans les requêtes SQL.

Tableau 3. Exemples de paramètres de flux.

Nom du paramètre de flux (exemple)	Stockage	Valeur du paramètre de flux	Interprété comme
PString	Chaîne	ss	'ss'
PInt	Entier	5	5
PReal	Réel	5.5	5.5
PTime	Heure	23:05:01	t{'23:05:01'}
PDate	Date	2011-03-02	d{'2011-03-02'}
PTimeStamp	Horodatage	2011-03-02 23:05:01	ts{'2011-03-02 23:05:01'}
PColumn	Inconnu	IntValue	IntValue

Dans la requête SQL, vous spécifiez un paramètre de flux de la même manière qu'une expression CLEM, à savoir par `'$P-<parameter_name>'`, où `<parameter_name>` est le nom qui a été défini pour le paramètre de flux.

Lors du référencement d'un champ, le type de stockage doit être défini comme Unknown, et la valeur du paramètre doit être entre guillemets, le cas échéant. Par exemple, à l'aide des exemples du tableau, si vous saisissez la requête SQL suivante :

```
select "IntValue" from Table1 where "IntValue" < '$P-PInt';
```

elle sera interprétée comme :

```
select "IntValue" from Table1 where "IntValue" < 5;
```

Si vous référencez le champ `IntValue` à l'aide du paramètre `PColumn`, vous devrez spécifier la requête comme suit pour obtenir le même résultat :

```
select "IntValue" from Table1 where "'$P-PColumn'" < '$P-PInt';
```

Noeud Délimité

Vous pouvez utiliser des noeuds Délimité pour lire les données de fichiers texte de longueur variable (fichiers dont les enregistrements contiennent un nombre fixe de champs et un nombre variable de caractères), aussi appelés fichiers texte délimités. Ce type de noeud est également utile pour les fichiers contenant des textes d'en-tête de longueur fixe et certains types d'annotation. Les enregistrement sont lus un à la fois et transmis via le flux jusqu'à la lecture de la totalité du fichier.

Remarque relative à la lecture dans les données géospatiales

Si le noeud contient des données géospatiales et qu'il a été créé en tant qu'exportation depuis un fichier à plat, vous devez effectuer des étapes supplémentaires de configuration des métadonnées géospatiales. Pour plus d'informations, voir «Importation de données géospatiales dans le noeud Délimité», à la page 30.

Remarques relatives à la lecture dans les données de texte délimité

- Les enregistrements doivent être délimités par un caractère de retour à la ligne à la fin de chaque ligne. Le caractère de retour à la ligne ne doit pas être utilisé à une autre fin (par exemple, dans un nom ou une valeur de champ). Les espaces de début et de fin doivent être supprimés pour économiser de l'espace. Ceci n'est cependant pas crucial. En option, ces espaces peuvent être supprimés par le noeud.
- Les champs doivent être délimités par une virgule ou un autre caractère utilisé uniquement comme délimiteur, ce qui signifie qu'il ne figure pas dans les noms ou les valeurs de champ. Si ce n'est pas

possible, tous les champs de texte peuvent être placés entre guillemets, si aucun nom de champ ou aucune valeur textuelle ne contient des guillemets. Si un nom de champ ou une valeur contient des guillemets, les champs de texte peuvent être placés entre apostrophes, si les noms ou les valeurs de champ ne comportent pas d'apostrophe. Si vous ne pouvez pas utiliser de guillemets ni d'apostrophes, les valeurs textuelles doivent être modifiées et vous devez supprimer ou remplacer le caractère délimiteur, les apostrophes ou les guillemets.

- Chaque ligne, dont celle d'en-tête, doit contenir le même nombre de champs.
- La première ligne doit contenir les noms de champ. Si tel n'est pas le cas, désélectionnez **Lire les noms des champs à partir du fichier** pour attribuer à chaque champ un nom générique tel que Champ1, Champ2, etc.
- La deuxième ligne doit contenir le premier enregistrement de données. Il ne doit y avoir ni ligne vierge ni commentaire.
- Les valeurs numériques ne doivent pas inclure le séparateur des milliers ou le symbole de groupement sans la virgule dans 3,000.00, par exemple. L'indicateur de décimale (point aux Etats-Unis et au Royaume-Uni) doit être utilisé si nécessaire.
- les valeurs de date et d'heure doivent être exprimées dans l'un des formats reconnus dans la boîte de dialogue des options de flux, par exemple `JJ/MM/AAAA` ou `HH:MM:SS`. Idéalement, tous les champs de date et d'heure dans le fichier doivent respecter le même format, et tout champ contenant une date doit utiliser le même format pour toutes les valeurs dans ce champ.

Définition des options pour le noeud Délimité

Vous définissez les options sous l'onglet Fichier de la boîte de dialogue noeud Délimité.

Fichier Spécifiez le nom du fichier. Vous pouvez entrer un nom de fichier ou cliquer sur le bouton représentant des points de suspension (**...**) pour sélectionner un fichier. Le chemin d'accès au fichier apparaît lorsque vous sélectionnez un fichier. Son contenu est affiché avec des délimiteurs dans le panneau situé en dessous.

Vous pouvez copier et coller l'exemple de texte affiché à partir de votre source de données dans les contrôles suivants : caractères de commentaires de fin de ligne et délimiteurs spécifiés par l'utilisateur. Utilisez les raccourcis clavier Ctrl+C et Ctrl+V pour effectuer le copier-coller.

Lire les noms des champs à partir du fichier Sélectionnée par défaut, cette option traite la première ligne dans le fichier de données en tant que libellés pour la colonne. Si la première ligne n'est pas un en-tête, désélectionnez l'option pour attribuer automatiquement à chaque champ du jeu de données un nom générique, comme *Champ1, Champ2*.

Indiquer le nombre de champs. Indiquez le nombre de champs dans chaque enregistrement. Le nombre de champs peut être détecté automatiquementsi les enregistrements se terminent par des caractères de nouvelle ligne. Vous pouvez également entrer directement un nombre.

Ignorer les caractères des en-têtes. Indiquez le nombre de caractères à ignorer au début du premier enregistrement.

Caractères de commentaires fin de ligne. Spécifiez quels caractères (comme # ou !) indiquent des annotations dans les données. Lorsque l'un de ces caractères apparaît dans le fichier, toutes les données situées entre ce caractère et le caractère de nouvelle ligne suivant (non inclus) sont ignorées.

Supprimer les espaces de début et de fin. Sélectionnez les options permettant la suppression des espaces situés en début et en fin des chaînes lors de l'importation.

Remarque. Les comparaisons entre des chaînes qui utilisent ou nom les conversions SQL peuvent générer des ensembles de résultats différents en présence d'espaces situés en fin des chaînes.

Caractères non valides. Sélectionnez **Supprimer** pour supprimer les caractères non valides de la source de données. Sélectionnez **Remplacer par** pour remplacer les caractères non valides par le symbole indiqué (un caractère uniquement). Les caractères non valides sont des caractères nuls (0) ou des caractères qui n'existent pas dans la méthode de codage spécifiée.

Codage. Indique la méthode de codage de texte employée. Vous pouvez choisir la valeur par défaut du système, la valeur par défaut du flux ou UTF-8.

- Si le système est exécuté en mode réparti, sa valeur par défaut est spécifiée dans le Panneau de configuration de Windows de l'ordinateur serveur.
- La valeur par défaut du flux est spécifiée dans la boîte de dialogue Propriétés du flux.

Symbole décimal Sélectionnez le type de séparateur décimal qui est utilisé dans votre source de données. **Valeur par défaut flux** est le caractère qui est sélectionné dans l'onglet Options de la boîte de dialogue des propriétés de flux. Sinon, sélectionnez **Point (.)** ou **Virgule (,)** pour lire toutes les données de cette boîte de dialogue à l'aide du caractère choisi comme séparateur décimal.

Le délimiteur de ligne est un caractère de nouvelle ligne Pour utiliser le caractère de retour à la ligne au lieu d'un délimiteur de zone, sélectionnez cette option. Par exemple, cette option peut être utile si une ligne est coupée du fait que le nombre de délimiteurs sur cette ligne est impair. Remarque : si vous sélectionnez cette option, vous ne pouvez pas sélectionner **Nouvelle ligne** dans la liste des délimiteurs.

Remarque : Si vous sélectionnez cette option, les valeurs vides à la fin des lignes de données sont supprimées.

Délimiteurs. A l'aide des cases à cocher répertoriées pour cette commande, vous pouvez spécifier quels caractères (comme la virgule) marquent les limites des champs dans le fichier. Vous pouvez indiquer plusieurs délimiteurs(, | par exemple) pour les enregistrements qui font appel à des délimiteurs multiples. Le séparateur par défaut est la virgule.

Remarque : si la virgule est également définie en tant que séparateur décimal, les paramètres par défaut fournis ne fonctionnent pas. Si la virgule sert à la fois de séparateur de champs et de séparateur décimal, sélectionnez **Autre** dans la liste Séparateurs. Ensuite, ajoutez manuellement une virgule dans le champ de saisie.

Sélectionnez **Autoriser plusieurs délimiteurs non renseignés** pour considérer plusieurs délimiteurs non renseignés adjacents comme un délimiteur unique. Par exemple, une séquence constituée d'une valeur de données suivie de quatre espaces, puis d'une autre valeur de données, sera considérée comme séquence à deux champs, et non comme une séquence à cinq champs.

Lignes à analyser pour la colonne et le type Spécifiez le nombre de lignes et de colonnes à analyser pour les types de données spécifiés.

Reconnaître automatiquement les dates et les heures Pour qu'IBM SPSS Modeler puisse tenter automatiquement de reconnaître les entrées de données en tant que dates et heures, sélectionnez cette case à cocher. Par exemple, cela signifie qu'une entrée telle que 07-11-1965 sera identifiée comme une date et que 02:35:58 sera identifié comme une heure. Cependant, des entrées telles que 07111965 ou 023558 seront affichées sous la forme d'un entier car il n'y a pas de délimiteurs entre les nombres.

Remarque : Pour éviter tout problème potentiel lié aux données lorsque vous utilisez des fichiers de données provenant de versions précédentes d'IBM SPSS Modeler, cette case est désélectionnée par défaut pour les informations qui sont sauvegardées dans des versions antérieures à la version 13.

Traiter les crochets comme des listes Si vous sélectionnez cette case à cocher, les données incluses entre les crochets ouvrant et fermant sont traitées comme une valeur unique, même si le contenu inclut des caractères délimiteurs tels que des virgules et des guillemets. Tel est le cas par exemple pour des données

géospatiales en deux ou trois dimensions, où les coordonnées contenues entre crochets sont traitées comme un élément de liste unique. Pour plus d'informations, voir «Importation de données géospatiales dans le noeud Délimité»

Guillemets. A l'aide des listes déroulantes, vous pouvez indiquer la façon dont les guillemets simples et doubles sont traités lors de l'importation. Vous pouvez choisir l'option **Supprimer** (supprime tous les guillemets), **Inclure comme texte** (inclut les guillemets dans la valeur du champ) ou **Apparier et supprimer** (supprime des paires de guillemets). Si un guillemet n'est pas apparié, un message d'erreur apparaît. Si vous sélectionnez **Supprimer** ou **Apparier et supprimer**, la valeur du champ (sans les guillemets) est stockée sous forme de chaîne.

Remarque : Lorsque vous utilisez **Apparier et supprimer**, les espaces sont conservés. Lorsque vous utilisez **Supprimer**, espaces en début et en fin à l'intérieur et à l'extérieur des guillemets sont supprimés (par exemple, `' " ab c" , "d ef " , " gh i " '` donnent `'ab c, d ef, gh i'`). Lorsque vous utilisez **Inclure comme texte**, les guillemets sont traités comme des caractères normaux. Les espaces situés en début et en fin sont supprimés naturellement.

A tout moment lorsque vous travaillez dans cette boîte de dialogue, cliquez sur **Rafraîchir** pour recharger les champs à partir de la source de données. Ceci est utile lorsque vous modifiez des connexions des données au noeud source ou lorsque vous utilisez les différents onglets de la boîte de dialogue.

Importation de données géospatiales dans le noeud Délimité

Si le noeud contient des données géospatiales, a été créé en tant qu'exportation depuis un fichier à plat et est utilisé dans le même flux que dans celui où il a été créé, il conserve les métadonnées géospatiales et aucune étape supplémentaire de configuration n'est nécessaire.

Toutefois, si le noeud est exporté et utilisé dans un flux différent, les données de liste géospatiales sont converties automatiquement au format chaîne ; vous devez suivre des étapes supplémentaires pour restaurer le type de stockage de liste et les métadonnées géospatiales associées.

Pour plus d'informations sur les listes, voir «Stockage de liste et niveaux de mesure associés», à la page 12.

Pour plus d'informations sur les détails que vous pouvez définir en tant que métadonnées géospatiales, voir «Sous-niveaux de mesure géospatiaux», à la page 147.

Pour configurer les métadonnées géospatiales, procédez comme suit.

1. Dans l'onglet Fichier du noeud Délimité, sélectionnez la case à cocher **Traiter les crochets comme des listes**. Si vous sélectionnez cette case à cocher, les données incluses entre les crochets ouvrant et fermant sont traitées comme une valeur unique, même si le contenu inclut des caractères délimiteurs comme des virgules et des guillemets. Si vous ne sélectionnez pas cette case à cocher, vos données sont lues comme un type de stockage de chaînes, les virgules dans le champ sont traitées comme des délimiteurs, et votre structure de données n'est pas interprétée correctement.
2. Si vos données incluent des apostrophes ou des guillemets, sélectionnez l'option **Apparier et supprimer** dans les champs **Guillemets simples** et **Guillemets doubles**, comme approprié.
3. Dans l'onglet Données du noeud Délimité, pour les champs de données géospatiales, sélectionnez la case à cocher **Remplacer** et changez le type **Stockage** de chaîne en liste.
4. Par défaut, le type **Stockage** de liste a pour valeur *Liste de nombres réels* et le type de stockage de valeur sous-jacent de la zone de liste est *Réel*. Pour changer le type de stockage de valeur sous-jacent ou la profondeur, cliquez sur **Spécifier** pour afficher la sous-boîte de dialogue Stockage.
5. Dans la sous-boîte de dialogue Stockage, vous pouvez modifier les paramètres suivants :
   - **Stockage** Spécifiez le type de stockage général du champ de données. Par défaut, le type de stockage a pour valeur Liste ; toutefois, la liste déroulante contient tous les autres types de stockage

(Chaîne, Entier, Réel, Date, Heure et Horodatage). Si vous sélectionnez un type de stockage autre que Liste, les options **Stockage de valeur** et **Profondeur** ne sont pas disponibles.

- **Stockage de valeur** Spécifiez les types de stockage des éléments de la liste, par opposition au champ entier. Lorsque vous importez des champs géospatiaux, les seuls types de stockage pertinents sont Réel et Entier ; le paramètre par défaut est Réel.
- **Profondeur** Spécifiez la profondeur de la zone de liste. Elle dépend du type de champ géospatial et respecte les critères suivants :
  - Point – 0
  - Chaîne – 1
  - Polygone – 1
  - Multipoint – 1
  - Multichaîne – 2
  - Multipolygone – 2

  Remarque : Vous devez connaître le type de champ géospatial que vous convertissez en liste et la profondeur pour ce type de champ. Si ces informations ne sont pas définies correctement, le champ ne peut pas être utilisé.

6. Dans l'onglet Types du noeud Délimité, pour le champ de données géospatiales, assurez-vous que la cellule **Mesure** contient le niveau de mesure correct. Pour changer le niveau de mesure, dans la cellule **Mesure**, cliquez sur **Spécifier** pour afficher la boîte de dialogue Valeurs.
7. Dans la boîte de dialogue Valeurs, les champs **Mesure**, **Stockage** et **Profondeur** sont affichés pour la liste. Sélectionnez l'option **Indiquer les valeurs et libellés** et dans la liste déroulante **Type**, sélectionnez le type correct pour **Mesure**. Selon la valeur sélectionnée pour **Type**, vous pouvez être invité à entrer d'autres détails, par exemple indiquer si les données représentent deux ou trois dimensions et le système de coordonnées utilisé.

Noeud Fixe

Vous pouvez utiliser des noeuds Fixe pour importer les données de fichiers texte de longueur fixe (fichiers dont les champs ne sont pas délimités, mais qui commencent au même endroit et sont de longueur fixe). Les données générées automatiquement ou héritées sont souvent stockées au format de longueur fixe. Grâce à l'onglet Fichier du noeud Fixe, vous pouvez facilement indiquer la position et la longueur des colonnes de vos données.

Définition des options du noeud Fixe

L'onglet Fichier du noeud Fixe vous permet d'importer des données dans IBM SPSS Modeler et de spécifier la position des colonnes et la longueur des enregistrements. Vous pouvez cliquer dans le panneau d'aperçu des données situé au centre de la boîte de dialogue pour ajouter des flèches indiquant les points d'arrêt entre les champs.

Fichier. Indiquez le nom du fichier. Vous pouvez entrer un nom de fichier ou cliquer sur le bouton représentant des points de suspension (**...**) pour sélectionner un fichier. Une fois que vous avez sélectionné un fichier, son chemin d'accès apparaît. Son contenu est affiché avec des séparateurs dans le panneau situé en dessous.

Vous pouvez utiliser le panneau d'aperçu des données pour spécifier la position et la longueur des colonnes. La règle située en haut de la fenêtre d'aperçu vous permet de mesurer la longueur des variables et de spécifier le point d'arrêt entre elles. Vous pouvez spécifier des lignes de points d'arrêt en cliquant dans la zone de la règle au-dessus des champs. Pour déplacer les points d'arrêt, faites-les glisser. Pour les supprimer, faites-les glisser hors de la zone d'aperçu des données.

- Chaque ligne de points d'arrêt ajoute automatiquement un nouveau champ à ceux du tableau situé en dessous.

- Les positions de départ indiquées par les flèches sont automatiquement ajoutées à la colonne Démarrer du tableau situé en dessous.

Orientée vers la ligne. Indiquez si vous souhaitez ignorer le caractère de nouvelle ligne à la fin de chaque enregistrement.

Ignorer les lignes des en-têtes. Indiquez le nombre de lignes à ignorer au début du premier enregistrement. Cette option est utile si vous souhaitez ignorer les en-têtes de colonne.

Longueur de l'enregistrement. Indiquez le nombre de caractères dans chaque enregistrement.

Champ. Tous les champs que vous avez définis pour ce fichier de données sont répertoriés ici. Vous pouvez définir des champs de deux manières :

- Spécifier les champs de façon interactive à l'aide du panneau d'aperçu des données situé au-dessus.
- Spécifier les champs manuellement en ajoutant des lignes de champs vides dans le tableau en dessous. Cliquer sur le bouton situé à droite du panneau des champs pour ajouter de nouveaux champs. Entrez ensuite dans le champ vide un nom de champ, une position de départ et une longueur. Ces options ajoutent automatiquement des flèches au panneau d'aperçu des données, que vous pouvez ajuster très facilement.

Pour supprimer un champ défini précédemment, sélectionnez-le dans la liste et cliquez sur le bouton de suppression rouge.

Démarrer. Indiquez la position du premier caractère dans le champ. Par exemple, dans le cas d'un enregistrement dont le second champ commence au seizième caractère, vous devez indiquer la valeur 16.

Longueur. Indiquez le nombre de caractères contenus dans la valeur la plus longue de chaque champ. Ceci permet de déterminer le point de césure pour le champ suivant.

Supprimer les espaces de début et de fin. Cochez cette case pour que les espaces situés en début et en fin des chaînes soient supprimés lors de l'importation.

Remarque. Les comparaisons entre des chaînes qui utilisent ou nom les conversions SQL peuvent générer des ensembles de résultats différents en présence d'espaces situés en fin des chaînes.

Caractères non valides. Sélectionnez **Supprimer** pour supprimer les caractères non valides de l'entrée de données. Sélectionnez **Remplacer par** pour remplacer les caractères non valides par le symbole indiqué (un caractère uniquement). Les caractères non valides sont des caractères nuls (0) ou des caractères qui n'existent pas dans le codage en cours.

Codage. Indique la méthode de codage de texte employée. Vous pouvez choisir la valeur par défaut du système, la valeur par défaut du flux ou UTF-8.

- Si le système est exécuté en mode réparti, sa valeur par défaut est spécifiée dans le Panneau de configuration de Windows de l'ordinateur serveur.
- La valeur par défaut du flux est spécifiée dans la boîte de dialogue Propriétés du flux.

Symbole décimal. Sélectionnez le type de séparateur décimal utilisé dans votre source de données. La **valeur par défaut du flux** est le caractère sélectionné dans l'onglet Options de la boîte de dialogue des propriétés du flux. Sinon, sélectionnez **Point (.)** ou **Virgule (,)** pour lire toutes les données de cette boîte de dialogue à l'aide du caractère choisi comme séparateur décimal.

Reconnaître automatiquement les dates et les heures. Pour permettre à IBM SPSS Modeler d'essayer de reconnaître automatiquement des entrées de date ou d'heures, cochez cette case. Par exemple, cela signifie qu'une entrée telle que 07-11-1965 sera identifiée comme une date et que 02:35:58 sera identifié comme

une heure. Cependant, des entrées telles que 07111965 ou 023558 seront affichées sous la forme d'un entier car il n'y a pas de délimiteurs entre les nombres.

Remarque : Pour éviter de rencontrer des problèmes relatifs aux données lors de l'utilisation de fichiers de données provenant de versions précédentes d'IBM SPSS Modeler, cette case n'est pas cochée par défaut pour les informations enregistrées dans les versions antérieures à la version 13.

Lignes à analyser pour le noeud Typer. Indiquez le nombre de lignes à traiter pour les types de données indiqués.

A tout moment lorsque vous travaillez dans cette boîte de dialogue, cliquez sur **Rafraîchir** pour recharger les champs à partir de la source de données. Ceci est utile lors de la modification des connexions des données au noeud source ou lors de l'utilisation des différents onglets de la boîte de dialogue.

Noeud Statistics

Vous pouvez utiliser le noeud Fichier Statistics pour lire des données directement à partir d'un fichier IBM SPSS Statistics enregistré (*.sav* ou *.zsav*). Ce format remplace le format de fichier cache qui était utilisé dans les versions précédentes d'IBM SPSS Modeler. Si vous souhaitez importer un fichier cache enregistré, vous devez utiliser un noeud IBM SPSS Statistics.

Importer le fichier. Indiquez le nom du fichier. Vous pouvez entrer un nom de fichier ou cliquer sur le bouton représentant des points de suspension (**...**) pour sélectionner un fichier. Le chemin d'accès du fichier apparaît une fois le fichier sélectionné.

Le fichier est chiffré par mot de passe. Cochez cette case si vous savez que le fichier est protégé par mot de passe ; vous êtes ensuite invité à saisir le **Mot de passe**. Si le fichier est protégé par mot de passe et que vous ne le saisissez pas, un message d'avertissement s'affiche lorsque vous tentez d'activer un autre onglet, d'actualiser les données, de prévisualiser le contenu du noeud ou d'exécuter un flux contenant le noeud.

Remarque : Les fichiers protégés par mot de passe ne peuvent être ouverts que par IBM SPSS Modeler version 16 ou supérieure.

Noms des variables. Sélectionnez une méthode de gestion des noms et des libellés de variable lors de l'importation d'un fichier IBM SPSS Statistics *.sav* ou *.zsav*. Les métadonnées que vous incluez sont conservées tout au long de votre travail dans IBM SPSS Modeler. Vous pouvez également les réexporter pour les utiliser dans IBM SPSS Statistics.

- **Lire les noms et les libellés.** Sélectionnez cette option afin de lire les noms et les libellés de variable dans IBM SPSS Modeler. Par défaut, cette option est sélectionnée et les noms de variable affichés dans le noeud Typer. Les libellés peuvent apparaître dans les graphiques, les navigateurs de modèle et d'autres types de sortie, selon les options spécifiées dans la boîte de dialogue des propriétés du flux. Par défaut, l'affichage de libellés dans la sortie est désactivé.
- **Lire les libellés sous forme de nom.** Sélectionnez cette option pour lire les libellés de variable descriptifs du fichier IBM SPSS Statistics *.sav* ou *.zsav* au lieu des noms de champ abrégés, puis utilisez ces libellés en tant que noms de variable dans IBM SPSS Modeler.

Valeurs. Sélectionnez une méthode de gestion des noms et des libellés lors de l'importation d'un fichier IBM SPSS Statistics *.sav* ou *.zsav*. Les métadonnées que vous incluez sont conservées tout au long de votre travail dans IBM SPSS Modeler. Vous pouvez également les réexporter pour les utiliser dans IBM SPSS Statistics.

- **Lire les données et les libellés.** Choisissez cette option pour les valeurs réelles et les libellés de valeur dans IBM SPSS Modeler. Par défaut, cette option est sélectionnée et les valeurs proprement dites

apparaissent dans le noeud Typer. Les libellés de valeur peuvent être affichées dans le Générateur de formules, les navigateurs de modèle et d'autres types de sortie, selon les options spécifiées dans la boîte de dialogue des propriétés du flux.

- **Lire les libellés sous forme de données.** Choisissez cette option si vous préférez utiliser les libellés de valeurs du fichier *.sav* ou *.zsav* plutôt que les codes numériques ou symboliques utilisés pour représenter les valeurs. Par exemple, si vous sélectionnez cette option pour les données dont le champ indiquant le genre a pour valeur 1 et 2 (représentant respectivement *masculin* et *féminin*), le champ sera converti en chaîne, et importera *masculin* et *féminin* comme valeurs réelles.

  Il est important de prendre en compte les valeurs manquantes dans vos données IBM SPSS Statistics avant de choisir cette option. Par exemple, si un champ numérique utilise des libellés uniquement pour les valeurs manquantes (0 = *Pas de réponse*, 99 = *Inconnu*) et que vous sélectionnez l'option ci-dessus, seules les libellés de valeurs *Pas de réponse* et *Inconnu* sont importées, et le champ est converti en chaîne. Dans ce cas, vous devez importer les valeurs elles-mêmes et définir les valeurs manquantes dans un noeud Typer.

Utilisez les informations de formats de champ pour déterminer le stockage. Si cette case est désélectionnée, les valeurs de champs formatées dans le fichier *.sav* en tant qu'entiers (c'est à dire des champs spécifiés comme F*n*.0 dans la Vue des variables d'IBM SPSS Statistics) sont importées à l'aide du stockage d'entier. Toutes les autres valeurs de champ, à l'exception des chaînes, sont importées en tant que nombres réels.

Si cette case est cochée (par défaut), toutes les valeurs de champ, à l'exception des chaînes, sont importées en tant que nombres réels, qu'elles soient formatées dans le fichier *.sav* sous la forme d'entiers ou non.

Ensembles de réponses multiples. Tout ensemble de réponses multiples défini dans le fichier IBM SPSS Statistics est automatiquement conservé lors de l'importation du fichier. Vous pouvez afficher et modifier les ensembles de réponses multiples dans n'importe quel noeud avec un onglet Filtrer. Pour plus d'informations, voir «Modification des ensembles de réponses multiples», à la page 161.

Data Collection Noeud

Les noeuds source Data Collection importent les données d'enquête en fonction du kit Survey Reporter Developer Kit fourni avec votre produit Data Collection. Ce format fait la distinction entre les *données d'observation* (réponses réelles fournies à des questions et collectées au cours d'une enquête) et les *métadonnées* qui décrivent le mode de collecte et d'organisation des données d'observation. Les métadonnées consistent en des informations diverses : texte des questions, nom et description de variables, définitions de variables à réponses multiples, traduction de chaînes de texte et définition de la structure des données d'observation.

Remarque : Ce noeud requiert Survey Reporter Developer Kit, fourni avec votre produit Data Collection. Mise à part l'installation de Developer Kit, aucune configuration supplémentaire n'est requise.

Commentaires

- Les données d'enquête sont lues à partir du format VDATA sous forme de tableau uniforme, ou à partir de sources au format HDATA hiérarchique; si elles incluent une source de métadonnées.
- Les types sont instanciés automatiquement à partir des informations contenues dans les métadonnées.
- Lorsque des données d'enquête sont importées dans SPSS Modeler, les questions sont affichées sous forme de champs et comportent un enregistrement par personne sondée.

Options de fichier d'importation Data Collection

Dans l'onglet Fichier du noeud Data Collection, vous pouvez définir des options pour les métadonnées et les données d'observation à importer.

Paramètres des métadonnées

Remarque : Pour visualiser la liste complète des types de fichier des fournisseurs disponibles, vous devez installer Survey Reporter Developer Kit, disponible avec le logiciel Data Collection.

Fournisseur de métadonnées. Les données d'enquête peuvent être importées dans l'un des divers formats pris en charge par votre kit Data Collection Survey Reporter Developer Kit. Les principaux types de fournisseur disponibles sont les suivants :

- **DataCollectionMDD.** Lit les métadonnées à partir d'un fichier de définition de questionnaire (*.mdd*). Il s'agit du format de modèle de données Data Collection standard.
- **Base de données ADO.** Lit les données d'observation et les métadonnées à partir de fichiers ADO. Indiquez le nom et l'emplacement du fichier *.adoinfo* contenant les métadonnées. Le nom interne de ce fichier DSC est *mrADODsc*.
- **Base de données In2data.** Lit les données d'observation et les métadonnées In2data. Le nom interne de ce fichier DSC est *mrI2dDsc*.
- **Fichier journal de collecte des données.** Lit les métadonnées issues d'un fichier journal Data Collection standard. Les fichiers journaux sont généralement dotés d'une extension *.tmp*. Toutefois, certains fichiers peuvent comporter une autre extension. Si nécessaire, vous pouvez renommer le fichier et lui attribuer l'extension *.tmp*. Le nom interne de ce fichier DSC est *mrLogDsc*.
- **Fichier de définitions Quancept.** Convertit les métadonnées en script Quancept. Indiquez le nom du fichier Quancept *.qdi*. Le nom interne de ce fichier DSC est *mrQdiDrsDsc*.
- **Base de données Quanvert.** Lit les données d'observation et les métadonnées Quanvert. Précisez le nom et l'emplacement du fichier *.qvinfo* ou *.pkd*. Le nom interne de ce fichier DSC est *mrQvDsc*.
- **Base de données de participation à la collecte de données.** Lit les tables exemple et les tables d'historique d'un projet et crée des variables catégorielles calculées correspondant aux colonnes de ces tables. Le nom interne de ce fichier DSC est *mrSampleReportingMDSC*.
- **Fichier Statistics.** Lit les données d'observation et les métadonnées issues d'un fichier IBM SPSS Statistics *.sav*. Ecrit les données d'observation dans un fichier IBM SPSS Statistics *.sav* en vue de leur analyse dans IBM SPSS Statistics. Ecrit les métadonnées provenant d'un fichier IBM SPSS Statistics *.sav* dans un fichier *.mdd*. Le nom interne de ce fichier DSC est *mrSavDsc*.
- **Fichier Surveycraft.** Lit les données d'observation et les métadonnées SurveyCraft. Indiquez le nom du fichier *.vq* SurveyCraft. Le nom interne de ce fichier DSC est *mrSCDsc*.
- **Fichier de script de collecte des données.** Lit les métadonnées contenues dans un fichier mrScriptMetadata. Ces fichiers sont généralement dotés de l'extension *.mdd* ou *.dms*. Le nom interne de ce fichier DSC est *mrScriptMDSC*.
- **Fichier XML Triple-S.** Lit les métadonnées à partir d'un fichier Triple-S au format XML. Le nom interne de ce fichier DSC est *mrTripleSDsc*.

Propriétés des métadonnées. Vous pouvez également sélectionner **Propriétés** pour préciser la version de l'enquête à importer, ainsi que la langue, le contexte et le type de libellé à utiliser. Pour plus d'informations, voir «Propriétés relatives aux métadonnées d'importation Data Collection», à la page 37.

Paramètres des données d'observation

Remarque : Pour visualiser la liste complète des types de fichier des fournisseurs disponibles, vous devez installer Survey Reporter Developer Kit, disponible avec le logiciel Data Collection.

Obtenir les paramètres des données d'observation. Lorsque le système ne lit que des métadonnées issues de fichiers *.mdd*, cliquez sur **Obtenir les paramètres des données d'observation** pour déterminer les sources de données d'observation associées aux métadonnées sélectionnées, ainsi que les paramètres nécessaires pour accéder à une source donnée. Cette option est disponible uniquement pour les fichiers *.mdd*.

Fournisseur de données d'observation. Les types de fournisseur suivants sont pris en charge :

- **Base de données ADO.** Lit les données d'observation par le biais de l'interface Microsoft ADO. Sélectionnez `UDL OLE-DB` comme type de donnée d'observation et indiquez une chaîne de connexion dans le champ UDL de données d'observation. Pour plus d'informations, voir «Chaîne de connexion de base de données», à la page 38. Le nom interne de ce fichier component est *mrADODsc*.
- **Fichier texte délimité (Excel).** Lit des données d'observation à partir d'un fichier délimité par des virgules (.CSV), tel qu'il peut être sorti par Excel. Le nom interne est *mrCsvDsc*.
- **Fichier de données de collecte des données.** Lit les données d'observation à partir d'un fichier de données Data Collection natif. Le nom interne est *mrDataFileDsc*.
- **Base de données In2data.** Lit les données d'observation et les métadonnées à partir d'un fichier de base de données In2data (*.i2d*). Le nom interne correspondant est *mrI2dDsc*.
- **Fichier journal de collecte des données.** Lit les données d'observation à partir d'un fichier journal Data Collection standard. Les fichiers journaux sont généralement dotés d'une extension *.tmp*. Toutefois, certains fichiers peuvent comporter une autre extension. Si nécessaire, vous pouvez renommer le fichier et lui attribuer l'extension *.tmp*. Le nom interne correspondant est *mrLogDsc*.
- **Fichier de données Quantum.** Lit les données d'observation provenant d'un fichier ASCII Quantum (*.dat*). Le nom interne correspondant est *mrPunchDsc*.
- **Fichier de données Quancept.** Lit les données d'observation issues d'un fichier *.drs*, *.drz* ou *.dru* Quancept. Le nom interne correspondante est *mrQdiDrsDsc*.
- **Base de données Quanvert.** Lit les données d'observation à partir d'un fichier *qvinfo* ou *.pkd*Quanvert. Le nom interne correspondant est *mrQvDsc*.
- **Base de données de collecte des données (MS SQL Server).** Lit les données d'observation pour une base de données relationnelle Microsoft SQL Server. Pour plus d'informations, voir «Chaîne de connexion de base de données», à la page 38. Le nom interne est *mrRdbDsc2*.
- **Fichier Statistics.** Lit les données d'observation et les métadonnées issues d'un fichier IBM SPSS Statistics *.sav*. Le nom interne correspondant est *mrSavDsc*.
- **Fichier SurveyCraft.** Lit les données d'observation provenant d'un fichier *.qdt* SurveyCraft. Les fichiers *.vq* et *.qdt* doivent se trouver dans le même répertoire et être accessibles en lecture et en écriture. Cela n'est pas le cas lorsque ces fichiers sont créés par défaut à l'aide de SurveyCraft ; par conséquent, vous devez déplacer l'un des fichiers pour pouvoir importer ensuite des données SurveyCraft. Le nom interne correspondant est *mrScDsc*.
- **Fichier de données Triple-S.** Lit les données d'observation à partir d'un fichier de données Triple-S, au format délimité par des virgules ou de longueur fixe. Le nom interne est *mr TripleDsc*.
- **Collecte de données XML.** Lit les données d'observation issues d'un fichier de données XML Data Collection. Vous pouvez généralement utiliser ce format pour transférer des données d'observation d'un emplacement vers un autre. Le nom interne correspondant est *mrXmlDsc*.

Type de données d'observation. Indique si les données d'observation sont lues à partir d'un fichier, d'un dossier, ou d'un type UDL OLE-DB ou DSN ODBC, et met à jour en conséquence les options de la boîte de dialogue. Les options valides dépendent du type de fournisseur. Pour les fournisseurs de base de données, vous pouvez définir les options de connexion OLE-DB ou ODBC. Pour plus d'informations, voir «Chaîne de connexion de base de données», à la page 38.

Projet de données d'observation. Lorsque vous lisez des données d'observation provenant d'une base de données Data Collection, vous pouvez fournir le nom du projet. Pour tous les autres types de données d'observation, ce paramètre doit rester vide.

Importation de variable

Importation de variables système. Indique si les variables système sont importées, y compris les variables qui indiquent l'état de l'entretien (en cours, terminé, date de fin, etc). Vous pouvez choisir **Aucune**, **Toutes** ou **Communes**.

Importation de variables «Codes». Contrôle l'importation de variables qui représentent des codes utilisés pour des réponses «Autre» ouvertes pour des variables catégorielles.

Importation de variables «SourceFile». Contrôle l'importation de variables qui contiennent les noms de fichiers d'images de réponses numérisées.

Importer des variables à réponses multiples sous la forme de. Des variables à réponses multiples peuvent être importées sous la forme de plusieurs champs booléens (un ensemble de dichotomies multiples), qui est la méthode par défaut pour les nouveaux flux. Les flux créés dans les versions d'IBM SPSS Modeler antérieures à 12.0, des réponses multiples importées dans un champ unique, avec des valeurs séparées par des virgules. L'ancienne méthode reste prise en charge pour permettre l'exécution des flux existants. Toutefois, la mise à jour des anciens flux pour utiliser la nouvelle méthode est recommandée. Pour plus d'informations, voir «Importation des ensembles de réponses multiples», à la page 38.

Propriétés relatives aux métadonnées d'importation Data Collection

Lorsque vous importez des données d'enquête Data Collection, dans la boîte de dialogue Propriétés des métadonnées, vous pouvez préciser la version de l'enquête à importer, ainsi que la langue, le contexte et le type de libellé à utiliser. Vous ne pouvez importer qu'une langue, un contexte et un type de libellé à la fois.

Version - Chaque version d'enquête peut être considérée comme un instantané des métadonnées utilisées pour collecter un ensemble précis de données d'observation. Au fil des changements apportés à un questionnaire, plusieurs versions différentes peuvent être créées. Vous pouvez importer la dernière version, toutes les versions ou une version particulière.

- **Toutes.** Sélectionnez cette option si vous souhaitez utiliser une combinaison (ou sur-ensemble) de toutes les versions disponibles. (C'est ce que l'on nomme parfois supervision). En cas de conflit entre versions, les versions les plus récentes sont généralement prioritaires sur les versions plus anciennes. Ainsi, si un libellé de catégorie diffère dans l'une des versions, c'est le texte de la version la plus récente qui est utilisé.
- **Version la plus récente.** Sélectionnez cette option pour n'utiliser que la version la plus récente.
- **Spécifier une version.** Sélectionnez cette option si vous souhaitez utiliser une version d'enquête particulière.

Sélectionner l'ensemble des versions s'avère utile, par exemple, lorsque vous souhaitez exporter des données d'observation à partir de plusieurs versions et que des changements ont été apportés aux définitions des variables et des catégories (autrement dit, lorsque les données d'observation collectées dans une version ne sont pas valides dans une autre version). En sélectionnant toutes les versions pour lesquelles effectuer une exportation des données d'observation, vous pouvez généralement exporter simultanément les données souhaitées collectées dans les différentes versions, sans rencontrer d'erreurs de validité dues à des différences entre versions. Toutefois, selon les changements apportés aux versions, certaines erreurs de validité risquent tout de même de se produire.

Langue - Les questions et le texte associé peuvent être stockés en plusieurs langues dans les métadonnées. Vous pouvez utiliser la langue par défaut de l'enquête ou indiquer une langue particulière. Si un élément n'est pas disponible dans la langue demandée, la langue par défaut est alors utilisée.

Contexte. Sélectionnez le contexte utilisateur souhaité. Il permet de déterminer les textes à afficher. Par exemple, sélectionnez **Question** pour afficher le texte des questions ou **Analyse** pour afficher une version abrégée des textes, mieux adaptée lors de l'analyse des données.

Type de libellé. Répertorie les types de libellé qui ont été définis. La valeur par défaut, **Libellé**, est utilisée pour le texte des questions dans le contexte utilisateur Question et la description des variables dans le contexte utilisateur Analyse. Vous pouvez définir d'autres types de libellé pour les instructions, les descriptions, etc.

Chaîne de connexion de base de données

Lorsque vous utilisez le noeud Data Collection pour importer des données d'observation à partir d'une base de données via une connexion OLE-DB ou ODBC, sélectionnez **Edition** dans l'onglet Fichier pour accéder à la boîte de dialogue de la chaîne de connexion et personnalisez la chaîne de connexion transmise au fournisseur afin d'affiner la connexion.

Propriétés avancées

Lorsque vous utilisez le noeud Data Collection afin d'importer des données d'observation à partir d'une base de données qui requiert une connexion explicite, sélectionnez **Options avancées** pour fournir un ID utilisateur et un mot de passe permettant d'accéder à la source de données.

Importation des ensembles de réponses multiples

Les variables à réponses multiples peuvent être importées à partir de Data Collection comme ensembles de plusieurs dichotomies, avec un champ booléen distinct pour chaque valeur possible de variable. Par exemple, s'il est demandé aux personnes interrogées de sélectionner les musées qu'ils ont visités dans une liste, l'ensemble inclut un champ booléen distinct pour chaque musée répertorié.

Une fois les données importées, vous pouvez ajouter ou modifier des ensembles de réponses multiples à partir d'un noeud quelconque qui inclut un onglet Filtrer. Pour plus d'informations, voir «Modification des ensembles de réponses multiples», à la page 161.

Importation de réponses multiples dans un champ unique (pour les flux créés dans des versions précédentes)

Dans les versions antérieures de SPSS Modeler, au lieu d'importer des réponses multiples comme décrit ci-dessus, elles ont été importées dans un champ unique, avec des valeurs séparées par des virgules. Cette méthode reste acceptée afin de prendre en charge les flux existants. Toutefois, il est recommandé de mettre à jour ces flux pour utiliser la nouvelle méthode.

Remarques sur l'importation de colonnes Data Collection

Les colonnes issues des données Data Collection sont lues dans SPSS Modeler de la manière indiquée dans le tableau suivant.

Tableau 4. Récapitulatif d'importation de la colonne Data Collection

Data Collection Type de colonne	Stockage de SPSS Modeler	Niveau de mesure
Commutateur booléen (yes/no (oui/non))	Chaîne	Commutateur (valeurs 0 et 1)
Catégorique	Chaîne	Nominal
Date ou horodatage	Horodatage	Continu
Double (valeur à virgule flottante comprise dans un intervalle défini)	Réel	Continu
Long (valeur entière comprise dans un intervalle défini)	Entier	Continu
Texte (description libre)	Chaîne	Sans type
Niveau (indique des grilles ou des boucles dans une question)	Ne s'applique pas aux données VDATA et n'est pas importé dans SPSS Modeler	
Objet (données binaires, comme la télécopie d'un texte griffonné ou un enregistrement sonore)	Pas importé dans SPSS Modeler	
Aucun (type inconnu)	Pas importé dans SPSS Modeler	

Tableau 4. Récapitulatif d'importation de la colonne Data Collection (suite)

Data Collection Type de colonne	Stockage de SPSS Modeler	Niveau de mesure
Colonne Respondent.Serial (associe un ID unique à chaque personne sondée)	Entier	Sans type

Pour éviter toute incohérence éventuelle entre les libellés de valeur lues dans les métadonnées et les valeurs réelles, toutes les valeurs des métadonnées sont converties en minuscules. Par exemple, le libellé de valeur *E1720_ans* est convertie en *e1720_ans*.

noeud source IBM Cognos

Le noeud source IBM Cognos permet d'inclure des données de base de données Cognos ou des rapports de liste unique dans votre session d'exploration de données. Ainsi, vous pouvez combiner les capacités de veille économique de Cognos aux capacités d'analyses prédictives d'IBM SPSS Modeler. Vous pouvez importer des données relationnelles, DMR (dimensionally-modeled relational) et OLAP.

A partir d'une connexion au serveur Cognos, commencez par sélectionner un emplacement à partir duquel importer les données ou les rapports. Un emplacement contient un modèle Cognos et tous les dossiers, requêtes, rapports, vues, raccourcis, URL et définitions de travaux associés à ce modèle. Un modèle Cognos définit les règles métier, les descriptions de données, les relations entre les données, les dimensions et les hiérarchies métier et d'autres tâches administratives.

Si vous importez des données, vous sélectionnez ensuite les objets que vous voulez importer depuis le package sélectionné. Parmi les objets que vous pouvez importer : les objets de requête (qui représentent les tables de la base de données) ou les éléments de requête individuels (qui représentent les colonnes de la table). Voir «Icônes d'objet Cognos» pour plus d'informations.

Si des filtres sont définis dans le package, vous pouvez importer un ou plusieurs d'entre eux. Si un des filtres que vous importez est associée à des données importées, ce filtre est appliqué avant l'importation des données. Les données à importer doivent être au format UTF-8.

Si vous importez un rapport, vous sélectionnez un package, ou un dossier dans un package, contenant un ou plusieurs rapports. Vous sélectionnez ensuite le rapport individuel que vous voulez importer. Seuls les rapports de liste unique peuvent être importés, les listes multiples ne sont pas prises en charge.

Si les paramètres ont été définis, soit pour un objet de données soit pour un rapport, vous pouvez spécifier les valeurs de ces paramètres avant d'importer l'objet ou le rapport.

Remarque : Le noeud source Cognos ne prend en charge que les packages Cognos CQM. Les packages DQM ne sont pas pris en charge.

Icônes d'objet Cognos

Les divers types d'objets pouvant être importés depuis une base de données Cognos Analytics sont représentés par différentes icônes, comme l'illustre le tableau suivant.

Tableau 5. Icônes d'objet Cognos.

Icône	Objet
	Package

Tableau 5. Icônes d'objet Cognos (suite).

Icône	Objet
	Espace de nommage
	Objet de requête
	Elément de requête
	Dimension de mesure
	Mesure
	Dimension
	Hiérarchie de niveau
	Niveau
	Filtrer
	Rapport
	Calcul autonome

Importation des données Cognos

Pour importer des données d'une base de données IBM Cognos Analytics (version 11 ou ultérieure prise en charge), sous l'onglet Données de la boîte de dialogue IBM Cognos, assurez-vous que le **Mode** est défini sur **Données**.

Connexion. Cliquez sur **Modifier** pour afficher une boîte de dialogue dans laquelle vous pourrez définir les détails d'une connexion Cognos à partir de laquelle importer les données ou les rapports. Si vous êtes déjà connecté à un serveur Cognos via IBM SPSS Modeler, vous pouvez également modifier les détails de la connexion actuelle. Pour plus d'informations, voir «Connexions Cognos», à la page 42.

Emplacement. Lorsque la connexion au serveur Cognos est établie, cliquez sur **Modifier** à côté de ce champ pour afficher une liste des packages disponibles depuis lesquels importer le contenu. Consultez «Sélection de l'emplacement de Cognos», à la page 42 pour plus d'informations.

Contenu. Affiche le nom du package sélectionné, avec les espaces de nommage associés au package. Double-cliquez sur un espace de nommage pour afficher les objets que vous pouvez importer. Les divers types d'objets sont représentés par différentes icônes. Voir «Icônes d'objet Cognos», à la page 39 pour plus d'informations.

Pour choisir un objet à importer, sélectionnez l'objet et cliquez sur la flèche supérieure des deux flèches droites pour déplacer l'objet dans le volet **Champs à importer**. La sélection d'un objet de requête importe tous ses éléments de requête. Double-cliquer sur un objet de requête le développe et vous pouvez ainsi choisir un ou plusieurs éléments de requête individuels. Vous pouvez effectuer plusieurs sélections avec Ctrl-clic (sélectionne des éléments individuels), Maj-clic (sélectionne un block d'éléments) et Ctrl-A (sélectionne tous les éléments).

Pour choisir un filtre à appliquer (si des filtres sont définis pour le package), accédez au filtre dans le volet Contenu, sélectionnez le filtre et cliquez sur la flèche inférieure des deux flèches de droite pour déplacer le filtre dans le volet **Filtres à appliquer**. Vous pouvez faire plusieurs choix avec Ctrl-clic (sélectionner des filtres individuels) et Shift-clic (sélectionner un bloc de filtres).

Champs à importer. Répertorie les objets de la base de données que vous avez choisi d'importer dans IBM SPSS Modeler pour être traités. Si un des objets spécifiques n'est plus requis, sélectionnez-le et cliquez sur la flèche vers la gauche pour le déplacer de nouveau vers le volet **Contenu**. Vous pouvez effectuer plusieurs sélections de la même façon que pour le **Contenu**.

Filtres à appliquer. Répertorie les filtres que vous avez choisi d'appliquer aux données avant de les importer. Si l'un des filtres spécifiques n'est plus requis, sélectionnez-le et cliquez sur la flèche vers la gauche pour le déplacer de nouveau vers le volet **Contenu**. Vous pouvez effectuer plusieurs sélections de la même façon que pour le **Contenu**.

Paramètres. Si ce bouton est activé, l'objet sélectionné a des paramètres définis. Vous pouvez utiliser des paramètres afin de procéder à des ajustements (par exemple, effectuer un calcul paramétré) avant l'importation des données. Si des paramètres sont définis mais qu'aucune valeur par défaut n'existe, le bouton affiche un triangle d'avertissement. Cliquez sur le bouton pour afficher les paramètres et les modifier le cas échéant. Si le bouton est désactivé, aucun paramètre n'est défini pour le rapport.

Agréger les données avant l'importation. Cochez cette case si vous souhaitez importer des données agrégées plutôt que des données brutes.

Importer des rapports Cognos

Pour importer un rapport prédéfini d'une base de données Cognos IBM, sous l'onglet Données de la boîte de dialogue IBM Cognos, assurez-vous que le **Mode** est défini sur **Données**. Seuls les rapports de liste unique peuvent être importés, les listes multiples ne sont pas prises en charge.

Connexion. Cliquez sur **Modifier** pour afficher une boîte de dialogue dans laquelle vous pourrez définir les détails d'une connexion Cognos à partir de laquelle importer les données ou les rapports. Si vous êtes déjà connecté à un serveur Cognos via IBM SPSS Modeler, vous pouvez également modifier les détails de la connexion actuelle. Pour plus d'informations, voir «Connexions Cognos», à la page 42.

Emplacement. Lorsque la connexion au serveur Cognos est établie, cliquez sur **Modifier** à côté de ce champ pour afficher une liste des packages disponibles depuis lesquels importer le contenu. Consultez «Sélection de l'emplacement de Cognos», à la page 42 pour plus d'informations.

Contenu. Affiche le nom du package ou dossier sélectionné contenant les rapports. Accédez à un rapport spécifique, sélectionnez-le et cliquez sur la flèche droite pour amener le rapport dans le champ **Rapport à importer**.

Rapport à importer. Indique le rapport que vous avez choisi d'importer dans IBM SPSS Modeler. Si vous n'avez plus besoin de ce rapport, sélectionnez-le et cliquez sur la flèche vers la gauche pour le déplacer de nouveau vers le volet **Contenu** ou insérer un autre rapport dans ce champ.

Paramètres. Si ce bouton est activé, les paramètres du rapport sélectionné sont définis. Vous pouvez utiliser les paramètres pour effectuer des réglages avant d'importer le rapport (par exemple, spécifier une date de début et de fin pour les données de rapport). Si des paramètres sont définis mais qu'aucune valeur par défaut n'existe, le bouton affiche un triangle d'avertissement. Cliquez sur le bouton pour afficher les paramètres et les modifier le cas échéant. Si le bouton est désactivé, aucun paramètre n'est défini pour le rapport.

Connexions Cognos

La boîte de dialogue Connexions Cognos vous permet de sélectionner le serveur Cognos Analytics (version 11 ou ultérieure prise en charge) à partir duquel vous souhaitez importer ou exporter les objets de base de données.

URL du serveur Cognos Entrez l'URL du serveur Cognos Analytics depuis lequel vous souhaitez importer ou exporter. Ceci est la valeur de la propriété d'environnement "URI de répartiteur externe" de la configuration Cognos IBM sur le serveur Cognos. Contactez votre administrateur système Cognos si vous ne savez pas quel URL choisir.

Mode Sélectionnez **Définir les données d'identification** si vous souhaitez vous connecter avec un espace de noms, un nom d'utilisateur et un mot de passe Cognos spécifique (par exemple en tant qu'administrateur). Sélectionnez **Utilisez une connexion anonyme** pour vous connecter sans identifiant. Dans ce cas, vous n'avez pas besoin de remplir les autres champs.

Sinon, si vous avez des données d'identification IBM Cognos stockée dans le référentiel IBM SPSS Collaboration and Deployment Services, vous pouvez les utiliser au lieu d'entrer le nom d'utilisateur et le mot de passe, ou de créer une connexion anonyme. Pour utiliser des données d'identification existantes, sélectionnez **Données d'identification stockées** et entrez le **Nom des données d'identification** ou recherchez-le.

L'espace de nom Cognos est modélisé par un domaine dans IBM SPSS Collaboration and Deployment Services.

ID d'espace de nom Indiquez le fournisseur de sécurité de l'authentification Cognos utilisé pour se connecter au serveur. Le fournisseur d'authentification sert à définir et à gérer les utilisateurs, groupes et rôles et à contrôler le processus d'authentification. Notez qu'il s'agit de l'ID d'espace de nom et non du nom d'espace de nom (l'ID n'est pas toujours identique au nom).

Nom d'utilisateur Entrez le nom d'utilisateur Cognos avec lequel effectuer la connexion au serveur.

Mot de passe Entrez le mot de passe associé au nom d'utilisateur défini.

Enregistrer par défaut Cliquez sur ce bouton pour stocker ces paramètres comme paramètres par défaut et ainsi éviter d'avoir à les saisir à chaque fois que vous ouvrez le noeud.

Sélection de l'emplacement de Cognos

La boîte de dialogue Spécifier l'emplacement vous permet de sélectionner un package Cognos à partir duquel importer des données, ou un package ou dossier à partir duquel importer des rapports.

Dossiers publics. Si vous importez des données, ceci répertorie les packages et dossiers disponibles sur le serveur choisi. Sélectionnez le package désiré et cliquez sur **OK**. Vous ne pouvez sélectionner qu'un seul package par noeud source Cognos.

Si vous importez des rapports, ceci répertorie les packages et dossiers contenant des rapports disponibles sur le serveur choisi. Sélectionnez un dossier de package ou de rapport et cliquez sur **OK**. Vous ne pouvez sélectionner qu'un seul dossier de package ou de rapport par noeud source Cognos, bien que les dossiers de rapport puissent contenir d'autres dossiers de rapport ainsi que des rapports individuels.

Spécification des paramètres pour les données ou les rapports

Si les paramètres ont été définis dans Cognos Analytics, soit pour un objet de données soit pour un rapport, vous pouvez spécifier les valeurs de ces paramètres avant d'importer l'objet ou le rapport. Par exemple, les paramètres d'un rapport peuvent être les dates de début et de fin relatives au contenu du rapport.

Nom. Le nom du paramètre tel qu'il est spécifié dans la base de données Cognos.

Type. Une description du paramètre.

Valeur. La valeur à affecter au paramètre. Pour entrer ou modifier une valeur, double-cliquez sur la cellule correspondante dans le tableau. Les valeurs ne sont pas validées ici, par conséquent les valeurs non valides sont détectées au moment de l'exécution.

Supprimer automatiquement les paramètres non valides de la table. Cette option est sélectionnée par défaut et supprime tout paramètre non valide détecté dans l'objet de données ou le rapport.

Noeud source IBM Cognos TM1

Le noeud source IBM Cognos TM1 permet d'inclure des données Cognos TM1 dans votre session d'exploration de données. Ainsi, vous pouvez combiner les fonctions de planification d'entreprise de Cognos aux fonctions d'analyse prédictive d'IBM SPSS Modeler. Vous pouvez importer une version à plat des données du cube OLAP multidimensionnel.

Remarque : L'utilisateur TM1 doit disposer des droits suivants : privilège d'écriture de cubes, privilège de lecture des dimensions et privilège d'écriture d'éléments de dimension. En outre, IBM Cognos TM1 10.2, groupe de correctifs 3 ou ultérieur, est requis pour que SPSS Modeler puisse importer et exporter des données Cognos TM1. Les flux existants basés sur des versions antérieures continueront de fonctionner.

Les informations d'identification de l'administrateur ne sont pas requises pour ce noeud. Elles le sont toutefois si vous utilisez toujours l'ancien noeud TM1 existant antérieur à la version 17.1.

Vous devez modifier les données dans TM1 avant de pouvoir les importer.

Remarque : Les données à importer doivent être au format UTF-8.

Depuis une connexion d'un hôte d'administration IBM Cognos TM1, vous sélectionnez d'abord un serveur TM1 à partir duquel importer les données. Un serveur contient un ou plusieurs cubes TM1. Sélectionnez ensuite le cube requis puis, à l'intérieur du cube, les colonnes et lignes à importer.

Remarque : Pour pouvoir utiliser les noeuds source ou d'exportation TM1 dans SPSS Modeler, vous devez vérifier certains paramètres dans le fichier `tm1s.cfg`. Il s'agit du fichier de configuration du serveur TM1 dans le répertoire racine du serveur TM1.

- HTTPPortNumber - définissez un numéro de port valide, généralement 1 à 65535. Notez qu'il ne s'agit pas du numéro de port que vous avez par la suite spécifié dans la connexion dans le noeud, mais d'un port interne utilisé par TM1, qui est désactivé par défaut. Si nécessaire, contactez votre administrateur afin qu'il vous indique le paramètre valide pour ce port.
- UseSSL - si vous le définissez sur *True*, HTTPS est utilisé comme protocole de transport. Dans ce cas, vous devez importer la certification TM1 vers l'environnement d'exécution Java SPSS Modeler Server.

Importation de données IBM Cognos TM1

Pour importer des données depuis une base de données IBM Cognos TM1, dans l'onglet Données de la boîte de dialogue IBM Cognos TM1, sélectionnez l'hôte d'administration TM1 et les détails des données, du cube et du serveur associés.

Remarque : Avant d'importer des données, vous devez exécuter certaines tâches de prétraitement dans TM1 pour vérifier que les données sont dans un format reconnaissable par IBM SPSS Modeler. Cela implique le filtrage des données à l'aide de l'éditeur de sous-ensemble pour afficher la vue dans une taille et une forme adaptées à l'importation.

Les valeurs zéro (0) importées depuis TM1 sont traitées comme des valeurs "null" (TM1 ne distingue pas les blancs des valeurs zéro). Notez également que les données non numériques (ou métadonnées) provenant de *dimensions ordinaires* peuvent être importées dans IBM SPSS Modeler. En revanche, l'importation de *mesures* non numériques n'est pas prise en charge actuellement.

Hôte admin. Entrez l'adresse URL de l'hôte d'administration où le serveur TM1 auquel vous voulez vous connecter est installé. L'hôte d'administration est défini en tant qu'adresse URL unique pour tous les serveurs TM1. A partir de cette adresse URL, tous les serveurs IBM Cognos TM1 installés et s'exécutant dans votre environnement peuvent être reconnus et sont accessibles.

Serveur TM1. Une fois que vous avez établi la connexion à l'hôte d'administration,sélectionnez le serveur qui contient les données à importer et cliquez sur **Connexion**. Si vous ne vous êtes pas déjà connecté à ce serveur, vous êtes invité à entrer votre **Nom d'utilisateur** et votre **Mot de passe**. Sinon, vous pouvez rechercher les détails de connexion déjà entrés que vous avez sauvegardés en tant que **Données d'identification stockées**.

Sélectionnez une vue de cube TM1 à importer. Affiche le nom des cubes du serveur TM1 à partir desquels vous pouvez importer des données. Cliquez deux fois sur un cube pour afficher les données de vue que vous pouvez importer.

Remarque :

Seuls les cubes avec une dimension peuvent être importés dans IBM SPSS Modeler.

Si un alias a été défini pour un élément dans votre cube TM1 (par exemple, si une valeur `23277` dispose d'un alias `Ventes`), la valeur sera importée, mais pas l'alias.

Pour choisir les données à importer, sélectionnez la vue et cliquez sur la flèche droite pour la déplacer vers le volet **Vue à importer**. Si la vue requise n'est pas visible, cliquez deux fois sur un cube pour développer sa liste de vues. Vous pouvez sélectionner une vue publique ou privée.

Dimension(s) de ligne. Répertorie le nom de la dimension de ligne dans les données que vous avez choisi d'importer. Faites défiler la liste des niveaux et sélectionnez celui dont vous avez besoin.

Dimension de colonne. Répertorie le nom de la dimension de colonne dans les données que vous avez choisi d'importer. Faites défiler la liste des niveaux et sélectionnez celui dont vous avez besoin.

Dimension(s) contextuelle(s). Affichage uniquement. Affiche les dimensions contextuelles associées aux lignes et colonnes sélectionnées.

Noeud source TWC

Le noeud source TWC importe les données météorologiques de The Weather Company, une entreprise IBM. Vous pouvez l'utiliser pour obtenir les données météorologiques historiques ou prévisionnelles d'un lieu. Cela peut vous aider à développer de meilleures solutions métier météorologiques pour une prise de décision plus avisée à l'aide des données météorologiques les plus précises disponibles.

Avec ce noeud, vous pouvez entrer des données météorologiques comme les suivantes : `latitude`, `longitude`, `time`, `day_ind` (indique la nuit ou le jour), `temp`, `dewpt` (point de rosée), `rh` (humidité relative), `feels_like` (température), `heat_index`, `wc` (direction du vent), `wx_phrase` (couvert, partiellement couvert, etc), `pressure`, `clds` (nuages), `vis` (visibilité), `wspd` (vitesse du vent), `gust`, `wdir` (direction du vent), `wdir_cardinal` (NO, NNO, N, etc), `uv_index` (index ultraviolet) et `uv_desc` (faibles, élevés, etc).

Le noeud source TWC utilise les API suivantes :

- TWC Historical Observations Airport (http://goo.gl/DplOKj) pour les données historiques de météorologie
- TWC Hourly Forecast (http://goo.gl/IJhhvZ) pour la prévision des données météorologiques

Emplacement

Latitude. Entrez la latitude du lieu dont vous souhaitez obtenir les données météorologiques, au format `[-90.0~90.0]`.

Longitude. Entrez la longitude du lieu dont vous souhaitez obtenir les données météorologiques, au format `[-180.0~180.0]`.

Divers

Clé de licence. Clé de licence requise. Entrez la clé de licence que vous avez obtenue auprès de The Weather Company. Si vous ne possédez pas de clé, contactez votre administrateur ou votre interlocuteur IBM.

Au lieu d'envoyer la clé à tous les utilisateurs, votre administrateur peut l'avoir spécifiée dans un nouveau fichier `config.cfg` sur IBM SPSS Modeler Server, auquel cas vous pouvez laisser cette zone vide. Si la clé est spécifiée dans les deux emplacements, celle de cette boîte de dialogue est prioritaire. Remarque destinée aux administrateurs : pour ajouter la clé de licence sur le serveur, créez un fichier intitulé `config.cfg` avec le contenu `LicenseKey=<CLELICENCE>` (<CLELICENCE> représentant la clé de licence) à l'emplacement `<InstallationModelerServer>\ext\bin\pasw.twcdata`.

Unités. Sélectionnez l'unité de mesure à utiliser : **Anglais**, **Métrique** ou **Hybride**. La valeur par défaut est **Métrique**.

Type de données

Historique. Si vous souhaitez importer les données météorologiques historiques, sélectionnez **Historique**, puis spécifiez une date de début et une date de fin au format `YYYYMMDD` (par exemple, `20120101` pour le 1er janvier 2012).

Prévision. Si vous souhaitez importer les prévisions de données météorologiques, sélectionnez **Prévision**, puis spécifiez les heures de la prévision.

Noeud source SAS

Cette fonction est disponible dans SPSS Modeler Professional et SPSS Modeler Premium.

Le noeud source SAS permet d'inclure des données SAS dans votre session d'exploration de données. Vous pouvez importer quatre types de fichier :

- SAS pour Windows/OS2 (*.sd2*)
- SAS pour UNIX (*.ssd*)
- Fichier Transport SAS (*.tpt*)
- SAS Version 7/8/9 (*.sas7bdat*)

Une fois les données importées, toutes les variables sont conservées et aucun type de variable n'est modifié. Toutes les observations sont sélectionnées.

Définition des options du noeud source SAS

Importer. Sélectionnez le type de fichier SAS à importer. Vous pouvez choisir **SAS pour Windows/OS2 (.sd2)**, **SAS pour UNIX (.SSD)**, **Fichier Transport SAUVEGARDES (.tpt)** ou **SAS Version 7/8/9 (.sas7bdat)**.

Importer le fichier. Indiquez le nom du fichier. Vous pouvez entrer un nom de fichier ou cliquer sur le bouton représentant des points de suspension (**...**) pour accéder à l'emplacement du fichier.

Membre. Choisissez un membre à importer depuis le fichier Transport SAS sélectionné au-dessus. Vous pouvez entrer un nom de membre ou cliquer sur **Sélectionner** pour parcourir tous les membres du fichier.

Lire les formats utilisateur à partir d'un fichier de données SAS. Cochez cette case pour que les formats utilisateur soient lus. Les données et les formats de données (comme les libellés de variables) sont stockés dans différents fichiers. La plupart du temps, il est conseillé d'importer également les formats. Cependant, dans le cas des jeux de données volumineux, il peut être préférable de désélectionner cette option afin d'économiser la mémoire.

Formater le fichier. Si un fichier format est requis, cette zone de texte est active. Vous pouvez entrer un nom de fichier ou cliquer sur le bouton représentant des points de suspension (**...**) pour accéder à l'emplacement du fichier.

Noms des variables. Sélectionnez une méthode de gestion des noms et libellés de variable lors de l'importation des données d'un fichier SAS. Les métadonnées que vous incluez sont conservées tout au long de votre travail dans IBM SPSS Modeler. Vous pouvez également les réexporter pour les utiliser dans SAS.

- **Lire les noms et les libellés.** Sélectionnez cette option afin de lire les noms et les libellés de variable dans IBM SPSS Modeler. Par défaut, cette option est sélectionnée et les noms de variable affichés dans le noeud Typer. Les libellés peuvent être affichées dans le Générateur de formules, les navigateurs de modèle et d'autres types de sortie, selon les options spécifiées dans la boîte de dialogue des propriétés du flux.
- **Lire les libellés sous forme de nom.** Sélectionnez cette option pour lire les libellés de variable descriptifs du fichier SAS au lieu des noms de champ abrégés, puis utilisez ces libellés en tant que noms de variable dans IBM SPSS Modeler.

Noeud source Excel

Le noeud source Excel permet d'importer des données de Microsoft Excel au format de fichier .xlsx.

Type de fichier. Sélectionnez le fichier de type Excel que vous souhaitez importer.

Importer le fichier. Indique le nom et l'emplacement de la feuille de calcul à importer.

Utiliser l'intervalle nommé. Permet d'indiquer un intervalle de cellules nommé, tel qu'il est défini dans la feuille de calcul Excel. Cliquez sur le bouton représentant des points de suspension (**...**) pour sélectionner la valeur souhaitée dans la liste des intervalles disponibles. Si vous utilisez un intervalle nommé, les autres paramètres de feuille de calcul et d'intervalle de données ne s'appliquent plus et sont, par conséquent, désactivés.

Choisir une feuille de calcul. Indique la feuille de calcul à importer, via un index ou un nom.

- **Par index** Définit la valeur d'index de la feuille de calcul à importer, 0 désignant la première feuille de calcul, 2, la deuxième et ainsi de suite.
- **Par nom.** Indiquez le nom de la feuille de calcul à importer. Cliquez sur le bouton représentant des points de suspension (**...**) pour sélectionner la valeur souhaitée dans la liste des feuilles de calcul disponibles.

Intervalle sur feuille de calcul. Vous pouvez importer des données en partant de la première ligne renseignée ou en indiquant un intervalle de cellules explicite.

- **L'intervalle commence à la première ligne non renseignée.** Repère la première cellule renseignée et l'utilise comme angle supérieur gauche de l'intervalle de données.
- **Intervalle de cellules explicite.** Vous permet de spécifier un intervalle explicite par ligne et par colonne. Par exemple, pour spécifier l'intervalle Excel `A1:D5` , vous pouvez entrer `A1` dans le premier champ et `D5` dans le second (ou bien, `R1C1` et `R5C4`). Toutes les lignes de l'intervalle indiqué sont renvoyées, y compris les lignes vides.

Sur les lignes non renseignées. Si plusieurs lignes vides sont rencontrées, vous pouvez **Arrêter la lecture** ou cliquer sur **Renvoyer des lignes non renseignées** pour poursuivre la lecture des données jusqu'à la fin de la feuille de calcul (lignes vides comprises).

La première ligne a des noms de colonne. Indique que la première ligne de l'intervalle spécifié doit être utilisée pour les noms de champ (de colonne). Si vous ne sélectionnez pas cette option, les noms de champ sont générés automatiquement.

Stockage de champ et niveau de mesure

Lors de la lecture de valeurs issues d'Excel, les champs de stockage numérique sont lus avec un niveau de mesure *continu* par défaut et les champs de chaîne sont lus avec un niveau *nominal*. Vous pouvez modifier manuellement le niveau de mesure (continu ou nominal) dans l'onglet Type, mais le stockage est, lui, déterminé automatiquement (il est toutefois possible, si nécessaire, de le modifier à l'aide d'une fonction de conversion, telle que `to_integer`, appliquée dans un noeud Remplacer ou Calculer). Pour plus d'informations, voir «Définition du stockage et du formatage des champs», à la page 9.

Par défaut, les champs comportant à la fois des valeurs numériques et des valeurs de type chaîne sont considérés comme numériques ; autrement dit, toute valeur de chaîne prendra la valeur nulle (manquante dans le système) dans IBM SPSS Modeler. Cela s'explique par le fait que contrairement à Excel IBM SPSS Modeler n'autorise pas les types de stockage mixtes dans un même champ. Pour éviter ce type de problème, vous pouvez définir manuellement le format de cellule sur `Texte` dans la feuille de calcul Excel ; toutes les valeurs (y compris les nombres) sont ainsi lues en tant que chaînes.

Noeud source XML

Cette fonction est disponible dans SPSS Modeler Professional et SPSS Modeler Premium.

Le noeud source XML vous permet d'importer des données depuis un fichier au format XML dans un flux IBM SPSS Modeler. XML est un langage standard d'échange de données, et représente pour beaucoup d'entreprises un format de choix approprié. Par exemple, un organisme gouvernemental d'imposition souhaite analyser des données provenant de déclarations de revenus soumises en ligne et dont les données sont au format XML (voir http://www.w3.org/standards/xml/).

L'importation de données XML dans un flux IBM SPSS Modeler vous permet d'exécuter une large gamme de fonctions d'analyse prédictive sur la source. Les données XML sont analysées dans un format tabulaire dans lequel les colonnes correspondent à différents niveaux d'imbrication des attributs et des éléments XML. Les éléments XML sont affichés au format XPath (voir http://www.w3.org/TR/xpath20/).

Important : le noeud source XML ne prend pas en compte la déclaration d'espace de nom. Ainsi, par exemple, vos fichiers XML ne peuvent pas contenir de signe deux-points (**:**) dans la balise `name`. Lors de leur exécution, vous recevriez des erreurs de caractères non valides.

Lire un fichier unique. Par défaut, SPSS Modeler lit un seul fichier, que vous spécifiez dans le champ **Source de données XML**.

Lire tous les fichiers XML d'un répertoire. Sélectionnez cette option si vous souhaitez lire tous les fichiers XML d'un répertoire particulier. Spécifiez l'emplacement dans le champ **Répertoire** qui s'affiche. Cochez la case **Inclure les sous-répertoires** pour lire des fichiers XML supplémentaires dans tous les sous-répertoires du répertoire spécifié.

Source de données XML. Saisissez le chemin complet et le nom du fichier de la source XML que vous souhaitez importer, ou utilisez le bouton Parcourir pour rechercher le fichier.

Schéma XML. (Facultatif) Spécifiez le chemin complet et le nom du fichier d'un fichier XSD ou DTD à partir duquel la structure XML est lue, ou utilisez le bouton Parcourir pour rechercher ce fichier. Si vous laissez ce champ vierge, la structure est lue à partir du fichier source XML. Un fichier XSD ou DTD peut avoir plus d'un élément racine. Dans ce cas, lorsque vous déplacez l'activation vers un autre champ, une boîte de dialogue s'affiche dans laquelle vous pouvez choisir l'élément racine à utiliser. Pour plus d'informations, voir la rubrique «Sélection de plusieurs éléments racine», à la page 49.

Remarque : Les indicateurs XSD sont ignorés par SPSS Modeler

Structure XML. Un arbre hiérarchique affichant la structure du fichier source XML (ou le schéma, si vous en avez spécifié un dans le champ **Schéma XML**). Pour définir une limite d'enregistrement, sélectionnez un élément et cliquez sur le bouton de la flèche droite pour copier l'élément dans le champ **Enregistrements**.

Afficher les attributs. Affiche ou masque les attributs, des éléments XML dans le champ **Structure XML**.

Enregistrements (expression XPath). Affiche la syntaxe XPath d'un élément copié à partir du champ de structure XML. Cet élément est alors mis en évidence dans la structure XML et définit la limite de l'enregistrement. A chaque fois que cet élément est rencontré dans le fichier source, un nouvel enregistrement est créé. Si ce champ est vide, le premier élément enfant sous la racine est utilisé comme limite d'enregistrement.

Lire toutes les données. Par défaut, toutes les données du fichier source sont lues dans le flux.

Spécifier les données à lire. Sélectionnez cette option pour importer des attributs, des éléments individuels ou les deux. Sélectionner cette option active le tableau Champs dans lequel vous pouvez spécifier les données que vous souhaitez importer.

Champs. Ce tableau répertorie les éléments et les attributs sélectionnés pour l'importation, si vous avez sélectionné l'option **Spécifier les données à lire**. Vous pouvez soit saisir la syntaxe XPath d'un élément ou d'un attribut directement dans la colonne XPath, soit sélectionner un élément ou un attribut dans la structure XML et cliquer sur le bouton de la flèche droite pour copier l'élément dans le tableau. Pour copier tous les éléments enfants et les attributs d'un élément, sélectionnez l'élément dans la structure XML et cliquez sur le bouton en forme de double-flèche.

- **XPath.** La syntaxe Xpath des éléments à importer.

- **Emplacement.** L'emplacement dans la structure XML des éléments à importer. **Chemin fixe** affiche le chemin de l'élément en relation avec l'élément mis en évidence dans la structure XML (ou le premier élément enfant sous la racine, si aucun élément n'est mis en évidence). **N'importe quel emplacement** indique un élément du nom donné à n'importe quel emplacement de la structure XML. **Personnalisé** s'affiche si vous saisissez l'emplacement directement dans la colonne XPath.

Sélection de plusieurs éléments racine

Alors qu'un fichier XML correctement formé ne peut contenir qu'un seul élément racine, un fichier XSD ou DTD peut en contenir plusieurs. Si l'un des éléments racines correspond à celui du fichier source XML, cet élément racine est utilisé, sinon vous devez en sélectionnez un.

Choisissez la racine à afficher. Sélectionnez l'élément racine à utiliser. L'élément par défaut est le premier élément racine dans la structure XSD ou DTD.

Suppression des espaces superflus des données source XML

Des sauts de ligne peuvent être implémentés dans les données source XML par une combinaison des caractères `[CR][LF]`. Dans certain cas, ces sauts de ligne peuvent apparaître au milieu d'une chaîne de texte, par exemple :

```
<description>An in-depth look at creating applications[CR][LF]
with XML.</description>
```

Ces sauts de ligne peuvent ne pas être visibles lorsque le fichier est ouvert dans certaines applications, par exemple dans un navigateur Web. Toutefois, lorsque les données sont lues dans le flux à travers le noeud source XML, les sauts de ligne sont convertis en une série de caractères d'espacement.

Vous pouvez corriger ceci en utilisant un noeud Remplacer et supprimer ces espaces superflus :

Voici un exemple de la manière de traiter ce problème :

1. Reliez un noeud Remplacer au noeud source XML.
2. Ouvrez le noeud Remplacer et utilisez le sélecteur de champs pour sélectionner le champ contenant les espace superflus.
3. Définissez **Remplacer** sur **Basé sur une condition** et **Condition** sur **true** (vrai).
4. Dans le champ **Remplacer par**, entrez `replace("  ","",@FIELD)` et cliquez sur OK.
5. Reliez un noeud Table au noeud Remplacer et exécutez le flux.

Dans la sortie du noeud Table, le texte apparaît maintenant sans les espaces supplémentaires.

Noeud Utilisateur

Le noeud Utilisateur représente une façon simple de créer des données synthétiques (à partir de zéro ou en modifiant des données existantes). Ceci est utile, par exemple, si vous souhaitez créer un jeu de données de test pour la modélisation.

Création intégrale de données

Le noeud Utilisateur est disponible dans la palette Sources. Vous pouvez l'ajouter directement à l'espace de travail de flux.

1. Cliquez sur l'onglet **Sources** de la palette de noeuds.
2. Faites glisser le noeud Utilisateur ou double-cliquez dessus pour l'ajouter à l'espace de travail de flux.
3. Double-cliquez pour ouvrir la boîte de dialogue correspondante, et spécifiez les champs et les valeurs.

Remarque : les noeuds Utilisateur sélectionnés depuis la palette Sources sont entièrement vides (ils ne contiennent aucun champ et aucune information sur les données). Ceci permet de créer intégralement des données synthétiques.

Génération de données à partir d'une source de données existante

Vous pouvez également générer un noeud Utilisateur à partir de tout noeud non terminal dans le flux :

1. Déterminez l'emplacement dans le flux du noeud à remplacer.
2. Cliquez avec le bouton droit de la souris sur le noeud qui alimentera le noeud Utilisateur en données, puis sélectionnez **Générer le noeud Utilisateur** dans le menu.
3. Le noeud Utilisateur sera associé à tous les processus en aval qui lui sont connectés, remplaçant le noeud existant à cet emplacement dans votre flux de données. Une fois généré, le noeud hérite de toute la structure des données et des informations de type de champ (le cas échéant) des métadonnées.

Remarque : si les données ne sont pas transmises par tous les noeuds du flux, ces derniers ne sont pas entièrement instanciés. Cela signifie que les valeurs de stockage et de données ne seront peut-être pas disponibles lors du remplacement par un noeud Utilisateur.

Définition des options du noeud Utilisateur

La boîte de dialogue d'un noeud Utilisateur contient différents outils que vous pouvez utiliser pour entrer des valeurs et définir la structure des données synthétiques. Pour un noeud généré, le tableau de l'onglet Données contient les noms de champ de la source de données d'origine. Pour un noeud ajouté à partir de la palette Sources, le tableau est vide. A l'aide des options du tableau, vous pouvez effectuer les opérations suivantes :

- Ajouter de nouveaux champs à l'aide du bouton Ajouter un nouveau champ situé à droite du tableau.
- Renommer des champs existants.
- Spécifier le stockage des données pour chaque champ.
- Indiquer des valeurs.
- Modifier l'ordre des champs affichés.

Saisie de données

Pour chaque champ, vous pouvez spécifier des valeurs ou en insérer à partir du jeu de données d'origine via le bouton de sélection des valeurs situé à droite du tableau. Pour plus d'informations sur la spécification des valeurs, reportez-vous aux règles décrites ci-dessous. Vous pouvez également choisir de laisser le champ vide. Les champs vides sont renseignés avec la valeur système nulle (`$null$`).

Pour indiquer des valeurs de chaîne, saisissez-les dans la colonne de valeurs, séparées par des espaces :

```
Fred Ethel Martin
```

Les chaînes qui incluent des espaces peuvent être mises entre doubles guillemets :

```
«Bill Smith» «Fred Martin» «Jack Jones»
```

Pour les champs numériques, vous pouvez entrer des valeurs multiples de la même manière (utiliser des espaces pour délimiter les valeurs) :

```
10 12 14 16 18 20
```

Vous pouvez également spécifier les mêmes valeurs en indiquant le premier et le dernier nombre (10, 20), et l'incrément qui les sépare (2). Dans ce cas, vous entrerez :

```
10,20,2
```

Ces deux méthodes peuvent être combinées par imbrication, comme suit :

```
1 5 7 10,20,2 21 23
```

Cette syntaxe produira les valeurs suivantes :

```
1 5 7 10 12 14 16 18 20 21 23
```

Les valeurs date et heure peuvent être saisies à l'aide du format par défaut actuel sélectionné dans la boîte de dialogue Propriétés du flux, par exemple :

```
11:04:00 11:05:00 11:06:00
2007-03-14 2007-03-15 2007-03-16
```

Pour les valeurs d'horodatage, qui comportent à la fois un composant date et heure, des guillemets doubles doivent être utilisés :

```
"2007-03-14 11:04:00" "2007-03-14 11:05:00" "2007-03-14 11:06:00"
```

Pour plus d'informations, reportez-vous aux commentaires sur le stockage des données ci-dessous.

Générer les données. Permet d'indiquer comment les enregistrements sont générés lorsque vous exécutez le flux.

- **Toutes les combinaisons.** Génère des enregistrements contenant toutes les combinaisons possibles des valeurs de champ de façon à ce que chaque valeur de champ apparaisse dans plusieurs enregistrements. Cette procédure peut parfois générer plus de données que ce que vous souhaitez. Par conséquent, vous pouvez faire suivre ce noeud d'un noeud Echantillonner.
- **Dans l'ordre**. Génère des enregistrements dans l'ordre dans lequel les valeurs du champ de données sont indiquées. Chaque valeur de champ n'apparaît que dans un enregistrement. Le nombre total d'enregistrements est égal au plus grand nombre de valeurs pour un seul champ. Lorsque le nombre de valeurs des champs est inférieur au plus grand nombre de valeurs, des valeurs non définies ($null$) sont insérées.

Exemple

Par exemple, les entrées suivantes généreront les enregistrements répertoriés dans les exemples de tableau ci-dessous.

- **Age.** `30,60,10`
- **TA.** `FAIBLE`
- **Cholestérol.** `NORMAL ELEVE`
- **Médicament.** (vide)

Tableau 6. Option Générer les données définie sur Toutes les combinaisons :.

Age	TA	Cholestérol	Médicament
30	FAIBLE	NORMAL	$null$
30	FAIBLE	ELEVEE	$null$
40	FAIBLE	NORMAL	$null$
40	FAIBLE	ELEVEE	$null$
50	FAIBLE	NORMAL	$null$
50	FAIBLE	ELEVEE	$null$
60	FAIBLE	NORMAL	$null$
60	FAIBLE	ELEVEE	$null$

Tableau 7. Option Générer les données définie sur Dans l'ordre :.

Age	TA	Cholestérol	Médicament
30	FAIBLE	NORMAL	$null$
40	$null$	ELEVEE	$null$
50	$null$	$null$	$null$
60	$null$	$null$	$null$

Stockage des données

Le stockage des données décrit la façon dont les données sont stockées dans un champ. Par exemple, un champ comportant les valeurs 1 et 0 stocke des nombres entiers. Il est à différencier du niveau de mesure, qui décrit l'utilisation des données et n'a aucune incidence sur le stockage. Par exemple, vous pouvez définir le niveau de mesure de nombre entier comportant les valeurs 1 et 0 comme étant un champ *indicateur*. En général, 1 correspond à la valeur *True (vrai)* et 0 à la valeur *False (faux)*. Alors que le stockage doit être déterminé au niveau de la source, le niveau de mesure peut être modifié à l'aide d'un noeud Typer en tout point du flux. Pour plus d'informations, voir la rubrique «Niveaux de mesure», à la page 145.

Les types de stockage disponibles sont les suivants :

- **Chaîne** Utilisé pour les champs qui contiennent des données non numériques, aussi appelées données alphanumériques. Une chaîne peut inclure n'importe quelle séquence de caractères, telle que *fred*, *Classe 2* ou *1234*. Notez que les nombres utilisés dans les chaînes ne peuvent pas être inclus dans les calculs.
- **Entier** Champ dont les valeurs sont des entiers.
- **Réel** Les valeurs sont des nombres qui peuvent inclure des décimales (non limitées à des entiers). Le format d'affichage est indiqué dans la boîte de dialogue Propriétés de flux et peut être ignoré pour des champs individuels dans un noeud Typer (onglet Format).
- **Date** Valeurs de date spécifiées au format standard, comme année, moi et jour (par exemple `2007-09-26`). Le format exact est indiqué dans la boîte de dialogue Propriétés de flux.
- **Temps** Temps mesuré en tant que durée. Par exemple, un appel de service ayant duré 1 heure, 26 minutes et 38 secondes peut être représenté sous la forme `01:26:38`, en fonction du format d'heure actuel indiqué dans la boîte de dialogue Propriétés de flux.
- **Horodatage** Valeurs qui incluent un composant de date et un composant d'heure, par exemple `2007–09–26 09:04:00`, selon les formats de date et d'heure définis dans la boîte de dialogue Propriétés du flux. Remarque : il se peut que les valeurs d'horodatage doivent être placées entre guillemets doubles pour être interprétées comme une valeur unique et non comme des valeurs date et heure distinctes. (Cela s'applique, par exemple, lors de la saisie de valeurs dans un noeud Utilisateur).
- **Liste** Introduit dans SPSS Modeler version 17, avec les nouveaux niveaux de mesure Géospatial et Collection, un champ de stockage Liste contient plusieurs valeurs pour un enregistrement unique. Il existe des versions de liste pour tous les autres types de stockage.

Tableau 8. Icônes des types de stockage Liste

Icône	Type de stockage
[A]	Liste de chaînes
[◌]	Liste de nombres entiers
[◌]	Liste de nombres réels

Tableau 8. Icônes des types de stockage Liste (suite)

Icône	Type de stockage
[]	Liste d'heures
[]	Liste de dates
[]	Liste d'horodatages
[[]]	Liste dont la profondeur est supérieure à zéro

De plus, pour une utilisation avec le niveau de mesure Collection, il existe des versions de liste pour les niveaux de mesure ci-dessous.

Tableau 9. Icônes des niveaux de mesure de liste

Icône	Niveau de mesure
[]	Liste d'éléments continus
[]	Liste d'éléments catégoriels
[]	Liste d'indicateurs
[]	Liste d'éléments nominaux
[]	Liste d'éléments ordinaux

Les listes peuvent être importées dans SPSS Modeler dans l'un des trois noeuds source (Analytic Server, Géospatial ou Délimité) ou créé dans vos flux à l'aide des noeuds d'opération de champ Calculer ou Remplacer.

Pour plus d'informations sur les listes et leur interaction avec les niveaux de mesure Collection et Géospatial, voir «Stockage de liste et niveaux de mesure associés», à la page 12

Conversion de stockages. Vous pouvez également convertir le stockage d'un champ à l'aide de diverses fonctions de conversion, comme `to_string` et `to_integer` dans un noeud Remplacer. Pour plus d'informations, voir la rubrique «Conversion du stockage à l'aide du noeud Remplacer», à la page 170. Notez que les fonctions de conversion (et toutes les autres fonctions qui nécessitent un type spécifique d'entrée, par exemple une valeur de date ou d'heure) dépendent des formats actuels indiqués dans la boîte de dialogue Propriétés de flux. Par exemple, si vous souhaitez convertir un champ de type chaîne avec des valeurs *Jan 2003*, *Fév 2003*, etc., en stockage de date, sélectionnez **MOIS AAAA** comme format de date par défaut pour le flux. Les fonctions de conversion sont également disponibles depuis le noeud Calculer pour la conversion temporaire lors d'un calcul. Vous pouvez également utiliser le noeud Calculer pour effectuer d'autres manipulations, telles que la modification du codage des champs de type chaîne contenant des valeurs catégorielles. Pour plus d'informations, voir la rubrique «Recodage des valeurs à l'aide du noeud Calculer», à la page 169.

Lecture de données mixtes. Au cours de la lecture des champs de stockage numérique (entier, nombre réel, heure, horodatage ou date), toutes les valeurs non numériques sont définies comme étant nulles ou manquantes dans le système. En effet, contrairement à certaines applications, IBM SPSS Modeler n'autorise pas les types de stockage mixtes au sein d'un champ. Pour éviter ce type de problème, faites en sorte que les champs comportant des données mixtes soient lus en tant que chaînes ; pour cela, modifiez le type de stockage dans le noeud source ou dans l'application externe.

Remarque : les noeuds Utilisateur générés peuvent déjà contenir ces informations de stockage, recueillies à partir du noeud source si ce dernier a été instancié. Un noeud non instancié ne contient pas d'informations de stockage ou de type d'utilisation.

Règles pour la spécification des valeurs

Dans le cas des champs symboliques, vous devez séparer les valeurs multiples par des espaces, par exemple :

```
ELEVE MOYEN FAIBLE
```

Pour les champs numériques, vous pouvez entrer des valeurs multiples de la même manière (utiliser des espaces pour délimiter les valeurs) :

```
10 12 14 16 18 20
```

Vous pouvez également spécifier les mêmes valeurs en indiquant le premier et le dernier nombre (10, 20), et l'incrément qui les sépare (2). Dans ce cas, vous entrerez :

```
10,20,2
```

Ces deux méthodes peuvent être combinées par imbrication, comme suit :

```
1 5 7 10,20,2 21 23
```

Cette syntaxe produira les valeurs suivantes :

```
1 5 7 10 12 14 16 18 20 21 23
```

Noeud Génération de simulation

Le noeud Génération de simulation permet de générer facilement des données simulées, soit sans données d'historique en utilisant des distributions statistiques spécifiées par l'utilisateur, soit automatiquement à l'aide des distributions obtenues via l'exécution d'un noeud Ajustement de simulation sur des données historiques existantes. La génération de données simulées s'avère utile si vous voulez évaluer le résultat d'un modèle prédictif en présence d'incertitude dans les entrées du modèle.

Création de données sans données d'historique

Le noeud Génération de simulation est disponible dans la palette Sources. Vous pouvez l'ajouter directement à l'espace de travail de flux.

1. Cliquez sur l'onglet **Sources** de la palette de noeuds.
2. Faites glisser le noeud Génération de simulation ou cliquez deux fois dessus pour l'ajouter à l'espace de travail de flux.
3. Cliquez deux fois dessus pour ouvrir sa boîte de dialogue et spécifier des champs, des types de stockage, des distributions statistiques et des paramètres de distribution.

Remarque : Les noeuds Génération de simulation sélectionnés dans la palette Sources sont entièrement vides (ils ne contiennent aucun champ et aucune information sur la distribution). Cela vous permet de créer entièrement des données simulées sans données d'historique.

Génération de données simulées à l'aide des données d'historique existantes

Un noeud Génération de simulation peut également être créé via l'exécution d'un noeud terminal Ajustement de simulation :

1. Cliquez avec le bouton droit de la souris sur le noeud Ajustement de simulation et sélectionnez **Exécuter** dans le menu.
2. Le noeud Génération de simulation apparaît sur le canevas de flux avec un lien de mise à jour vers le noeud Ajustement de simulation.
3. Une fois généré, le noeud Génération de simulation hérite tous les champs, types de stockage et informations sur la distribution statistique du noeud Ajustement de simulation.

Définition d'un lien de mise à jour vers un noeud Ajustement de simulation

Vous pouvez créer un lien entre un noeud Génération de simulation et un noeud Ajustement de simulation. Cela s'avère utile si vous voulez mettre à jour un ou plusieurs champs à l'aide des informations de la distribution la mieux adaptée, qui est déterminée par l'ajustement aux données d'historique.

1. Cliquez avec le bouton droit de la souris sur le noeud Génération de simulation.
2. Dans le menu, sélectionnez **Définir le lien Mettre à jour**. Le curseur change et devient un curseur de lien.
3. Cliquez sur un autre noeud. Si ce noeud est un noeud Ajustement de simulation, un lien est est créé. Sinon, aucun lien n'est créé et le curseur redevient normal.

Si les champs figurant dans le noeud Ajustement de simulation sont différents de ceux qui se trouvent dans le noeud Génération de simulation, un message vous informant de la différence s'affiche.

Lorsque le noeud Ajustement de simulation est utilisé pour mettre à jour le noeud Génération de simulation lié, le résultat varie suivant si les mêmes champs sont présents ou non dans les deux noeuds et si les champs sont déverrouillés ou non dans le noeud Génération de simulation. Les résultats de la mise à jour d'un noeud Ajustement de simulation sont présentés dans le tableau ci-dessous.

Tableau 10. Résultats de la mise à jour d'un noeud Ajustement de simulation

	Champ du noeud	**Ajustement de simulation**
Champ du noeud Génération de simulation	**Présent**	**Manquant**
Présent et déverrouillé.	Le champ est remplacé.	Le champ est supprimé.
Manquant.	Le champ est ajouté.	Aucune modification.
Présent et verrouillé.	La distribution du champ n'est pas remplacée. Les informations figurant dans la boîte de dialogue Informations sur l'ajustement et les corrélations sont mises à jour.	Le champ n'est pas remplacé. Les corrélations sont définies sur zéro.
La case **Ne pas effacer Min et Max lors du réajustement** est cochée.	Le champ est remplacé, sauf	les valeurs figurant dans la colonne Min, Max.
La case **Ne pas recalculer les corrélations lors du réajustement** est cochée.	Si le champ est déverrouillé, il est remplacé.	Les corrélations ne sont pas remplacées.

Suppression d'un lien de mise à jour vers un noeud Ajustement de simulation

Vous pouvez supprimer un lien entre un noeud Génération de simulation et un noeud Ajustement de simulation en procédant comme suit.

1. Cliquez avec le bouton droit de la souris sur le noeud Génération de simulation.
2. Dans le menu, sélectionnez **Supprimer le lien Mettre à jour**. Le lien est supprimé.

Définition des options du noeud Génération de simulation

Vous pouvez utiliser les options de l'onglet Données de la boîte de dialogue du noeud Génération de simulation pour exécuter les opérations suivantes :

- Afficher, spécifier et éditer les informations de distribution statistiques pour les champs.
- Afficher, spécifier et éditer les corrélations entre les champs.
- Spécifier le nombre d'itérations et d'observations à simuler.

Sélectionner un élément. Permet de basculer entre les trois vues du noeud Génération de simulation : Champs simulés, Corrélations et Options avancées.

Vue Champs simulés

Si le noeud Génération de simulation a été généré ou mis à jour à partir d'un noeud Ajustement de simulation à l'aide des données d'historique, la vue Champs simulés permet d'afficher et d'éditer les informations de distribution statistiques pour chaque champ. Les informations suivantes relatives à chaque champ sont copiées dans l'onglet **Types** du noeud Génération de simulation à partir du noeud Ajustement de simulation :

- Niveau de mesure
- Valeurs
- Manquant
- Vérifier
- Rôle

Si vous ne disposez pas de données d'historique, vous pouvez définir des champs et indiquer leurs distributions en sélectionnant un type de stockage et un type de distribution, puis en saisissant les paramètres requis. Avec cette méthode de génération de données, les informations relatives au niveau de mesure de chaque champ ne sont pas disponibles tant que les données ne sont pas instanciées, par exemple sur l'onglet **Types** ou dans un noeud Typer.

La vue Champs simulés contient plusieurs outils, que vous pouvez utiliser pour exécuter les tâches suivantes :

- Ajouter et supprimer des champs.
- Modifier l'ordre des champs affichés.
- Spécifier un type de stockage pour chaque champ.
- Spécifier une distribution statistique pour chaque champ.
- Spécifier les valeurs de paramètre pour la distribution statistique de chaque champ.

Champs simulés. Cette table contient une ligne vide si le noeud Génération de simulation a été ajouté au canevas de flux à partir de la palette Sources. Lorsque cette ligne est éditée, une nouvelle ligne vide est ajoutée au bas de la table. Si le noeud Génération de simulation a été créé à partir d'un noeud Ajustement de simulation, cette table contiendra une ligne pour chaque champ des données d'historique. Des lignes supplémentaires peuvent être ajoutées à la table en cliquant sur l'icône **Ajouter un nouveau champ**.

La table Champs simulés est composée des colonnes suivantes :

- **Champ.** Contient les noms des champs. Les noms de champ peuvent être modifiés en saisissant d'autres données dans les cellules.
- **Stockage.** Les cellules de cette colonne contiennent une liste déroulante des types de stockage. Les types de stockage disponibles sont les suivants : **Chaîne**, **Entier**, **Réel**, **Heure**, **Date** et **Horodatage**. Le choix du type de stockage détermine les distributions disponibles dans la colonne Distribution. Si le noeud Génération de simulation a été créé à partir d'un noeud Ajustement de simulation, le type de stockage est copié à partir du noeud Ajustement de simulation.

  Remarque : Pour les champs dont le type de stockage est Date/Heure, vous devez spécifier les paramètres de distribution sous forme d'entiers. Par exemple, pour indiquer le 1er janvier 1970 comme date moyenne, utilisez l'entier 0. L'entier signé représente le nombre de secondes depuis (ou avant) le 1er janvier 1970 à minuit.
- **Statut.** Les icônes figurant dans la colonne Statut indiquent le statut d'ajustement pour chaque champ.

Aucune distribution n'a été spécifiée pour le champ, ou un ou plusieurs paramètres de distribution sont manquants. Pour exécuter la simulation, vous devez indiquer une distribution pour ce champ et saisir des valeurs valides pour les paramètres.

Le champ est défini sur la distribution d'ajustement la mieux adaptée.
Remarque : Cette icône ne peut s'afficher que si le noeud Génération de simulation est créé à partir d'un noeud Ajustement de simulation.

La distribution d'ajustement la mieux adaptée a été remplacée par une autre distribution provenant de la sous-boîte de dialogue Informations sur l'ajustement. Pour plus d'informations, voir la rubrique «Informations sur l'ajustement», à la page 62.

La distribution a été spécifiée ou éditée manuellement et peut inclure un paramètre spécifié à plusieurs niveaux.

- **Verrouillé.** Le verrouillage d'un champ simulé, en cochant la case située dans la colonne contenant l'icône de verrouillage, exclut le champ de la mise à jour automatique par un noeud Ajustement de simulation lié. Cela s'avère particulièrement utile lorsque vous spécifiez manuellement une distribution et que vous voulez vous assurer qu'elle ne sera pas affectée par l'ajustement de distribution automatique lors de l'exécution d'un noeud Ajustement de simulation lié.
- **Distribution.** Les cellules de cette colonne contiennent une liste déroulante des distributions statistiques. Le choix du type de stockage détermine les distributions disponibles dans cette colonne pour un champ donné. Pour plus d'informations, voir «Distributions», à la page 65.

  Remarque : Vous ne pouvez pas spécifier la distribution Fixe pour chaque champ. Si vous voulez que chaque champ figurant dans vos données générées soit fixe, vous pouvez utiliser un noeud Utilisateur suivi d'un noeud équilibrer.
- **Paramètres.** Les paramètres de distribution associés à chaque distribution ajustée s'affichent dans cette colonne. Lorsqu'il existe plusieurs valeurs pour un paramètre, elles sont séparées par une virgule. La spécification de plusieurs valeurs pour un paramètre génère plusieurs itérations pour la simulation. Pour plus d'informations, voir «Itérations», à la page 64. L'icône affichée dans la colonne Statut reflète si des paramètres sont manquants. Pour spécifier des valeurs pour les paramètres, cliquez sur cette colonne dans la ligne correspondant au champ qui vous intéresse et sélectionnez **Spécifier** dans la liste.

La sous-boîte de dialogue Spécifier les paramètres s'ouvre. Pour plus d'informations, voir «Spécifier les paramètres», à la page 63. Cette colonne est désactivée si l'option Empirique est choisie dans la colonne Distribution.

- **Min, Max.** Pour certaines distributions, cette colonne vous permet de spécifier une valeur minimale et/ou une valeur maximale pour les données simulées. Les données simulées inférieures à la valeur minimale et supérieures à la valeur maximale sont rejetées, même si elles sont valides pour la distribution spécifiée. Pour spécifier des valeurs minimale et maximale, cliquez sur cette colonne dans la ligne correspondant au champ qui vous intéresse et sélectionnez **Spécifier** dans la liste. La sous-boîte de dialogue Spécifier les paramètres s'ouvre. Pour plus d'informations, voir «Spécifier les paramètres», à la page 63. Cette colonne est désactivée si l'option Empirique est choisie dans la colonne Distribution.

Utiliser l'ajustement le plus proche. Cette option est activée uniquement si le noeud Génération de simulation a été créé automatiquement à partir d'un noeud Ajustement de simulation à l'aide des données d'historique et qu'une seule ligne est sélectionnée dans la table Champs simulés. Elle remplace les informations relatives au champ sur la ligne sélectionnée par les informations de la distribution d'ajustement la mieux adaptée pour le champ. Si les informations figurant sur la ligne sélectionnée ont été modifiées, appuyez sur ce bouton pour restaurer la distribution d'ajustement la mieux adaptée qui a été déterminée par le noeud Ajustement de simulation.

Informations sur l'ajustement. Cette option est activée uniquement si le noeud Génération de simulation a été créé automatiquement à partir d'un noeud Ajustement de simulation. La sous-boîte de dialogue Informations sur l'ajustement s'ouvre. Pour plus d'informations, voir «Informations sur l'ajustement», à la page 62.

Plusieurs tâches utiles peuvent être exécutées à l'aide des icônes situées à droite de la vue Champs simulés. Ces icônes sont décrites dans le tableau suivant.

Tableau 11. Icônes de la vue Champs simulés.

Icône	Infobulle	Description
	**Editer les paramètres de distribution**	Cette option est activée uniquement lorsqu'une seule ligne est sélectionnée dans la table Champs simulés. La sous-boîte de dialogue Spécifier les paramètres s'ouvre pour la ligne sélectionnée. Pour plus d'informations, voir «Spécifier les paramètres», à la page 63.
	**Ajouter un nouveau champ**	Cette option est activée uniquement lorsqu'une seule ligne est sélectionnée dans la table Champs simulés. Elle ajoute une nouvelle ligne vide au bas de la table Champs simulés.
	**Créer plusieurs copies**	Cette option est activée uniquement lorsqu'une seule ligne est sélectionnée dans la table Champs simulés. La sous-boîte de dialogue Champ clone s'ouvre. Pour plus d'informations, voir «Champ clone», à la page 61.
	**Supprimer le champ sélectionné**	Supprime la ligne sélectionnée de la table Champs simulés.

Tableau 11. Icônes de la vue Champs simulés (suite).

Icône	Infobulle	Description
	**Placer en haut**	Cette option est activée uniquement si la ligne sélectionnée n'est pas déjà la première ligne de la table Champs simulés. Elle déplace la ligne sélectionnée en haut de la table Champs simulés. Cette action affecte l'ordre des champs dans les données simulées.
	**Monter d'un niveau**	Cette option est activée uniquement si la ligne sélectionnée n'est pas la première ligne de la table Champs simulés. Elle déplace la ligne sélectionnée d'un niveau vers le haut dans la table Champs simulés. Cette action affecte l'ordre des champs dans les données simulées.
	**Descendre d'un niveau**	Cette option est activée uniquement si la ligne sélectionnée n'est pas la dernière ligne de la table Champs simulés. Elle déplace la ligne sélectionnée d'un niveau vers le bas dans la table Champs simulés. Cette action affecte l'ordre des champs dans les données simulées.
	**Placer en bas**	Cette option est activée uniquement si la ligne sélectionnée n'est pas déjà la dernière ligne de la table Champs simulés. Elle déplace la ligne sélectionnée au bas de la table Champs simulés. Cette action affecte l'ordre des champs dans les données simulées.

Ne pas effacer Min et Max lors du réajustement. Lorsque cette option est sélectionnée, les valeurs minimale et maximale ne sont pas remplacées quand les distributions sont mises à jour via l'exécution d'un noeud Ajustement de simulation connecté.

Vue Corrélations

Les champs d'entrée dans les modèles prédictifs sont souvent corrélés ; par exemple, la hauteur et la pondération. Les corrélations entre les champs à simuler doivent être prises en compte pour s'assurer que les valeurs simulées les conservent.

Si le noeud Génération de simulation a été généré ou mis à jour à partir d'un noeud Ajustement de simulation à l'aide des données d'historique, la vue Corrélations permet d'afficher et d'éditer les corrélations calculées entre les paires de champs. Si vous ne disposez pas de données d'historique, vous pouvez spécifier manuellement les corrélations en fonction de la façon dont les champs sont corrélés.

Remarque : Avant toute génération de données, la matrice de corrélation est automatiquement vérifiée pour voir si elle est positive semi-définie et si elle peut donc être inversée. Une matrice peut être inversée si ses colonnes sont indépendantes linéairement. Si la matrice de corrélation ne peut pas être inversée, elle sera automatiquement ajustée pour devenir réversible.

Vous pouvez choisir d'afficher les corrélations au format matrice ou liste.

Matrice de corrélation. Cette option affiche les corrélations entre les paires de champs d'une matrice. Les noms de champ sont répertoriés par ordre alphabétique, le long d'une diagonale partant du haut vers la gauche de la matrice. Seules les cellules situées en dessous de la diagonale peuvent être éditées ; une valeur comprise entre -1,000 et 1,000 inclus doit être saisie. La cellule située au dessus de la diagonale est actualisée lorsque sa cellule en miroir en dessous de la diagonale n'est plus mise en évidence ; les deux cellules affichent alors la même valeur. Les cellules diagonales sont toujours désactivées et ont toujours une corrélation de 1,000. La valeur par défaut pour toutes les autres cellules est 0,000. La valeur 0,000 indique qu'il n'existe pas de corrélation entre la paire de champs associée. Seuls les champs continus et ordinaux sont inclus dans la matrice. Les champs nominaux, catégoriels et indicateurs, ainsi que les champs auxquels est affectée la distribution Fixe ne s'affichent pas dans la table.

Liste de Corrélations. Cette option affiche les corrélations entre les paires de champs d'une table. Chaque ligne de la table affiche la corrélation entre une paire de champs. Vous ne pouvez pas ajouter ni supprimer de lignes. Les colonnes avec les en-têtes Champ 1 et Champ 2 contiennent les noms de champ, qui ne sont pas modifiables. La colonne Corrélation contient les corrélations, qui peuvent être éditées ; une valeur comprise entre -1,000 et 1,000 inclus doit être saisie. La valeur par défaut pour toutes les cellules est 0,000. Seuls les champs continus et ordinaux sont inclus dans la liste. Les champs nominaux, catégoriels et indicateurs, ainsi que les champs auxquels est affectée la distribution Fixe ne s'affichent pas dans la liste.

Redéfinir les corrélations. La boîte de dialogue Redéfinir les corrélations s'ouvre. Si des données d'historique sont disponibles, vous pouvez choisir l'une des trois options suivantes :

- **Ajusté.** Remplace les corrélations actuelles par celles calculées à l'aide des données d'historique.
- **Zéros.** Remplace les corrélations actuelles par des zéros.
- **Annuler.** Permet de fermer la boîte de dialogue. Les corrélations ne sont pas modifiées.

Si les données d'historique ne sont pas disponibles mais que vous avez apporté des modifications aux corrélations, vous pouvez choisir de remplacer les corrélations actuelles par des zéros, ou les annuler.

Afficher comme. Sélectionnez **Table** pour afficher les corrélations sous forme de matrice. Sélectionnez **Liste** pour afficher les corrélations sous forme de liste.

Ne pas recalculer les corrélations lors du réajustement. Sélectionnez cette option pour spécifier manuellement des corrélations et empêcher leur remplacement lors de l'ajustement automatique des distributions à l'aide d'un noeud Ajustement de simulation et des données d'historique.

Utiliser la table de contingence à plusieurs facteurs ajustée pour les entrées avec une distribution qualitative. Par défaut, tous les champs comportant une distribution catégorielle sont inclus dans une table de contingence (ou une table de contingence à plusieurs facteurs, suivant le nombre de champs dotés d'une distribution catégorielle). La table de contingence est créée, comme les corrélations, lors de l'exécution d'un noeud Ajustement de simulation. La table de contingence ne peut pas être affichée. Lorsque cette option est sélectionnée, les champs comportant une distribution catégorielle sont simulés à l'aide des pourcentages réels provenant de la table de contingence. Par conséquent, les associations entre les champs nominaux sont recréées dans les nouvelles données simulées. Lorsque cette option est désélectionnée, les champs comportant des distributions catégorielles sont simulés à l'aide des pourcentages attendus provenant de la table de contingence. Si vous modifiez un champ, il est supprimé de la table de contingence.

Vue Options avancées

Nombre d'observations à simuler. Affiche les options permettant de spécifier le nombre d'observations à simuler, ainsi que le mode de dénomination des itérations.

- **Nombre maximal d'observations.** Cette option indique le nombre maximal d'observations de données simulées et les valeurs cible associées à générer. La valeur par défaut est 10000, la valeur minimale est 1000 et la valeur maximale est 2 147 483 647.

- **Itérations.** Ce nombre est calculé automatiquement et n'est pas modifiable. Une itération est créée automatiquement chaque fois que plusieurs valeurs sont spécifiées pour un paramètre de distribution.
- **Nombre total de lignes.** Cette option est activée uniquement lorsque le nombre d'itérations est supérieur à 1. Ce nombre est calculé automatiquement, à l'aide de l'équation présentée, et n'est pas modifiable.
- **Créer un champ d'itération.** Cette option est activée uniquement lorsque le nombre d'itérations est supérieur à 1. Lorsqu'elle est sélectionnée, le champ **Nom** est activé. Pour plus d'informations, voir «Itérations», à la page 64.
- **Nom.** Cette option est activée uniquement lorsque la case **Créer un champ d'itération** est cochée et que le nombre d'itérations est supérieur à 1. Pour éditer le nom du champ d'itération, saisissez-en un autre dans ce champ de texte. Pour plus d'informations, voir «Itérations», à la page 64.

Valeur de départ aléatoire. Définir une valeur de départ aléatoire vous permet de dupliquer votre simulation.

- **Dupliquer les résultats**. Lorsque cette option est sélectionnée, le bouton **Générer** et le champ **Valeur de départ aléatoire** sont activés.
- **Valeur de départ aléatoire.** Cette option est activée uniquement lorsque la case **Dupliquer les résultats** est cochée. Ce champ vous permet de spécifier un entier à utiliser comme valeur de départ aléatoire. La valeur par défaut est 629 111 597.
- **Générer**. Cette option est activée uniquement lorsque la case **Dupliquer les résultats** est cochée. Elle crée un entier pseudo-aléatoire compris entre 1 et 999 999 999 inclus, dans le champ **Valeur de départ aléatoire**.

Champ clone

La boîte de dialogue Champ clone vous permet de spécifier le nombre de copies du champ sélectionné à créer, ainsi que le nom de chacune d'elles. Il est utile de disposer de plusieurs copies des champs lors de l'examen d'effets composés, par exemple les taux d'intérêt ou de croissance sur un certain nombre de périodes successives.

La barre de titre de la boîte de dialogue contient le nom du champ sélectionné.

Nombre de copies à effectuer. Contient le nombre de copies du champ à créer. Cliquez sur les flèches pour sélectionner le nombre de copies à créer. Le nombre minimal de copies est 1 et le nombre maximal, 512. La valeur par défaut est 10.

Copier le(s) caractère(s) de suffixe. Contient les caractères qui sont ajoutés à la fin du nom de champ de chaque copie. Ces caractères séparent le nom de champ du numéro de copie. Les caractères de suffixe peuvent être modifiés en saisissant des caractères dans ce champ. Ce champ peut rester vide ; dans ce cas, il n'y aura aucun caractère entre le nom de champ et le numéro de copie. Le caractère par défaut est un trait de soulignement.

Numéro de copie initiale. Contient le numéro de suffixe de la première copie. Cliquez sur les flèches pour sélectionner le numéro de copie initiale. Le numéro de copie initiale minimal est 1 et le numéro maximal, 1000. La valeur par défaut est 1.

Etape de numéro de copie. Contient l'incrément pour les numéros de suffixe. Cliquez sur les flèches pour sélectionner l'incrément. L'incrément minimal est 1 et l'incrément maximal, 255. La valeur par défaut est 1.

Champs. Contient un aperçu des noms de champs pour les copies, qui est mis à jour si l'un des champs de la boîte de dialogue Champ clone est édité. Ce texte est généré automatiquement et n'est pas modifiable.

OK. Génère toutes les copies indiquées dans la boîte de dialogue. Les copies sont ajoutées à la table Champs simulés dans la boîte de dialogue du noeud Génération de simulation, directement en dessous de la ligne contenant le champ copié.

Annuler. Permet de fermer la boîte de dialogue. Les modifications effectuées sont annulées.

Informations sur l'ajustement

La boîte de dialogue Informations sur l'ajustement est disponible uniquement si le noeud Génération de simulation a été créé ou mis à jour via l'exécution d'un noeud Ajustement de simulation. Elle affiche les résultats de l'ajustement de distribution automatique pour le champ sélectionné. Les distributions sont ordonnées par qualité d'ajustement, la distribution d'ajustement la plus appropriée étant répertoriée en tête de liste. Cette boîte de dialogue vous permet d'effectuer les tâches suivantes :

- Examiner les distributions ajustées aux données d'historique.
- Sélectionner l'une des distributions ajustées.

Champ. Contient le nom du champ sélectionné. Ce texte n'est pas modifiable.

Traiter comme (Mesure). Affiche le type de mesure du champ sélectionné. Il provient de la table Champs simulés située dans la boîte de dialogue du noeud Génération de simulation. Le type de mesure peut peut être modifié en cliquant sur la flèche et en sélectionnant un autre type dans la liste déroulante. Il existe trois options : **Continu**, **Nominal** et **Ordinal**.

Distributions. La table Distributions affiche toutes les distributions adaptées au type de mesure. Les distributions qui ont été ajustées aux données d'historique sont ordonnées selon la qualité d'ajustement, de la meilleure à la pire. La qualité d'ajustement est déterminée par la statistique d'ajustement choisie dans le noeud Ajustement de simulation. Les distributions qui n'ont pas été ajustées aux données d'historique sont répertoriées dans la table par ordre alphabétique, en dessous de celles qui l'ont été.

La table Distribution contient les colonnes suivantes :

- **Utiliser.** Le bouton radio sélectionné indique la distribution actuellement choisie pour le champ. Vous pouvez remplacer la distribution d'ajustement la mieux adaptée en sélectionnant le bouton radio correspondant à la distribution souhaitée dans la colonne Utiliser. La sélection d'un bouton radio dans la colonne Utiliser permet également d'afficher un graphique de la distribution superposé sur un histogramme (ou graphique à barres) des données d'historique pour le champ sélectionné. Vous ne pouvez sélectionner qu'une seule distribution à la fois.
- **Distribution.** Contient le nom de la distribution. Cette colonne n'est pas modifiable.
- **Statistiques d'ajustement.** Contient les statistiques d'ajustement calculées pour la distribution. Cette colonne n'est pas modifiable. Le contenu de la cellule dépend du type de mesure du champ :
  - **Continu.** Contient les résultats des tests Anderson-Darling et Kolmogorov-Smirnoff. Les valeurs p associées aux tests sont également affichées. La statistique d'ajustement choisie comme critère de qualité d'ajustement dans le noeud Ajustement de simulation s'affiche en tête et est utilisée pour ordonner les distributions. Les statistiques Anderson-Darling sont affichées sous la forme A=*aval* P=*pval*. Les statistiques Kolmogorov-Smirnoff sont affichées sous la forme K=*kval* P=*pval*. Si une statistique ne peut pas être calculée, un point est affiché à la place d'un nombre.
  - **Nominal et Ordinal.** Contient les résultats du test Khi-deux. La valeur p associée au test est également affichée. Les statistiques sont affichées sous la forme Khi-deux=*val* P=*pval*. Si la distribution n'a pas été ajustée, `Non ajusté` s'affiche. Si la distribution ne peut pas être ajustée mathématiquement, `Ne peut être ajusté` s'affiche.

  Remarque : La cellule est toujours vide pour la distribution Empirique.
- **Paramètres.** Contient les paramètres de distribution associés à chaque distribution ajustée. Les paramètres sont affichés sous la forme *nom_paramètre* = *valeur_paramètre*, les paramètres étant séparés par un espace. Pour la distribution catégorielle, les noms de paramètres sont les catégories et les

valeurs de paramètre, les probabilités associées. Si la distribution n'a pas été ajustée aux données d'historique, la cellule est vide. Cette colonne n'est pas modifiable.

Miniature Histogramme. Affiche un graphique de la distribution sélectionnée superposé sur un histogramme des données d'historique du champ sélectionné.

Miniature Distribution. Affiche une explication et une illustration de la distribution sélectionnée.

OK. Ferme la boîte de dialogue et met à jour les valeurs des colonnes Mesure, Distribution, Paramètres et Min, Max de la table Champs simulés pour le champ sélectionné à l'aide des informations provenant de la distribution sélectionnée. L'icône figurant dans la colonne Statut est également actualisée pour indiquer si la distribution sélectionnée est la distribution avec le meilleur ajustement aux données.

Annuler. Permet de fermer la boîte de dialogue. Les modifications effectuées sont annulées.

Spécifier les paramètres

La boîte de dialogue Spécifier les paramètres permet de spécifier manuellement les valeurs de paramètre pour la distribution du champ sélectionné. Vous pouvez également choisir une distribution différente pour le champ sélectionné.

La boîte de dialogue Spécifier les paramètres peut être ouverte de trois façons :

- Cliquez deux fois sur un nom de champ dans la table Champs simulés située dans la boîte de dialogue du noeud Génération de simulation.
- Cliquez sur la colonne Paramètres ou Min, Max de la table Champs simulés et sélectionnez **Spécifier** dans la liste.
- Dans la table Champs simulés, sélectionnez une ligne, puis cliquez sur l'icône **Editer les paramètres de distribution**.

Champ. Contient le nom du champ sélectionné. Ce texte n'est pas modifiable.

Distribution. Contient la distribution du champ sélectionné. Elle provient de la table Champs simulés. La distribution peut être modifiée en cliquant sur la flèche et en sélectionnant une autre distribution dans la liste déroulante. Les distributions disponibles dépendent du type de stockage du champ sélectionné.

Côtés. Cette option est disponible uniquement lorsque la distribution Dé à jouer est sélectionnée dans le champ **Distribution**. Cliquez sur les flèches pour sélectionner le nombre de côtés, ou catégories, à utiliser pour fractionner le champ. Le nombre minimal de côtés est deux et le nombre maximal, 20. La valeur par défaut est 6.

Paramètres de distribution. La table Paramètres de distribution contient une ligne pour chaque paramètre de la distribution choisie.

Remarque : La distribution utilise un paramètre de taux avec un paramètre de forme égale à K et un paramètre d'échelle inversée égal à 1.
Ce tableau contient deux colonnes :

- **Paramètre.** Contient les noms des paramètres. Cette colonne n'est pas modifiable.
- **Valeur(s).** Contient les valeurs des paramètres. Si le noeud Génération de simulation a été créé ou mis à jour à partir d'un noeud Ajustement de simulation, les cellules de cette colonne contiennent les valeurs de paramètre qui ont été déterminées par l'ajustement de la distribution aux données d'historique. Si le noeud Génération de simulation a été ajouté au canevas de flux à partir de la palette Noeuds source, les cellules de cette colonne sont vides. Les valeurs peuvent être modifiées en saisissant d'autres données dans les cellules. Pour plus d'informations sur les paramètres requis par chaque distribution et les valeurs de paramètre acceptables, consultez la rubrique «Distributions», à la page 65.

Si un paramètre contient plusieurs valeurs, elles doivent être séparées par des virgules. La spécification de plusieurs valeurs pour un paramètre définit plusieurs itérations pour la simulation. Vous pouvez uniquement spécifier plusieurs valeurs pour un paramètre.

Remarque : Pour les champs dont le type de stockage est Date/Heure, vous devez spécifier les paramètres de distribution sous forme d'entiers. Par exemple, pour indiquer le 1er janvier 1970 comme date moyenne, utilisez l'entier 0.

Remarque : Lorsque la distribution Dé à jouer est sélectionnée, la table Paramètres de distribution est légèrement différente. Elle contient une ligne par côté (ou catégorie), ainsi qu'une colonne Valeur et une colonne Probabilité. La colonne Valeur contient un libellé pour chaque catégorie. Les valeurs par défaut pour les libellés sont les entiers 1-N, où N correspond au nombre de côtés. Les libellés peuvent être modifiés en saisissant d'autres données dans les cellules. Vous pouvez entrer n'importe quelle valeur dans les cellules. Si vous voulez utiliser une valeur non numérique, le type de stockage du champ de données doit être modifié en chaîne, le cas échéant. La colonne Probabilité contient la probabilité de chaque catégorie. Les probabilités ne sont pas modifiables et sont calculées sous la forme 1/N.

Aperçu. Affiche un exemple de graphique de distribution, qui se fonde sur les paramètres spécifiés. Si deux valeurs au moins ont été spécifiées pour un paramètre, des exemples de graphique s'affichent pour chaque valeur du paramètre. Si des données d'historique sont disponibles pour le champ sélectionné, le graphique de la distribution vient se superposer à l'histogramme des données d'historique.

Paramètres facultatifs. Utilisez ces options pour spécifier une valeur minimale et/ou une valeur maximale pour les données simulées. Les données simulées inférieures à la valeur minimale et supérieures à la valeur maximale sont rejetées, même si elles sont valides pour la distribution spécifiée.

- **Spécifier le minimum.** Sélectionnez cette option pour activer le champ **Rejeter les valeurs en dessous de**. La case à cocher est désactivée si la distribution Empirique est sélectionnée.
- **Rejeter les valeurs en dessous de.** Cette option est activée uniquement si **Spécifier le minimum** est sélectionné. Saisissez une valeur minimale pour les données simulées. Toute valeur simulée inférieure à cette valeur sera rejetée.
- **Spécifier le maximum.** Sélectionnez cette option pour activer le champ **Rejeter les valeurs au-dessus de**. La case à cocher est désactivée si la distribution Empirique est sélectionnée.
- **Rejeter les valeurs au-dessus de.** Cette option est activée uniquement si **Spécifier le maximum** est sélectionné. Saisissez une valeur maximale pour les données simulées. Toute valeur simulée supérieure à cette valeur sera rejetée.

OK. Ferme la boîte de dialogue et met à jour les valeurs des colonnes Distribution, Paramètres et Min, Max de la table Champs simulés pour le champ sélectionné. L'icône figurant dans la colonne Statut est également actualisée pour refléter la distribution sélectionnée.

Annuler. Permet de fermer la boîte de dialogue. Les modifications effectuées sont annulées.

Itérations

Si vous avez spécifié plusieurs valeurs pour un champ fixe ou un paramètre de distribution, un ensemble indépendant d'observations simulées - en réalité, une simulation distincte - est généré pour chaque valeur spécifiée. Cela vous permet d'observer l'effet que produit le changement du champ ou du paramètre. Chaque ensemble d'observations simulées est appelé *itération*. Dans les données simulées, les itérations sont empilées.

Si la case **Créer un champ d'itération** dans la vue Options avancées de la boîte de dialogue du noeud Génération de simulation est cochée, un champ d'itération est ajouté aux données simulées en tant que champ nominal avec stockage numérique. Pour modifier le nom de ce champ, saisissez-en un autre dans le champ **Nom** de la vue Options avancées. Ce champ contient un libellé indiquant à quelle itération appartient chaque observation simulée. Le format des libellés dépend du type d'itération :

- **Itération d'un champ fixe.** Le libellé est le nom du champ, suivi d'un signe égal, puis de la valeur du champ pour cette itération, à savoir

  nom_champ = valeur_champ
- **Itération d'un paramètre de distribution.** Le libellé est le nom du champ, suivi du signe deux-points, puis du nom du paramètre itéré, suivi d'un signe égal et de la valeur du paramètre pour cette itération, à savoir

  nom_champ:*nom_paramètre = valeur_paramètre*
- **Itération d'un paramètre de distribution pour une distribution catégorielle ou de plage.** Le libellé est le nom du champ, suivi d'un signe deux-points, suivi du terme "Itération" et du numéro d'itération, à savoir

  nom_champ: Itération *numéro_itération*

Distributions

Vous pouvez spécifier manuellement la distribution de probabilité pour un champ en ouvrant la boîte de dialogue Spécifier les paramètres pour ce champ, en sélectionnant la distribution souhaitée dans la liste **Distribution** et en saisissant les paramètres de distribution dans la table **Paramètres de distribution**. Voici quelques remarques sur des distributions particulières :

- **Catégorielle.** La distribution catégorielle décrit un champ d'entrée doté d'un nombre fixe de valeurs numériques, appelées catégories. Chaque catégorie possède une probabilité associée, de sorte que la somme des probabilités de toutes les catégories soit égale à 1.

  Remarque : Si vous spécifiez des probabilités pour les catégories dont la somme n'est pas égale à 1, vous recevez un avertissement.
- **Binomiale négative - Echecs.** Décrit la distribution du nombre d'échecs dans une séquence d'essais avant qu'un nombre spécifique de réussites ne soit observé. Le paramètre *Seuil* représente le nombre de réussites spécifié, tandis que le paramètre *Probabilité* représente la probabilité de réussite dans un essai donné.
- **Binomiale négative - Essais.** Décrit la distribution du nombre d'essais requis avant qu'un nombre spécifié de réussites ne soit observé. Le paramètre *Seuil* représente le nombre de réussites spécifié, tandis que le paramètre *Probabilité* représente la probabilité de réussite dans un essai donné.
- **Intervalle.** Cette distribution est composée d'un ensemble d'intervalles, une probabilité étant affectée à chacun d'eux de sorte que la somme des probabilités sur tous les intervalles soit égale à 1. Les valeurs situées au sein d'un intervalle donné proviennent d'une distribution uniforme définie sur cet intervalle. Les intervalles sont spécifiés en entrant une valeur minimale, une valeur maximale et une probabilité.

  Par exemple, vous pensez que le coût d'une matière première a 40 % de chances de se situer dans la plage 10 à 15 par unité et 60 % de chances d'être compris entre 15 et 20 par unité. Vous voulez modéliser le coût avec une distribution Plage composée des deux intervalles [10 - 15] et [15 - 20], en définissant la probabilité associée au premier intervalle sur 0,4 et la probabilité associée au deuxième intervalle sur 0,6. Les intervalles n'ont pas besoin d'être contigus et peuvent même se chevaucher. Par exemple, vous pouvez spécifier les intervalles 10 à 15 et 20 à 25 ou 10 à 15 et 13 à 16.€.
- **Weibull.** Le paramètre *Emplacement* est un paramètre d'emplacement facultatif, qui indique où se situe l'origine de la distribution.

Le tableau ci-dessous présente les distributions disponibles pour l'ajustement de distribution personnalisé, ainsi que les valeurs acceptables pour les paramètres. Certaines de ces distributions sont disponibles pour l'ajustement personnalisé à des types de stockage particuliers, même si elles ne sont pas ajustées automatiquement à ces types de stockage par le noeud Ajustement de simulation.

Tableau 12. Distributions disponibles pour l'ajustement personnalisé

Proportion	Type de stockage pris en charge pour l'ajustement personnalisé	Paramètres	Limites du paramètre	Remarques
Bernoulli	Entier, réel ou date/heure	Probabilité	0 ≤ *Probabilité* ≤ 1	
Bêta	Entier, réel ou date/heure	Forme 1 Forme 2 Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Les paramètres Minimum et Maximum sont facultatifs.
Binomiale	Entier, réel ou date/heure	Nombre d'essais (n) Probabilité Minimum Maximum	> 0, entier 0 ≤ *Probabilité* ≤ 1 < *Maximum* > *Minimum*	Le nombre d'essais doit être un entier. Les paramètres Minimum et Maximum sont facultatifs.
Catégorielle	Entier, réel, date/heure ou chaîne	Nom (ou libellé) de la catégorie	0 ≤ *Valeur* ≤ 1	La valeur représente la probabilité de la catégorie. La somme des valeurs doit être égale à 1 ; sinon, un avertissement est généré.
Dé à jouer	Entier ou chaîne	Côtés	2 ≤ *Côtés* ≤ 20	La probabilité de chaque catégorie (côté) est calculée sous la forme 1/N, où N représente le nombre de côtés. Les probabilités ne sont pas modifiables.
Empirique	Entier, réel ou date/heure			Vous ne pouvez pas éditer la distribution empirique, ni la sélectionner comme type. La distribution Empirique est disponible uniquement lorsque des données d'historique sont présentes.
Exponentielle	Entier, réel ou date/heure	Echelle Minimum Maximum	> 0 < *Maximum* > *Minimum*	Les paramètres Minimum et Maximum sont facultatifs.

Tableau 12. Distributions disponibles pour l'ajustement personnalisé (suite)

Proportion	Type de stockage pris en charge pour l'ajustement personnalisé	Paramètres	Limites du paramètre	Remarques
Fixe	Entier, réel, date/heure ou chaîne	Valeur		Vous ne pouvez pas spécifier la distribution Fixe pour chaque champ. Si vous voulez que chaque champ de vos données générées soit fixe, vous pouvez utiliser un noeud Utilisateur suivi d'un noeud équilibrer.
Gamma	Entier, réel ou date/heure	Forme Echelle Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Les paramètres Minimum et Maximum sont facultatifs. La distribution utilise un paramètre de taux avec un paramètre de forme égale à K et un paramètre d'échelle inversée égal à 1.
Lognormal	Entier, réel ou date/heure	Forme 1 Forme 2 Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Les paramètres Minimum et Maximum sont facultatifs.
Binomiale négative - Echecs	Entier, réel ou date/heure	Seuil Probabilité Minimum Maximum	≥ 0 0 ≤ *Probabilité* ≤ 1 < *Maximum* > *Minimum*	Les paramètres Minimum et Maximum sont facultatifs.
Binomiale négative - Essais	Entier, réel ou date/heure	Seuil Probabilité Minimum Maximum	≥ 0 0 ≤ *Probabilité* ≤ 1 < *Maximum* > *Minimum*	Les paramètres Minimum et Maximum sont facultatifs.
Normale	Entier, réel ou date/heure	Moyenne Ecart type Minimum Maximum	≥ 0 > 0 < *Maximum* > *Minimum*	Les paramètres Minimum et Maximum sont facultatifs.
Poisson	Entier, réel ou date/heure	Moyenne Minimum Maximum	≥ 0 < *Maximum* > *Minimum*	Les paramètres Minimum et Maximum sont facultatifs.
Plage	Entier, réel ou date/heure	Début(X) Fin(X) Probabilité(X)	0 ≤ *Valeur* ≤ 1	X est l'index de chaque intervalle. La somme des valeurs de probabilité doit être égale à 1.
Triangulaire	Entier, réel ou date/heure	Mode Minimum Maximum	*Minimum* ≤ *Valeur* ≤ *Maximum* < *Maximum* > *Minimum*	

Tableau 12. Distributions disponibles pour l'ajustement personnalisé (suite)

Proportion	Type de stockage pris en charge pour l'ajustement personnalisé	Paramètres	Limites du paramètre	Remarques
Uniforme	Entier, réel ou date/heure	Minimum Maximum	< *Maximum* > *Minimum*	
Weibull	Entier, réel ou date/heure	Taux Echelle Emplacement Minimum Maximum	> 0 > 0 ≥ 0 < *Maximum* > *Minimum*	Les paramètres Emplacement, Maximum et Minimum sont facultatifs.

Noeud Importation d'extension

Avec le noeud Importation d'extension, vous pouvez exécuter des scripts R ou Python for Spark pour importer des données.

Noeud Importation d'extension - Onglet Syntaxe

Sélectionnez le type de syntaxe – **R** ou **Python for Spark**. Ensuite, entrez ou collez votre script personnalisé pour importer des données. Lorsque votre syntaxe est prête, vous pouvez cliquer sur **Exécuter** pour exécuter le noeud Importation d'extension.

Noeud Importation d'extension - Onglet Sortie de la console

L'onglet **Sortie de la console** contient les sorties reçues lorsque le script R ou le script Python for Spark de l'onglet Syntaxe est exécuté (par exemple, si un script R est utilisé, il affiche la sortie reçue de la console R lorsque le script R du champ **Syntaxe R** de l'onglet **Syntaxe** est exécuté). La sortie peut contenir des messages d'erreur ou d'avertissement R ou Python générés lors de l'exécution du script R ou Python. Cette sortie permet essentiellement de déboguer le script. L'onglet **Sortie de la console** contient également le script du champ **Syntaxe R** ou **Syntaxe Python**.

A chaque exécution du script Importation d'extension, le contenu de l'onglet **Sortie de la console** est écrasé par la sortie reçue de la console R ou Python for Spark. La sortie ne peut pas être éditée.

Filtrage ou modification du nom des champs

Vous pouvez renommer ou exclure des champs à tout stade d'un flux. Par exemple, en tant que chercheur en médecine, vous n'êtes peut-être pas intéressé par le niveau de potassium (données de niveau champ) des patients (données de niveau enregistrement) ; vous pouvez donc filtrer le champ `K` correspondant. Vous pouvez réaliser ceci à l'aide d'un noeud Filtrer distinct ou d'un onglet Filtrer sur un noeud source ou de sortie. Cette fonctionnalité est identique quel que soit le noeud à partir duquel vous y accéder.

- A partir de noeuds source, tels que Délimité, Fixe, Fichier Statistiques, XML Ou importation d'extension, vous pouvez renommer ou filtrer les champs à mesure que les données sont lues dans IBM SPSS Modeler.
- Le noeud Filtrer permet de renommer ou de filtrer les champs en tout point du flux.
- A partir des noeuds Export Statistiques, Transformation Statistiques, Modèle Statistiques et Sortie Statistiques, vous pouvez filtrer ou renommer les champs pour respecter les conventions de dénomination IBM SPSS Statistics. Pour plus d'informations, voir «Changement du nom ou filtrage des champs pour IBM SPSS Statistics», à la page 383.
- Vous pouvez utiliser l'onglet Filtrer dans l'un des noeuds susmentionnés pour définir ou modifier des ensembles de réponses multiples. Pour plus d'informations, voir «Modification des ensembles de réponses multiples», à la page 161.

- Finalement, vous pouvez utiliser un noeud Filtrer pour mapper les champs entre un noeud source et un autre.

Noeud Vue de données

Utilisez le noeud Vue de données pour inclure dans votre flux les données définies dans une vue de données analytiques IBM SPSS Collaboration and Deployment Services. Une vue de données analytiques définit une structure d'accès aux données qui décrit les entités utilisées dans les modèles prédictifs et les règles métier. La vue associe la structure de données à des sources de données physiques pour l'analyse.

L'analyse prédictive requiert l'organisation de données en tables dans lesquelles chaque ligne correspond à une entité pour laquelle des prévisions sont effectuées. Chaque colonne de la table représente un attribut mesurable de l'entité. Certains attributs peuvent être dérivés par l'agrégation des valeurs d'un autre attribut. Par exemple, les lignes d'une table peuvent représenter des clients, les colonnes correspondant au nom du client, à son sexe, à son code postal et au nombre de fois où il a effectué un achat de plus de 500 € l'année dernière. La dernière colonne est dérivée de l'historique des commandes du client, qui est généralement stocké dans une ou plusieurs tables associées.

Le processus d'analyse prédictive implique l'utilisation de différents jeux de données au cours du cycle de vie d'un modèle. Lors du développement initial d'un modèle prédictif, vous utilisez des données historiques dont les résultats sont souvent connus pour l'événement prévu. Pour évaluer l'efficacité et l'exactitude du modèle, validez un modèle candidat par rapport aux différentes données. Après avoir validé le modèle, déployez-le en production pour générer des scores pour plusieurs entités dans un traitement par lots ou pour des entités uniques dans un processus en temps réel. Si vous combinez le modèle avec des règles métier dans un processus de gestion décisionnelle, utilisez les données simulées pour valider les résultats de la combinaison. Toutefois, bien que les données utilisées varient au cours des étapes du processus de développement de modèle, chaque jeu de données doit fournir le même ensemble d'attributs pour le modèle. L'ensemble d'attributs demeure constant, tandis que les enregistrements de données en cours d'analyse changent.

Une vue de données analytiques se compose des éléments suivants, qui répondent aux besoins spécifiques de l'analyse prédictive :

- Un modèle de données ou un schéma de vue de données qui définit l'interface logique permettant d'accéder aux données sous la forme d'un ensemble d'attributs organisés en tables associées. Les attributs du modèle peuvent être dérivés d'autres attributs.
- Un ou plusieurs plans d'accès aux données qui fournissent des valeurs physiques aux attributs du modèle de données. Contrôlez les données auxquelles le modèle de données a accès en indiquant quel plan d'accès aux données est actif pour une application donnée.

Important : Pour utiliser le noeud Vue de données, IBM SPSS Collaboration and Deployment Services Repository doit d'abord être installé et configuré sur votre site. La vue de données analytiques référencée par le noeud est généralement créée et stockée dans le référentiel à l'aide d'IBM SPSS Deployment Manager.

Définition des options du noeud Vue de données

Utilisez les options situées sur l'onglet **Données** de la boîte de dialogue du noeud Vue de données afin de spécifier les paramètres de données pour une vue de données analytiques sélectionnée dans IBM SPSS Collaboration and Deployment Services Repository.

Vue de données analytiques. Cliquez sur le bouton représentant des points de suspension (**...**) pour sélectionner une vue de données analytiques. Si vous n'êtes pas connecté actuellement à un serveur de référentiel, indiquez l'URL du serveur dans la boîte de dialogue Référentiel : serveur, cliquez sur **OK** et spécifiez vos identifiants de connexion dans la boîte de dialogue Référentiel : identifiants. Pour plus d'informations sur la connexion au référentiel et l'extraction d'objets, consultez le Guide d'utilisation d'IBM SPSS Modeler.

Nom de la table. Sélectionnez une table dans le modèle de données de la vue de données analytiques. Chaque table du modèle de données représente un concept ou une entité entrant en jeu dans le processus d'analyse prédictive. Les champs des tables correspondent aux attributs des entités représentées par les tables. Par exemple, si vous analysez des commandes client, votre modèle de données peut inclure une table pour les clients et une table pour les commandes. La table clients peut comporter des attributs pour l'identificateur du client, son âge, son sexe, sa situation de famille et son pays de résidence. La table commandes peut contenir des attributs pour l'identificateur de la commande, le nombre d'articles de la commande, son coût total et l'identificateur du client qui a passé la commande. L'attribut de l'identificateur client peut être utilisé pour associer les clients dans la table clients à leurs commandes dans la table commandes.

Plan d'accès aux données. Sélectionnez un plan d'accès aux données dans la vue de données analytiques. Un plan d'accès aux données associe les tables du modèle de données figurant dans une vue de données analytiques à des sources de données physiques. Une vue de données analytiques contient généralement plusieurs plans d'accès aux données. Lorsque vous modifiez le plan d'accès aux données utilisé, vous modifiez également les données utilisées par votre flux. Par exemple, si la vue de données analytiques contient un plan d'accès aux données pour la formation d'un modèle et un plan d'accès aux données pour le test d'un modèle, vous pouvez passer des données de formation aux données de test en modifiant le plan d'accès aux données utilisé.

Attributs facultatifs. Si un attribut spécifique n'est pas requis par l'application utilisant la vue de données analytiques, vous pouvez marquer l'attribut comme facultatif. Contrairement aux attributs obligatoires, les attributs facultatifs peuvent inclure des valeurs nulles. Vous devrez peut-être ajuster votre application afin d'inclure le traitement des valeurs nulles pour les attributs facultatifs. Par exemple, lors de l'appel d'une règle métier créée dans IBM Operational Decision Manager, IBM Analytical Decision Management demande au service de règle de déterminer les entrées obligatoires. Si l'enregistrement à évaluer contient une valeur nulle pour l'un des champs obligatoires du service de règle, la règle n'est pas appelée et les champs de sortie de la règle sont renseignés à l'aide des valeurs par défaut. Si un champ facultatif contient une valeur nulle, la règle est appelée. La règle peut rechercher les valeurs nulles pour contrôler le traitement.

Pour spécifier des attributs comme facultatifs, cliquez sur **Attributs facultatifs** et sélectionnez les attributs facultatifs.

Inclure les données XML dans la zone. Sélectionnez cette option pour créer un champ contenant les données XML du modèle d'objet exécutable pour chaque ligne de données. Ces informations sont requises si les données doivent être utilisées avec IBM Operational Decision Manager. Indiquez le nom de ce nouveau champ.

Noeud source Géospatial

Vous utilisez le noeud source Géospatial pour intégrer des données de carte ou spatiales dans votre session d'exploration de données. Vous pouvez importer des données de l'une des façons suivantes :

- Dans un fichier de forme (`.shp`)
- En vous connectant à un serveur ESRI qui contient un système hiérarchique de fichiers incluant des fichiers de carte.

  Remarque : Vous pouvez uniquement vous connecter à des services de carte publics.

Les modèles de prévision spatio-temporelle (STP) peuvent inclure des éléments de carte ou spatiaux dans leurs prévisions. Pour plus d'informations sur ces modèles, voir la rubrique intitulée "Noeud de modélisation Prévision spatio-temporelle" dans la section Modèles de séries temporelles du guide des noeuds de modélisation Modeler (ModelerModelingNodes.pdf).

Définition des options pour le noeud source Géospatial

Type de source de données Vous pouvez importer des données depuis un **fichier de forme** (`.shp`) ou vous connecter à un **service de carte**.

Si vous utilisez un **fichier de forme**, entrez le nom de fichier et son chemin d'accès ou recherchez-le. Le fichier doit se trouver dans le répertoire local ou être accessible à partir d'une unité mappée. Vous ne pouvez pas accéder au fichier en utilisant un chemin UNC (convention de dénomination universelle).

Remarque : Les données de forme requièrent un fichier `.shp` et un fichier `.dbf`. Les deux fichiers doivent avoir le même nom et se trouver dans le même dossier. Le fichier `.dbf` est importé automatiquement lorsque vous sélectionnez le fichier `.shp`. De plus, un fichier `.prj` peut spécifier le système de coordonnées pour les données de forme.

Si vous utilisez un **service de carte**, entrez l'adresse URL du service et cliquez sur **Connecter**. Une fois que vous êtes connecté au service, les couches de ce service sont affichées au bas de la boîte de dialogue dans une arborescence dans le panneau **Cartes disponibles** ; développez l'arborescence et sélectionnez la couche de votre choix.

Remarque : Vous pouvez uniquement vous connecter à des services de carte publics.

Définition automatique des données géospatiales

Par défaut, SPSS Modeler définit automatiquement, si possible, les champs de données géospatiales sur le noeud source avec les métadonnées correctes. Les métadonnées peuvent inclure le niveau de mesure du champ géospatial (comme Point ou Polygone) et le système de coordonnées qui est utilisé par les champs, notamment des détails tels que le point d'origine (par exemple latitude 0, longitude 0) et les unités de mesure. Pour plus d'informations sur les niveaux de mesure, voir «Sous-niveaux de mesure géospatiaux», à la page 147.

Les fichiers `.shp` et `.dbf` qui constituent le fichier de forme contiennent un champ d'identificateur commun qui est utilisé comme clé. Par exemple, le fichier `.shp` peut contenir des pays, auquel cas le champ de nom de pays est utilisé comme identificateur, et le fichier `.dbf` peut contenir des informations sur ces pays, auquel cas le nom du pays est également utilisé comme identificateur.

Remarque : Si le système de coordonnées n'est pas identique au système de coordonnées SPSS Modeler par défaut, il peut être nécessaire de reprojeter les données afin d'utiliser le système de coordonnées requis. Pour plus d'informations, voir «Noeud de reprojection», à la page 195.

Onglets communs des noeuds source

Les options suivantes peuvent être spécifiées pour tous les noeuds source en cliquant sur l'onglet correspondant :

- **Onglet Données.** Onglet Données.Permet de modifier le type de stockage par défaut.
- **Onglet Filtrer.** Permet d'éliminer ou de renommer des champs de données. Cet onglet offre les mêmes fonctions que le noeud Filtrer. Pour plus d'informations, voir «Définition des options de filtrage», à la page 159.
- **Onglet Types.** Permet de définir les niveaux de mesure. Cet onglet offre les mêmes fonctions que le noeud Typer.
- **Onglet Annotations.** Utilisé pour tous les noeuds, cet onglet propose des options permettant de renommer les noeuds, de créer des info-bulles personnalisées et de stocker de longues annotations.

Définition des niveaux de mesure dans le noeud source

Les propriétés de champ peuvent être indiquées dans un noeud source ou dans un noeud Typer distinct. Les fonctionnalités sont similaires dans les deux noeuds. Les propriétés suivantes sont disponibles :

- **Champ** Cliquez deux fois sur un nom de champ pour spécifier des libellés de valeur et de champ pour les données dans IBM SPSS Modeler. Par exemple, vous pouvez consulter ou modifier ici les métadonnées de champ importées à partir de IBM SPSS Statistics. De même, vous pouvez créer des libellés pour les champs et leurs valeurs. La présence des libellés indiqués ici dans IBM SPSS Modeler dépend des sélections effectuées dans la boîte de dialogue Propriétés du flux.
- **Mesure** Il s'agit du niveau de mesure utilisé pour décrire les caractéristiques des données dans un champ précis. Si tous les détails d'un champ sont connus, il est dit **complètement instancié**. Pour plus d'informations, voir «Niveaux de mesure», à la page 145.

  Remarque : Le niveau de mesure d'un champ est différent de son type de stockage, qui indique si les données sont stockées sous forme de chaînes, d'entiers, de nombres réels, de dates, d'heures, d'horodatages ou de listes.
- **Valeurs** Cette colonne vous permet de spécifier des options pour la lecture des valeurs de données depuis le jeu de données ou d'utiliser l'option **Spécifier** afin de spécifier des niveaux de mesure et des valeurs dans une boîte de dialogue distincte. Vous pouvez également choisir de transférer les champs sans lire leurs valeurs. Pour plus d'informations, voir «Valeurs de données», à la page 150.

  Remarque : Vous ne pouvez pas modifier la cellule dans cette colonne si l'entrée **Champ** correspondante contient une liste.
- **Manquantes** Utilisé pour spécifier la façon dont les valeurs manquantes pour le champ sont traitées. Pour plus d'informations, voir «Définition de valeurs manquantes», à la page 155.

  Remarque : Vous ne pouvez pas modifier la cellule dans cette colonne si l'entrée **Champ** correspondante contient une liste.
- **Vérifier** Dans cette colonne, vous pouvez définir des options pour vous assurer que les valeurs de champ sont conformes aux valeurs ou plages spécifiées. Pour plus d'informations, voir «Vérification des valeurs de type», à la page 155.

  Remarque : Vous ne pouvez pas modifier la cellule dans cette colonne si l'entrée **Champ** correspondante contient une liste.
- **Rôle** Utilisé pour indiquer aux noeuds de modélisation si les champs sont des champs d'**entrée** (champs prédicteurs) ou **cible** (champs prédits) pour un processus d'apprentissage automatique). Sont également disponibles les rôles **Les deux** et **Aucun**, et l'option **Partition**. Cette dernière signale les champs utilisés pour partitionner les enregistrements en échantillons distincts à des fins d'apprentissage, de test et de validation. La valeur **Diviser** spécifie que des modèles séparés seront construits pour chaque valeur possible du champ. Pour plus d'informations, voir «Définition du rôle de champ», à la page 155.

Pour plus d'informations, voir «Noeud Typer», à la page 144.

A quel moment procéder à l'instanciation au niveau du noeud source ?

Vous pouvez obtenir des informations sur le stockage et les valeurs de données de vos champs de deux façons différentes. Cette **instanciation** peut se produire au niveau du noeud source lorsque vous introduisez des données pour la première fois dans IBM SPSS Modeler, ou lorsque vous ajoutez un noeud Typer dans le flux de données.

L'instanciation au niveau du noeud source est utile dans les cas suivants :

- L'jeu de données n'est pas volumineux.
- Vous prévoyez de calculer de nouveaux champs à l'aide du Générateur de formules (l'instanciation rend les valeurs des champs disponibles à partir du Générateur de formules).

En général, si votre jeu de données n'est pas trop volumineux et si vous ne prévoyez pas d'ajouter des champs au flux par la suite, l'instanciation au niveau du noeud source est la méthode la plus pratique.

Filtrage des champs à partir du noeud source

L'onglet Filtrer de la boîte de dialogue d'un noeud source vous permet d'exclure des champs des opérations en aval en fonction de votre examen initial des données. Ceci est utile, par exemple, s'il existe des champs en double dans les données ou si vous êtes déjà suffisamment familiarisé avec les données pour exclure les champs qui ne sont pas pertinents. Vous pouvez également ajouter ultérieurement au flux un noeud Filtrer distinct. Les fonctionnalités sont similaires dans les deux cas. Pour plus d'informations, voir «Définition des options de filtrage», à la page 159.

Chapitre 3. Noeuds d'opérations sur les lignes

Présentation des noeuds d'opérations sur les lignes

Les noeuds d'opérations sur les lignes permettent d'apporter des changements aux enregistrements de données. Ces opérations sont importantes durant les phases de **compréhension des données** et de **préparation des données** de l'exploration de données parce qu'elles vous permettent d'adapter les données à vos besoins métier.

Par exemple, selon les résultats de l'audit que vous avez mené à l'aide du noeud Audit données (palette Sortie), vous pouvez décider de fusionner les enregistrements achat client des trois derniers mois. Le noeud Fusionner permet de fusionner les enregistrements en fonction des valeurs d'un champ-clé, par exemple *ID client*. Vous pouvez également constater qu'une base de données d'informations relatives à la fréquentation d'un site Web devient impossible à gérer lorsqu'elle comporte plus d'un million d'enregistrements. Dans ce cas, utilisez un noeud Echantillonner pour sélectionner un sous-ensemble de données à utiliser lors de la modélisation.

La palette Opérations sur les lignes contient les noeuds suivants :

Le noeud Sélectionner permet de sélectionner ou d'exclure des sous-ensembles d'enregistrements d'un flux de données sur la base d'une condition spécifique. Par exemple, vous pouvez sélectionner les enregistrements qui appartiennent à un secteur de ventes particulier.

Le noeud Echantillonner sélectionne un sous-ensemble d'enregistrements. Divers types d'échantillons sont pris en charge, notamment les échantillons stratifiés, en cluster et non aléatoires (structurés). L'échantillonnage peut être utile pour améliorer les performances et pour sélectionner des groupes d'enregistrements associés ou des transactions pour analyse.

Le noeud Equilibrer corrige les déséquilibres survenant dans un jeu de données, de manière à respecter une condition précise. La règle d'équilibrage ajuste la proportion d'enregistrements présentant une condition True (vrai) par rapport au facteur indiqué.

Le noeud Agréger remplace une séquence d'enregistrements d'entrée par des enregistrements de sortie abrégés et agrégés.

Le noeud agrégé Recency, Frequency, Monetary (RFM) vous permet de prendre les données de l'historique des transactions d'un client, d'en éliminer les éventuelles données inutilisées et de combiner le reste des données de transaction sur une seule ligne qui indique la date de la dernière consultation, le nombre de transactions réalisées et la valeur monétaire totale de ces transactions.

Le noeud Trier trie les enregistrements par ordre croissant ou décroissant, en fonction de la valeur d'un ou de plusieurs champs.

Le noeud Fusionner permet de créer, à partir de plusieurs enregistrements d'entrée, un seul enregistrement de sortie contenant tout ou partie des champs d'entrée. Il sert notamment à fusionner des données provenant de différentes sources, telles que les données client internes et les données démographiques acquises.

Le noeud Ajouter réalise la concaténation d'ensembles d'enregistrements. Il permet de combiner des jeux de données dont les structures sont similaires, mais les données différentes.

Le noeud Distinguer supprime les enregistrements en double, soit en incluant le premier enregistrement dans le flux de données, soit en le supprimant et en incluant ses doublons dans le flux de données.

Le noeud Streaming Time Series génère et évalue les modélisations des séries chronologiques en une seule étape. Vous pouvez l'utiliser avec des données dans un environnement ou distribué ; dans un environnement distribué, vous pouvez exploiter la puissance d'IBM SPSS Analytic Server

Les noeuds Boîtes espace-temps constituent une extension des emplacements spatiaux avec Geohash. Un noeud Boîtes espace-temps est en particulier une chaîne alphanumérique qui représente une zone d'espace et de temps de forme régulière.

Le noeud Streaming TCM génère et évalue les modèles de causalité temporelle en une étape.

Le noeud Optimisation CPLEX permet d'utiliser l'optimisation mathématique complexe (CPLEX) via un fichier de modèle OPL (Optimization Programming Language). Cette fonctionnalité est disponible dans le produit IBM Analytical Decision Management, mais désormais, vous pouvez également utiliser le noeud CPLEX dans SPSS Modeler sans IBM Analytical Decision Management.

Pour plus d'informations sur l'optimisation CPLEX et OPL, consultez la documentation d'IBM Analytical Decision Management.

La plupart des noeuds de la palette Opérations sur les lignes nécessitent l'utilisation d'expressions CLEM. Si vous connaissez CLEM, vous pouvez saisir une expression dans le champ. Chaque champ d'expression comporte toutefois un bouton permettant d'ouvrir le Générateur de formules CLEM ; ce dernier vous aide à créer automatiquement de telles expressions.

Figure 1. Bouton du Générateur de formules

Noeud Sélectionner

Ce noeud permet de sélectionner ou de supprimer un sous-ensemble d'enregistrements du flux de données en fonction d'une condition spécifique, du type `TA` (tension artérielle) `== ELEVEE`.

Mode. Indique si les enregistrements répondant à la condition seront inclus dans le flux de données ou s'ils en seront exclus.

- **Enlever.** Permet d'inclure les enregistrements qui répondent à la condition de sélection.
- **Annuler.** Permet d'exclure les enregistrements qui répondent à la condition de sélection.

Condition. Affiche la condition de sélection, spécifiée à l'aide d'une expression CLEM, qui sera utilisée pour tester les enregistrements. Entrez une expression dans la fenêtre ou utilisez le Générateur de formules en cliquant sur le bouton en forme de calculatrice situé à droite de la fenêtre.

Si vous choisissez d'ignorer des enregistrements en fonction d'une condition, comme dans l'exemple suivant :

```
(var1='value1' and var2='value2')
```

le noeud Sélectionner par défaut ignore également les enregistrements ayant des valeurs nulles pour tous les champs de sélection. Pour éviter cela, ajoutez la condition suivante à la condition d'origine :

```
and not(@NULL(var1) and @NULL(var2))
```

Les noeuds Sélectionner sont également utilisés pour choisir une proportion d'enregistrements. Normalement, cette opération est effectuée à l'aide d'un noeud Echantillonner. Cependant, si les paramètres disponibles ne sont pas adaptés à la complexité de la condition que vous souhaitez spécifier, vous pouvez créer cette dernière à l'aide d'un noeud Sélectionner. Une condition semblable à la suivante peut être créée :

```
BP = "HIGH" and random(10) <= 4
```

Avec cette condition, environ 40 % des enregistrements présentant une tension artérielle élevée seront sélectionnés et transmis aux noeuds en aval pour être analysés plus en détail.

Noeud Echantillonner

Vous pouvez utiliser des noeuds Echantillonner pour sélectionner un sous-groupe d'enregistrements à analyser ou définir une proportion d'enregistrements à supprimer. Divers types d'échantillons sont pris en charge, notamment les échantillons stratifiés, en cluster et non aléatoires (structurés). Vous pouvez utiliser l'échantillonnage à diverses fins :

- Pour améliorer les performances en évaluant les modèles d'un sous-groupe de données. Les modèles évalués à partir d'un échantillon sont souvent aussi précis que ceux issus du jeu de données complet et plus encore si l'amélioration des performances permet de tester différentes méthodes que vous ne testeriez normalement pas.

- Pour sélectionner des groupes d'enregistrements ou de transactions associés à analyser, tels que tous les articles d'un panier en ligne ou toutes les propriétés d'un voisinage donné.
- Pour identifier des unités ou des observations pour une vérification aléatoire pour le contrôle de qualité, la prévention des fraudes ou la sécurité.

Remarque : si vous souhaitez simplement diviser les données dans des échantillons d'apprentissage et de test à des fins de validation, vous pouvez utiliser un noeud Partitionner à la place. Pour plus d'informations, voir «Noeud Partitionner», à la page 185.

Types d'échantillons

Echantillons en cluster. Echantillonnent des groupes ou des clusters et non des unités individuelles. Supposons que vous disposiez d'un fichier de données comportant un enregistrement pour chaque élève. Si vous classez en fonction de l'école et que la taille de l'échantillon est 50 %, 50 % des écoles seront choisies et tous les élèves de chacune des écoles sélectionnées seront sélectionnés. Les élèves des écoles non sélectionnées seront alors rejetés. En moyenne, 50 % des élèves devraient être sélectionnés, mais étant donné que les écoles ont des tailles différentes, le pourcentage peut ne pas être exact. De même, vous pouvez classer les articles d'un panier en fonction de l'ID de la transaction pour pouvoir conserver tous les articles des transactions sélectionnées. Pour un exemple de classement des propriétés en fonction de la ville, voir le flux d'échantillon *complexsample_property.str*.

Echantillons stratifiés. Sélectionnent les échantillons indépendamment dans des sous-groupes sans chevauchement de population, ou strates. Vous pouvez, par exemple, faire en sorte que tous les hommes et femmes soient échantillonnés dans des proportions égales ou que chaque région ou groupe socio-économique d'une population soient représentés. Vous pouvez également définir une taille d'échantillon différente pour chaque strate (par exemple, si vous pensez qu'un groupe est sous-représenté dans les données d'origine). Pour un exemple de stratification des propriétés en fonction du pays, voir le flux d'échantillon *complexsample_property.str*.

Echantillonnage systématique ou 1 en n Lorsque la sélection aléatoire est difficile à obtenir, vous pouvez échantillonner les unités de manière systématique (à une fréquence fixe) ou de manière séquentielle.

Pondérations d'échantillonnage. Des pondérations d'échantillonnage sont calculées automatiquement lors de la création du graphique d'un échantillon complexe et elles correspondent approximativement à "l'effectif" que chaque unité échantillonnée représente dans les données d'origine. Par conséquent, la somme des pondérations sur l'échantillon doit évaluer la taille des données d'origine.

Cadre d'échantillonnage

Un cadre d'échantillonnage définit la source des observations potentielles à inclure dans un échantillon ou une étude. Dans certains cas, il peut être possible d'identifier chaque membre d'une population et d'inclure n'importe quel membre dans un échantillon, par exemple, lors de l'échantillonnage des éléments qui proviennent d'une chaîne de production. Dans la plupart des cas, vous ne pourrez pas accéder à chacune des observations possibles. Par exemple, vous ne pouvez pas savoir qui va voter dans une élection tant que l'élection n'a pas eu lieu. Dans ce cas, vous pouvez utiliser le registre des inscrits comme cadre d'échantillonnage, même si certaines personnes inscrites ne voteront pas, sachant que des personnes peuvent voter bien qu'elles ne figurent pas dans le registre au moment où vous vérifiez le registre. Toute personne ne figurant pas dans le cadre d'échantillonnage ne peut pas être échantillonnée. La représentation de la population à évaluer par le cadre d'échantillonnage doit être traitée pour chaque observation réelle.

Options de noeud échantillon

Vous pouvez choisir la méthode **simple** ou **complexe** en fonction de vos besoins.

Options d'échantillonnage simples

La méthode simple permet de sélectionner un pourcentage aléatoire d'enregistrements, de sélectionner des enregistrements contigus ou de sélectionner chaque *nième* enregistrement.

Mode. Choisissez de transmettre (inclure) ou de supprimer (exclure) les enregistrements pour les modes suivants :

- **Inclure l'échantillon.** Inclut les enregistrements sélectionnés dans le flux de données et supprime tous les autres. Par exemple, si vous définissez le mode sur **Inclure l'échantillon** et l'option **1 en n** sur la valeur 5, un enregistrement sur cinq sera inclus pour donner un jeu de données dont la taille est égale à environ un cinquième de la taille d'origine. Il s'agit du mode par défaut de l'échantillonnage des données et du seul mode disponible avec la méthode complexe.
- **Retirer l'échantillon.** Exclut les enregistrements sélectionnés et inclut tous les autres. Par exemple, si vous définissez le mode sur **Retirer l'échantillon** et l'option **1 en n** sur la valeur 5, un enregistrement sur cinq sera exclu. Ce mode est disponible uniquement avec la méthode simple.

Echantillon. Sélectionnez la méthode d'échantillonnage à partir des options suivantes :

- **Premier.** Permet d'utiliser l'échantillonnage de données adjacentes. Si, par exemple, la taille d'échantillon maximale est 10 000, les 10 000 premiers enregistrements seront sélectionnés.
- **Tous les.** Sélectionnez cette option pour échantillonner les données en incluant ou en excluant un enregistrement sur *n*. Si, par exemple, *n* a la valeur 5, un enregistrement sur 5 sera sélectionné.
- **% aléatoire.** Permet d'échantillonner un pourcentage aléatoire de données. Par exemple, si vous indiquez la valeur 20, 20 % des données seront incluses dans le flux de données ou en seront exclues, selon le mode sélectionné. Dans le champ, indiquez le pourcentage d'échantillonnage. Vous pouvez également spécifier une valeur de départ à l'aide de l'option **Définir une valeur de départ aléatoire**.

  Utiliser l'échantillonnage des niveaux de bloc (dans la base de données uniquement). Cette option n'est activée que si vous choisissez un échantillonnage de pourcentage aléatoire lors de l'exploration d'une base de données Oracle ou IBM Db2. Dans ce cas, l'échantillonnage des niveaux de bloc peut être plus efficace.

  Remarque : Vous n'obtenez pas un nombre exact de lignes renvoyées chaque fois que vous exécutez les mêmes paramètres d'échantillonnage aléatoires. Cela est dû au fait que chaque enregistrement d'entrée a une probabilité de `N/100` d'être inclus dans l'échantillon (où `N` est le **% aléatoire** que vous spécifiez dans le noeud) et que les probabilités sont indépendantes. Par conséquent, les résultats ne sont pas exactement `N%`.

Taille maximale de l'échantillon. Indique le nombre maximum d'enregistrements à inclure dans l'échantillon. Cette option est redondante et par conséquent désactivée si **Premiers** et **Inclure** sont sélectionnés. Notez également que lorsqu'il est utilisé avec l'option **% aléatoire**, ce paramètre peut vous empêcher de sélectionner certains enregistrements. Par exemple, si vous disposez de 10 millions d'enregistrements dans votre jeu de données et que vous sélectionnez 50 % d'enregistrements avec une taille maximale d'échantillon de 3 millions d'enregistrements, 50 % des 6 premiers millions d'enregistrements seront sélectionnés et les quatre millions d'enregistrements restants ne le seront pas. Pour éviter cette limitation, sélectionnez la méthode d'échantillonnage **complexe** et demandez un échantillon aléatoire de trois millions d'enregistrements sans définir une variable de cluster ou de strate.

Options d'échantillonnage complexes

Les options d'échantillonnage complexe permettent d'affiner le contrôle de l'échantillon, notamment des échantillons en cluster, stratifiés et pondérés avec d'autres options.

Classer et stratifier. Permet de définir des champs de classification, de stratification et de pondération d'entrée, si nécessaire. Pour plus d'informations, voir «Paramètres de classification et de stratification», à la page 80.

Type d'échantillon.

- **Aléatoire.** Sélectionne des clusters ou des enregistrements dans chaque strate de manière aléatoire.
- **Systématique.** Sélectionne des enregistrements à une fréquence fixe. Cette option fonctionne comme la méthode *1 en n*, sauf que la position du premier enregistrement change en fonction d'une valeur de départ aléatoire. La valeur *n* est définie automatiquement en fonction de la taille ou de la proportion d'échantillonnage.

Unité d'échantillonnage. Vous pouvez sélectionner des proportions ou des nombres comme unité d'échantillonnage de base.

Taille d'échantillon. Vous pouvez définir la taille d'échantillonnage de différentes manières :

- **Colonne fixe.** Permet de définir la taille globale de l'échantillon sous la forme d'un nombre ou d'une proportion.
- **Personnalisé.** Permet de définir la taille d'échantillonnage de chaque sous-groupe ou strate. Cette option est disponible uniquement si un champ de stratification a été défini dans la sous-boîte de dialogue Classer et stratifier.
- **Variable.** Permet à l'utilisateur de sélectionner un champ qui définit la taille d'échantillonnage de chaque sous-groupe ou strate. Ce champ doit avoir la même valeur pour chaque enregistrement d'une strate. Si, par exemple, l'échantillon est stratifié en fonction du pays, tous les enregistrements ayant *county = Surrey* doivent avoir la même valeur. Le champ doit être numérique et sa valeur doit correspondre à l'unité d'échantillonnage sélectionnée. Pour les proportions, les valeurs doivent être supérieures à 0 et inférieures à 1. Pour les nombres, la valeur minimale est 1.

Echantillon minimum par strate. Définit le nombre minimum d'enregistrements (ou le nombre minimum de clusters si un champ de cluster est défini).

Echantillon maximum par strate. Définit le nombre maximum d'enregistrements ou de clusters. Si vous sélectionnez cette option sans définir un champ de cluster ou de strate, un échantillon aléatoire ou systématique de la taille définie est sélectionné.

Définir une valeur de départ aléatoire. Lors de l'échantillonnage ou du partitionnement d'enregistrements en fonction d'un pourcentage aléatoire, cette option vous permet de dupliquer les mêmes résultats dans une autre session. Indiquez la valeur de départ utilisée par le générateur de nombres aléatoires pour vous assurer que les mêmes enregistrements sont affectés à chaque exécution du noeud. Entrez la valeur de départ souhaitée ou cliquez sur le bouton **Générer** pour générer automatiquement une valeur aléatoire. Si cette option n'est pas sélectionnée, un échantillon différent est généré à chaque exécution du noeud.

Remarque : Lorsque vous utilisez l'option **Définir une valeur de départ aléatoire** avec des enregistrements lus à partir d'une base de données, il peut s'avérer nécessaire d'exécuter un noeud Trier avant de procéder à l'échantillonnage afin de garantir le même résultat à chaque exécution du noeud. Cela s'explique par le fait que la valeur de départ aléatoire dépend de l'ordre des enregistrements, et qu'il n'est pas garanti que cet ordre reste inchangé dans une base de données relationnelle. Pour plus d'informations, voir «Noeud Trier», à la page 88.

Paramètres de classification et de stratification

La boîte de dialogue Classer et stratifier permet de sélectionner des champs de cluster, de stratification et de pondération lors de la création d'un graphique d'échantillon complexe.

Clusters. Définit un champ catégoriel utilisé pour classer les enregistrements. Les enregistrements sont échantillonnés en fonction de leur appartenance aux clusters, certains clusters étant inclus et d'autres exclus. Toutefois, si un enregistrement d'un cluster est inclus, tous les enregistrements sont inclus. Lors de l'analyse des associations de produits d'un panier, par exemple, vous pouvez classer les articles en fonction de l'ID de transaction pour que tous les articles des transactions sélectionnées soient préservés.

Au lieu d'échantillonner les enregistrements, ce qui détruirait les informations sur les articles vendus ensemble, vous pouvez échantillonner les transactions pour que tous les enregistrements des transactions sélectionnées soient préservés.

Stratifier par. Définit un champ catégoriel utilisé pour stratifier les enregistrements pour que les échantillons soient sélectionnés de manière indépendante dans les sous-groupes sans chevauchement de population, ou strates. Si vous sélectionnez un échantillon de 50 % stratifié en fonction du sexe, par exemple, deux échantillons de 50 % sont utilisés, un pour les hommes et un autre pour les femmes. Les strates, par exemple, peuvent être des groupes socio-économiques, des catégories d'emplois ou des groupes ethniques permettant de disposer de tailles d'échantillons adéquates pour les sous-groupes d'intérêt. S'il existe trois fois plus de femmes que d'hommes dans le jeu de données d'origine, ce rapport est conservé en échantillonnant séparément depuis chaque groupe. Vous pouvez également définir plusieurs champs de stratification (par exemple, échantillonnage de lignes de produits dans les régions ou vice versa).

Remarque : Si vous stratifiez les données en fonction d'un champ ayant des valeurs manquantes (valeurs nulles ou système manquantes, chaînes vides, espaces et blancs ou valeurs définies par l'utilisateur manquantes), vous ne pouvez pas définir des tailles d'échantillons personnalisées pour les strates. Si vous voulez utiliser des tailles d'échantillon personnalisées lors de la stratification en fonction d'un champ ayant des valeurs manquantes ou vides, vous devez les définir en amont.

Utiliser la pondération d'entrée. Définit un champ utilisé pour pondérer les enregistrements avant l'échantillonnage. Si, par exemple, le champ de pondération a des valeurs comprises entre 1 et 5, les enregistrements pondérés 5 ont cinq fois plus de chance d'être sélectionnés. Les valeurs de ce champ sont remplacées par les pondérations d'entrée finales générées par le noeud (voir le paragraphe suivant).

Nouvelle pondération de sortie. Définit le nom du champ où les pondérations finales sont écrites si aucun champ de pondération d'entrée n'est défini. (Si un champ de pondération d'entrée est défini, ses valeurs sont remplacées par les pondérations finales comme indiqué ci-dessus, et aucun champ de pondération de sortie distinct n'est créé.) Les valeurs de pondération de sortie indiquent le nombre d'enregistrements représentés par chaque enregistrement échantillonné dans les données d'origine. La somme des valeurs de pondération donne l'estimation de la taille d'échantillon. Si, par exemple, un échantillon aléatoire de 10 % est utilisé, la pondération de sortie est 10 pour tous les enregistrements, indiquant que chaque enregistrement échantillonné représente environ dix enregistrements dans les données d'origine. Dans un échantillon stratifié ou pondéré, les valeurs de pondération de sortie peuvent varier en fonction de la proportion d'échantillonnage de chaque strate.

Commentaires

- L'échantillonnage en cluster est utile si vous ne pouvez pas obtenir la liste complète de la population à échantillonner, mais pouvez obtenir les listes complètes de certains groupes ou clusters. Il est également utilisé lorsqu'un échantillonnage aléatoire produit une liste de sujets de test difficile à contacter. Par exemple, il est plus simple de rendre visite à tous les fermiers d'un pays qu'à des fermiers dispersés dans le pays.
- Vous pouvez définir des champs de classification et de stratification pour échantillonner des clusters de manière indépendante dans chaque strate. Par exemple, vous pouvez échantillonner les valeurs de propriétés stratifiées en fonction du pays et effectuer une classification en fonction de la ville dans chaque pays. Ainsi, vous pouvez créer un échantillon indépendant des villes dans chaque pays. Certaines villes seront incluses et d'autres pas, mais pour chaque ville incluse, toutes les propriétés de la ville seront incluses.
- Pour sélectionner un échantillon aléatoire d'unités dans chaque cluster, vous pouvez enchaîner deux noeuds Echantillonner. Par exemple, vous pouvez échantillonner en premier les villes stratifiées en fonction du pays, comme indiqué ci-dessus, puis lier un second noeud Echantillonner et sélectionner *ville* comme champ de stratification, ce qui permet d'échantillonner une proportion d'enregistrements dans chaque ville.

- Si une combinaison de champs est nécessaire pour identifier de manière unique les clusters, vous pouvez générer un nouveau champ en utilisant un noeud Calculer. Si, par exemple, plusieurs boutiques utilisent le même système de numérotation des transactions, vous pouvez calculer un nouveau champ qui concatène les ID de boutique et de transaction.

Tailles d'échantillons des strates

Lorsque vous créez un échantillon stratifié, l'option par défaut consiste à échantillonner la même proportion d'enregistrements ou de clusters pour chaque strate. Si un groupe contient plus de membres qu'un autre par un facteur de 3, par exemple, vous voulez généralement préserver le même rapport dans l'échantillon. Si tel n'est pas le cas, vous pouvez définir la taille d'échantillon séparément pour chaque strate.

La boîte de dialogue Tailles d'échantillons des strates contient les valeurs du champ de stratification permettant de remplacer la valeur par défaut de la strate. S vous sélectionnez plusieurs champs de stratification, toutes les combinaisons de valeurs possibles sont affichées pour vous permettre de définir la taille de chaque groupe ethnique dans chaque ville, par exemple, ou chaque ville dans chaque pays. Les tailles sont définies sous forme de proportions ou de nombres, tel que défini par le paramètre en cours dans le noeud Echantillonner.

Pour définir les tailles d'échantillons des strates

1. Dans le noeud Echantillonner, sélectionnez **Complexe** et un ou plusieurs champs de stratification. Pour plus d'informations, voir «Paramètres de classification et de stratification», à la page 80.
2. Sélectionnez **Personnaliser** et **Définir des tailles**.
3. Dans la boîte de dialogue Tailles d'échantillons des strates, cliquez sur le bouton **Lire les valeurs** dans la partie inférieure gauche de l'écran. Si nécessaire, vous pouvez être amené à instancier des valeurs dans une source en amont ou un noeud Typer. Pour plus d'informations, voir «Qu'est-ce que l'instanciation ?», à la page 149.
4. Cliquez sur une ligne pour remplacer la taille par défaut de la strate.

Remarques sur les tailles d'échantillons

Des tailles d'échantillons personnalisées peuvent s'avérer utiles si des strates différentes ont des variances différentes, par exemple, pour que les tailles d'échantillons soient proportionnelles à l'écart-type. (Si les observations dans la strate varie davantage, vous devez les échantillonner davantage pour obtenir un échantillon représentatif.) Ou bien, si une strate est petite, vous pouvez utiliser une proportion d'échantillonnage supérieure pour inclure un nombre minimum d'observations.

Remarque : Si vous stratifiez les données en fonction d'un champ ayant des valeurs manquantes (valeurs nulles ou système manquantes, chaînes vides, espaces et blancs ou valeurs définies par l'utilisateur manquantes), vous ne pouvez pas définir des tailles d'échantillons personnalisées pour les strates. Si vous voulez utiliser des tailles d'échantillon personnalisées lors de la stratification en fonction d'un champ ayant des valeurs manquantes ou vides, vous devez les définir en amont.

Noeud Equilibrer

Le noeud Equilibrer permet de corriger les déséquilibres dans les jeux de données, de sorte que ceux-ci soient conformes aux critères de test spécifiés. Supposons, par exemple, qu'un jeu de données présente uniquement deux valeurs, *faible* ou *élevée*, et que 90 % des occurrences sont *faibles* tandis que seulement 10 % des occurrences sont *élevées*. De nombreuses techniques de modélisation ne parviennent pas à gérer ce type de données biaisées car elles ont tendance à ne retenir que la valeur *faible* et à ignorer la valeur *élevée*, qui est plus rare. Si les données sont équilibrées, avec des nombres approximativement égaux de valeurs *faibles* et *élevées*, les modèles pourront plus facilement trouver des tendances qui distinguent les deux groupes. Dans ce cas, vous pouvez utiliser un noeud Equilibrer pour créer une directive qui diminue le nombre d'observations de la valeur *faible*.

L'équilibrage est obtenu en dupliquant et en supprimant des enregistrements en fonction de conditions spécifiées. Les enregistrements pour lesquels aucune condition n'est vérifiée sont toujours ignorés. Dans la mesure où ce processus implique la duplication et/ou l'exclusion d'enregistrements, la séquence d'origine de vos données est perdue au cours d'opérations effectuées en aval. Veillez à calculer toutes les valeurs dépendant directement de la séquence de vos données avant d'ajouter un noeud Equilibrer au flux de données.

Remarque : les noeuds Equilibrer peuvent être automatiquement générés à partir de graphiques de distribution et Histogramme. Vous pouvez, par exemple, équilibrer les données pour indiquer les proportions égales dans toutes les catégories d'un champ catégoriel, comme indiqué dans une courbe de distribution.

Exemple. Lors de la création d'un flux RFM pour identifier les clients récents qui ont répondu positivement à des campagnes de publicité antérieures, le service Marketing d'une société utilise un noeud Equilibrer pour équilibrer les différences entre les réponses vraies et les réponses fausses dans les données.

Définition des options du noeud Equilibrer

Règles d'équilibrage d'enregistrements. Affiche les règles d'équilibrage en cours. Chaque directive inclut à la fois un facteur et une condition qui indiquent au logiciel "d'augmenter la proportion d'enregistrements d'un facteur spécifié lorsque la condition est vraie". Un facteur inférieur à 1,0 signifie que la proportion d'enregistrements va être réduite. Par exemple, si vous souhaitez réduire le nombre d'enregistrements pour lesquels le médicament Y est utilisé, vous pouvez créer une règle d'équilibrage avec un facteur de 0,7 et la condition `Médicament = "médY"`. Cette directive indique que le nombre d'enregistrements pour lesquels le médicament Y est utilisé sera réduit de 70 % pour toutes les opérations en aval.

Remarque : les facteurs d'équilibrage de la réduction peuvent avoir quatre décimales. Les facteurs définis en dessous de 0,0001 donnent des résultats erronés car ils ne sont pas calculés correctement.

- **Créez des conditions** en cliquant sur le bouton situé à droite du champ de texte. Une ligne vide est insérée ; elle vous permet de saisir les nouvelles conditions. Pour créer une expression CLEM pour la condition, cliquez sur le bouton Générateur de formules.
- **Supprimez des directives** à l'aide du bouton de suppression rouge.
- **Triez les directives** à l'aide des flèches vers le haut ou vers le bas.

Equilibrer uniquement les données d'apprentissage. Si un champ de partition figure dans le flux, cette option équilibre les données dans la partition d'apprentissage uniquement. Cela peut s'avérer utile, notamment, si vous générez des scores de propension ajustée qui nécessitent une partition de test ou de validation non équilibrée. Si aucun champ de partition ne figure dans le flux (ou si plusieurs champs de partition sont définis), cette option est ignorée et toutes les données sont équilibrées.

Noeud Agréger

L'agrégation est une tâche de préparation des données fréquemment utilisée pour réduire la taille d'un jeu de données. Avant d'effectuer l'agrégation, vous devez prendre le temps de nettoyer les données, en vous concentrant notamment sur les valeurs manquantes. Une fois l'agrégation terminée, les informations éventuellement utiles concernant les valeurs manquantes risquent d'être perdues.

Les noeuds Agréger permettent de remplacer une séquence d'enregistrements d'entrée par des enregistrements de sortie récapitulatifs et agrégés. Prenons par exemple les enregistrements de ventes d'entrée, tels que ceux affichés dans le tableau ci-après.

Tableau 13. Exemple d'entrée d'un enregistrement de vente

Age	Sexe	Région	Filiale	Ventes
23	M	S	8	4
45	M	S	16	4
37	M	S	8	5
30	M	S	5	7
44	M	S	4	9
25	M	S	2	11
29	F	S	16	6
41	F	S	4	8
23	F	S	6	2
45	F	S	4	5
33	F	S	6	10

Vous pouvez utiliser les champs-clés *Sexe* et *Région* pour agréger ces enregistrements. Agrégez ensuite *Age* avec le mode **Moyenne** et *Ventes* avec le mode **Somme**. Sélectionnez **Inclure le comptage des enregistrements dans le champ** dans la boîte de dialogue du noeud Agréger. Vous obtenez alors ce qui est indiqué dans le tableau ci-après.

Tableau 14. Exemple d'enregistrement agrégé

Âge (moyenne)	Sexe	Région	Ventes (somme)	Nombre d'enregistrements
35.5	F	S	25	4
29	F	S	6	1
34.5	M	S	20	2
33.75	M	S	20	4

Ceci vous apprend par exemple que l'âge moyen des quatre membres féminins de l'équipe de vente dans la région nord est de 35,5 ans et que le montant total de leurs ventes était de 25 unités.

Remarque : les champs comme *Filiale* sont automatiquement exclus lorsqu'aucun mode d'agrégation n'est spécifié.

Définition des options du noeud agrégé

Dans le noeud Agréger, vous spécifiez ce qui suit.

- Un ou plusieurs champs clés à utiliser comme catégories d'agrégation
- Un ou plusieurs champs agrégés pour lesquels calculer les valeurs agrégées
- Un ou plusieurs modes d'agrégation (types d'agrégation) de sortie pour chaque champ agrégé

Vous pouvez également spécifier les modes d'agrégation par défaut à utiliser pour les nouveaux champs ajoutés et utiliser des expressions (similaires aux formules) pour catégoriser l'agrégation.

Notez que, pour obtenir de meilleures performances au niveau des opérations d'agrégation, vous pouvez activer le traitement parallèle.

Champs-clés. Affiche les champs qui peuvent être utilisés comme catégories pour l'agrégation. Les champs continus (numériques) et catégoriels sont autorisés en tant que clés. Si vous sélectionnez plusieurs champs-clés, les valeurs seront combinées de façon à produire une valeur-clé qui sera utilisée pour l'agrégation des enregistrements. Pour chaque champ-clé unique, un enregistrement agrégé est

généré. Par exemple, si vous avez choisi les champs-clés *Sexe* et *Région*, chaque combinaison unique de *M* et *F* avec les régions *N* et *S* (quatre combinaisons uniques) est associée à un enregistrement agrégé. Pour ajouter un champ-clé, utilisez le sélecteur de champs situé à droite dans la fenêtre.

Le reste de la boîte de dialogue est divisée en deux zones principales, **Agrégats de base** et **Expressions agrégées**.

Agrégats de base

Champs d'agrégation. Affiche les champs dont les valeurs seront agrégées, ainsi que les modes d'agrégation sélectionnés. Pour ajouter des champs à cette liste, utilisez le sélecteur de champs situé à droite. Les modes d'agrégation suivants sont disponibles.

Remarque : Certains modes ne s'appliquent pas aux champs non numériques (par exemple, **Somme** pour un champ de date/heure). Les modes qui ne peuvent pas être utilisés avec un champ agrégé sélectionné sont désactivés.

- **Somme.** Cochez cette case pour obtenir les valeurs additionnées de chaque combinaison de champs-clés. La somme ou le total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.
- **Moyenne.** Cochez cette case pour obtenir les valeurs moyennes de chaque combinaison de champs-clés. La moyenne est une mesure de tendance centrale et est la moyenne arithmétique (la somme divisée par le nombre de cas).
- **Min.** Cochez cette case pour obtenir les valeurs minimales de chaque combinaison de champs-clés.
- **Max.** Cochez cette case pour obtenir les valeurs maximales de chaque combinaison de champs-clés.
- **Ecart-type.** Cochez cette case pour obtenir l'écart-type de chaque combinaison de champs-clés. L'écart-type est la mesure de la dispersion des valeurs autour de la moyenne, égale à la racine carrée de la variance.
- **Médiane.** Cochez cette case pour obtenir les valeurs médianes de chaque combinaison de champs-clés. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées. Elle est également nommée 50ème centile ou 2ème quartile.
- **Comptage.** Cochez cette case pour obtenir le nombre de valeurs non nulles pour chaque combinaison de champs-clés.
- **Variance.** Cochez cette case pour obtenir les valeurs de variance de chaque combinaison de champs-clés. La variance est une mesure de dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un.
- **1er quartile.** Cochez cette case pour obtenir les valeurs du 1er quartile (25ème centile) pour chaque combinaison de champs-clés.
- **3ème quartile.** Cochez cette case pour obtenir les valeurs du 3ème quartile (75ème centile) pour chaque combinaison de champs-clés.

Remarque : Lorsque vous exécutez un flux contenant un noeud Agréger, les valeurs renvoyées pour le premier et le troisième quartiles lors de la conversion des instructions SQL dans une base de données Oracle peuvent différer de celles renvoyées en mode natif.

Mode par défaut. Indiquez le mode d'agrégation par défaut à utiliser pour les nouveaux champs ajoutés. Si vous utilisez souvent le même type d'agrégation, sélectionnez un ou plusieurs modes, et utilisez le bouton Appliquer partout situé à droite pour appliquer les modes sélectionnés à tous les champs répertoriés.

Extension du nom du nouveau champ. Permet d'ajouter un suffixe ou un préfixe, tel que "1" ou "nouveau" pour dupliquer les champs agrégés. Par exemple, l'agrégation des valeurs minimales du champ *Age* produit un champ appelé *Age_Min_1* si vous avez sélectionné l'option du suffixe et spécifié 1 comme extension. *Remarque* : les extensions d'agrégation telles que *_Min* ou *Max_* sont

automatiquement ajoutées au nouveau champ, indiquant ainsi le type d'agrégation exécuté. Sélectionnez **Suffixe** ou **Préfixe** pour indiquer le style d'extension voulu.

Inclure le comptage des enregistrements dans le champ. Permet d'inclure un champ supplémentaire dans chaque enregistrement intitulé *Record_Count*, par défaut. Pour chaque enregistrement de sortie, ce champ indique le nombre d'enregistrements d'entrée qui ont été agrégés. Indiquez le nom de votre choix pour ce champ dans le champ d'édition.

Remarque : les valeurs système nulles sont exclues lors du calcul de l'agrégation, mais sont incluses dans le nombre d'enregistrements. En revanche, les valeurs non renseignées sont incluses dans l'agrégation et dans le nombre d'enregistrements. Pour exclure les valeurs non renseignées, vous pouvez utiliser un noeud Remplacer pour remplacer les valeurs non renseignées par des valeurs nulles. Vous pouvez également supprimer les blancs à l'aide d'un noeud Sélectionner.

Expressions agrégées

Les expressions sont similaires à des formules créées à partir de valeurs, de noms de champs, d'opérateurs et de fonctions. Contrairement aux fonctions qui agissent sur un seul enregistrement à la fois, les expressions agrégées agissent sur un groupe, un ensemble ou une collection d'enregistrements.

Remarque : Vous ne pouvez créer des expressions agrégées que si le flux inclut une connexion de base de données (via un noeud de source de base de données).

Les nouvelles expressions sont créées sous forme de champs dérivés. Pour créer une expression, vous utilisez les fonctions *Agrégats de base de données* qui sont disponibles dans le Générateur de formules.

Pour plus d'informations sur le générateur de formules, voir le guide d'utilisation d'IBM SPSS Modeler (ModelerUsersGuide.pdf).

Notez qu'il existe une connexion entre les **Champs-clés** et les expressions agrégées que vous créez car les expressions agrégées sont regroupées par le champs-clé.

Les expressions agrégées valides sont celles dont l'évaluation aboutit à des résultats agrégés. Voici des exemples d'expressions agrégées valides et des règles qui les régissent :

- Vous pouvez utiliser des fonctions scalaires pour combiner plusieurs fonctions d'agrégation en vue de générer un résultat d'agrégation unique. Par exemple :
  ```
  max(C01) - min(C01)
  ```
- Une fonction d'agrégation peut agir sur le résultat de plusieurs fonctions scalaires. Par exemple :
  ```
  sum (C01*C01)
  ```

Agrégation des paramètres d'optimisation

Dans l'onglet Optimisation, spécifiez les valeurs ci-dessous.

Les clés sont adjacentes. Sélectionnez cette option si vous savez que tous les enregistrements ayant les mêmes valeurs de clés sont regroupés dans l'entrée (si, par exemple, l'entrée est triée sur les champs de clé). Ainsi, vous améliorez les performances.

Autoriser une approximation pour la valeur médiane et les quartiles. Les statistiques d'ordre (valeur médiane, premier quartile et troisième quartile) ne sont pas prises en charge actuellement en cas de traitement des données dans Analytic Server. Si vous utilisez Analytic Server, vous pouvez sélectionner cette case à cocher pour utiliser une valeur approximative pour ces statistiques au lieu de la valeur calculée en discrétisant les données puis en calculant une estimation pour la statistique en fonction de la distribution dans les casiers. Par défaut, cette option n'est pas sélectionnée.

Nombre de casiers. Disponible seulement si vous sélectionnez la case à cocher **Autoriser une approximation pour la valeur médiane et les quartiles**. Sélectionnez le nombre de casiers à utiliser lors de l'estimation de la statistique ; ce nombre a un impact sur **% d'erreurs maximal**. Par défaut, le nombre de casiers est de 1000, ce qui correspond à une erreur maximale de 0,1 pour cent de la plage.

Noeud Agréger RFM

Le noeud Agréger RFM (Recency, Frequency, Monetary) permet d'utiliser les données historiques des transactions des clients, de supprimer les données inutiles et de combiner toutes leurs données de transaction restantes dans une seule ligne, en utilisant leur ID de client unique comme clé, qui indique leur dernier contact avec vous (Recency), le nombre de transactions qu'ils ont effectuées (Frequency) et la valeur totale des transactions (Monetary).

Avant d'effectuer une agrégation, vous devez nettoyer les données en vous concentrant notamment sur les valeurs manquantes.

Après avoir identifié et transformé les données en utilisant le noeud Agréger RFM, vous pouvez utiliser un noeud Analyse RFM pour effectuer d'autres analyses. Pour plus d'informations, voir «Noeud Analyse RFM», à la page 181.

Notez qu'une fois que le fichier de données a été exécuté via le noeud Agréger RFM, il ne dispose plus de valeurs cible. Par conséquent, pour pouvoir l'utiliser comme entrée pour effectuer d'autres analyses prédictives avec des noeuds de modélisation, tels que C5.0 ou CHAID, vous devez le fusionner avec d'autres données client (par exemple, en faisant correspondre les ID client). Pour plus d'informations, voir «Noeud Fusionner», à la page 89.

Les noeuds Agréger RFM et Analyse RFM d'IBM SPSS Modeler sont configurés pour utiliser la création d'intervalles indépendants ; en d'autres termes, ils classent et espacent les données sur chaque mesure de valeur de proximité dans le temps, d'effectif et de valeur monétaire, sans tenir compte de leur valeur ni des deux autres mesures.

Définition des options du noeud Agréger RFM

L'onglet Paramètres du Noeud Agréger RFM contient les champs suivants.

Calculer la récence par rapport à Spécifie la date à partir de laquelle la récence des transactions sera calculée. Il peut s'agir d'une **date fixe** que vous entrez ou de la **date du jour**, telle que définie par votre système. La **date du jour** est entrée par défaut et elle est mise à jour automatiquement lors de l'exécution du noeud.

Remarque : L'affichage de la **Date fixe** peut être différente en fonction des environnements locaux. Par exemple, si la valeur `2007-8-10` est stockée dans votre flux sous la forme `Fri Aug 10 00:00:00 CST 2007`, il s'agit d'une date et d'une heure dans le fuseau horaire 'UTC+8'. Toutefois, elle s'affiche sous la forme `Thu Aug 9 12:00:00 EDT 2007` dans le fuseau horaire 'UTC-8'.

Les ID sont contigus Si vos données sont prétriées de façon à ce que tous les enregistrements avec le même ID apparaissent ensemble dans le flux de données, sélectionnez cette option pour accélérer le traitement. Si vos données ne sont pas pré-triées (ou si vous n'en êtes pas certain), ne sélectionnez pas cette option ; le noeud triera automatiquement les données.

ID Sélectionnez le champ à utiliser pour identifier le client et ses transactions. Pour afficher les champs à sélectionner, utilisez le bouton Sélecteur de champs sur la droite.

Date Sélectionnez le champ de date à utiliser pour calculer la récence. Pour afficher les champs à sélectionner, utilisez le bouton Sélecteur de champs sur la droite.

Notez que cela nécessite d'utiliser un champ avec l'enregistrement de date, ou horodatage, dans le format approprié comme entrée. Par exemple, si vous disposez d'un champ de chaîne contenant des valeurs telles que *Jan 2007*, *Fév 2007*, etc., vous pouvez le convertir en champ de date en utilisant un noeud Filtrer et la fonction `to_date()` Pour plus d'informations, voir «Conversion du stockage à l'aide du noeud Remplacer», à la page 170.

Valeur Sélectionnez le champ à utiliser pour calculer la valeur monétaire totale des transactions du client. Pour afficher les champs à sélectionner, utilisez le bouton Sélecteur de champs sur la droite. *Remarque* : Il doit s'agir d'une valeur numérique.

Extension du nom du nouveau champ Ajoutez un suffixe ou un préfixe, tel que "12_month", dans les nouveaux champs générés de récence, d'effectif et monétaire. Sélectionnez **Suffixe** ou **Préfixe** pour indiquer le style d'extension voulu. Par exemple, cela peut s'avérer utile pour examiner différentes périodes.

Supprimer les enregistrements avec une valeur inférieure Si nécessaire, vous pouvez définir une valeur minimale en dessous de laquelle les informations de transaction ne sont pas utilisées pour calculer les totaux RFM. L'unité de valeur fait référence au champ **Valeur** sélectionné.

Inclure uniquement les transactions récentes Si vous analysez une base de données volumineuse, vous pouvez indiquer que seuls les derniers enregistrements doivent être utilisés. Vous pouvez utiliser les données enregistrées après une date donnée ou au cours d'une période récente :

- **Date de transaction après** Spécifie la date de transaction après laquelle les enregistrements seront inclus dans votre analyse.
- **Transaction au cours des derniers** Définissez le nombre et le type des périodes (jours, semaines, mois ou années) par rapport à la date **Calculer la récence par rapport à** après laquelle les enregistrements seront inclus dans l'analyse.

Enregistrer la date de la deuxième transaction la plus récente Si vous voulez connaître la date de la deuxième transaction la plus récente de chaque client, cochez cette case. En outre, vous pouvez cochez la case **Enregistrer la date de la troisième transaction la plus récente**. Elle permet, par exemple, d'identifier les clients qui peuvent avoir exécuté un grand nombre de transactions il y a longtemps, mais une seule transaction récente.

Noeud Trier

Le noeud Trier permet de classer les enregistrements par ordre croissant ou décroissant, en fonction de la valeur d'un ou de plusieurs champs. Les noeuds Trier sont souvent utilisés pour visualiser et sélectionner les enregistrements contenant les valeurs de données les plus répandues. La première opération consiste à agréger les données à l'aide d'un noeud Agréger, puis à utiliser un noeud Trier pour les trier par ordre décroissant de nombre d'enregistrements. Les résultats apparaissent dans un tableau, dans lequel vous pouvez examiner les données et les manipuler (pour sélectionner les enregistrements relatifs aux dix clients les plus fidèles, par exemple).

L'onglet Paramètres du noeud Trier contient les champs suivants.

Trier par. Tous les champs sélectionnés en tant que clé de tri apparaissent dans un tableau. Le tri est plus efficace lorsque le champ-clé est numérique.

- **Ajoutez des champs** à cette liste en utilisant le sélecteur de champs situé à droite.
- **Sélectionnez un ordre** en cliquant sur la flèche **Croissant** ou **Décroissant** de la colonne *Ordre* du tableau.
- **Supprimez des champs** à l'aide du bouton de suppression rouge.
- **Triez les directives** à l'aide des flèches vers le haut ou vers le bas.

Ordre de tri par défaut. Sélectionnez **Croissant** ou **Décroissant** pour définir l'ordre de tri par défaut lorsque de nouveaux champs sont ajoutés.

Remarque : Le noeud Trier n'est pas appliqué s'il existe un noeud Distinguer en aval du flux de modèle. Pour des informations sur le noeud Distinguer, voir «Noeud Distinguer», à la page 98.

Paramètres d'optimisation du tri

Si vous savez que les données avec lesquelles vous travaillez ont déjà été triées en fonction de certains champs-clés, vous pouvez indiquer les champs ayant déjà fait l'objet d'un tri ; le système peut ainsi trier le reste des données de manière plus efficace. Il se peut, par exemple, que vous souhaitiez effectuer un tri sur les champs *Age* (décroissant) et *Médicament* (croissant) mais que vous sachiez que les données ont déjà été triées en fonction du champ *Age* (décroissant).

Les données ont fait l'objet d'un tri préalable. Indique si les données sont déjà triées en fonction d'un ou de plusieurs champs.

Spécifier l'ordre de tri existant. Indiquez les champs déjà triés. Dans la boîte de dialogue Sélectionner les champs, ajoutez des champs à la liste. Dans la colonne *Ordre*, précisez si chaque champ est trié dans l'ordre croissant ou décroissant. Si vous entrez ici plusieurs champs, veillez à les répertorier dans le bon ordre. Cliquez sur les flèches situées à droite de la liste pour trier les champs dans l'ordre souhaité. Si vous définissez un ordre de tri existant incorrect, un message d'erreur apparaît lors de l'exécution du flux, indiquant le numéro d'enregistrement au niveau duquel le tri n'est pas conforme à ce que vous avez indiqué.

Remarque : L'activation du traitement parallèle peut profiter à la vitesse de tri.

Noeud Fusionner

Le noeud Fusionner permet de créer à partir de plusieurs enregistrements d'entrée un seul enregistrement de sortie contenant la totalité ou une partie des champs d'entrée. Cette opération permet notamment de fusionner des données provenant de différentes sources, telles que les données client internes et les données démographiques acquises. Vous pouvez fusionner des données des manières suivantes :

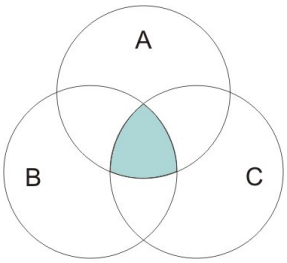
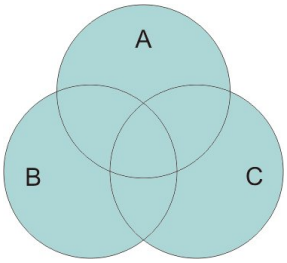
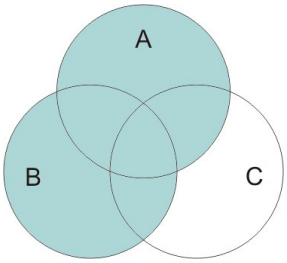
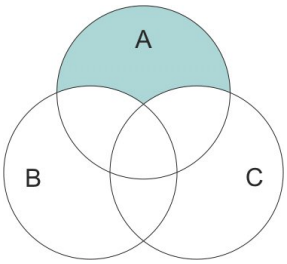
- La fusion par **Ordre** concatène les enregistrements correspondants issus de toutes les sources dans l'ordre d'entrée, jusqu'à ce que la plus petite source de données soit épuisée. Si vous utilisez cette option, il est important d'avoir trié vos données à l'aide d'un noeud Trier.
- La fusion à l'aide d'un champ **Clé**, tel que *ID client*, vous permet de spécifier le mode de mise en correspondance des enregistrements d'une source de données avec ceux d'autres sources de données. Plusieurs types de jointures sont disponibles, notamment les jointures interne, externe complète, externe partielle et anti-jointure. Pour plus d'informations, voir «Types de jointure».
- La fusion par **Condition** signifie que vous pouvez spécifier une condition à remplir pour que la fusion ait lieu. La condition peut être spécifiée directement dans le noeud, ou il est possible de la générer à l'aide du Générateur de formules.
- La fusion par **Condition classée** est une jointure externe gauche dans laquelle vous spécifiez une condition à remplir pour que la fusion ait lieu et une expression de classement qui procède à un tri par ordre croissant. Elle est le plus souvent utilisée pour fusionner des données géospatiales et vous pouvez spécifier la condition directement dans le noeud, ou construire la condition à l'aide du générateur de formules.

Types de jointure

Lorsque vous utilisez un champ-clé pour la fusion des données, prenez le temps nécessaire pour choisir les enregistrements à inclure et à exclure. Il existe diverses jointures qui sont présentées en détail ci-dessous.

Les deux principaux types de jointure sont appelés jointures internes et jointures externes. Ces méthodes sont souvent utilisées pour fusionner des tables à partir de jeux de données associés, sur la base des valeurs communes d'un champ-clé, tel que *ID client*. Les jointures internes permettent d'obtenir des fusions "propres", ainsi qu'un jeu de données de sortie n'incluant que les enregistrements complets. Les jointures externes comprennent également des enregistrements complets issus des données fusionnées, mais elles vous permettent également d'inclure des données uniques provenant d'une ou de plusieurs tables d'entrée.

Les types de jointures autorisés sont décrits en détail ci-dessous.

	Une **jointure interne** inclut uniquement les enregistrements dans lesquels une valeur pour le champ-clé est commune à toutes les tables d'entrée. Cela signifie que les enregistrements sans correspondance ne sont pas inclus dans le jeu de données de sortie.
	Une **jointure externe complète** comprend tous les enregistrements, ceux qui ont une correspondance comme ceux qui n'en ont pas, des tables d'entrée. Les jointures externes gauche et droite sont appelées jointures externes partielles et sont décrites ci-dessous.
	Une **jointure externe partielle** inclut tous les enregistrements mis en correspondance à l'aide du champ-clé, ainsi que les enregistrements sans correspondance issus des tables spécifiées. (Autrement dit, elle inclut tous les enregistrements de certaines tables et uniquement les enregistrements correspondants d'autres tables.) Les tables (par exemple, A et B) peuvent être sélectionnées pour être incluses dans la jointure externe à l'aide du bouton Sélectionner de l'onglet Fusionner. Les jointures partielles sont également appelées jointures externes gauche ou droite lorsque deux tables seulement sont fusionnées. IBM SPSS Modeler autorisant la fusion de plusieurs tables, cette jointure est appelée une jointure externe partielle.
	Une **anti-jointure** ne comprend que les enregistrements sans correspondance de la première table d'entrée (table A dans cet exemple). Ce type de jointure est le contraire de la jointure interne ; elle n'inclut pas les enregistrements complets dans le jeu de données de sortie.

Par exemple, si un jeu de données contient des informations sur des fermes et qu'un autre comporte des déclarations de sinistre relatives aux fermes, vous pouvez mettre en correspondance les enregistrements de la première source et ceux de la seconde à l'aide des options de fusion.

Pour déterminer si un client inclus dans cet exemple de fermes a émis une déclaration de sinistre, utilisez l'option de jointure interne pour renvoyer la liste des correspondances de tous les ID de ces deux jeux de

données.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalue
1	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0...
2	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0...
3	id620	name620	north	880.000	74.000	6.000	426988.000	rapeseed	arable_de	118885.0...

Figure 2. Exemple de sortie pour une fusion réalisée par jointure interne

L'option de jointure externe complète permet de renvoyer à partir des tables d'entrée les enregistrements avec et sans correspondance. La valeur manquante (`$null$`) du système est utilisée pour les valeurs incomplètes.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalu
1	id601	$null$	$null$	$null$	$null$	$null$	$null$	$null$	decomm...	74703.1(
2	id602	name602	north	1780.000	42.000	9.000	734118.000	maize	$null$	$nul
3	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0
4	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0
5	id606	$null$	$null$	$null$	$null$	$null$	$null$	$null$	arable_de	122135.0

Figure 3. Exemple de sortie pour une fusion réalisée par jointure externe complète

Une jointure externe partielle inclut tous les enregistrements mis en correspondance à l'aide du champ-clé, ainsi que les enregistrements sans correspondance issus des tables spécifiées. Le tableau affiche tous les enregistrements mis en correspondance à partir du champ d'ID, ainsi que ceux mis en correspondance à partir du premier jeu de données.

	id	claimtype	claimvalue	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	$null$	$null$	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id604	decomm...	281082.0...	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes
3	id605	decomm...	122006.0...	name605	north	1700.000	46.000	8.000	621148.000	wheat
4	id607	$null$	$null$	name607	southeast	1820.000	29.000	6.000	211605.000	maize
5	id608	$null$	$null$	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
6	id609	$null$	$null$	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
7	id615	$null$	$null$	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
8	id618	$null$	$null$	name618	southeast	1180.000	98.000	3.000	368646.000	maize

Figure 4. Exemple de sortie pour une fusion réalisée par jointure externe partielle

Si vous utilisez l'option d'anti-jointure, la table ne renvoie que les enregistrements sans correspondance issus de la première table d'entrée.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id607	name607	southeast	1820.000	29.000	6.000	211605.000	maize
3	id608	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
4	id609	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
5	id615	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
6	id618	name618	southeast	1180.000	98.000	3.000	368646.000	maize
7	id619	name619	north	840.000	64.000	8.000	457552.000	potatoes

Figure 5. Exemple de sortie pour une fusion réalisée par anti-jointure

Spécification d'une méthode de fusion et des clés

L'onglet Fusion du noeud Fusionner contient les champs suivants.

Méthode de fusion Sélectionnez la méthode à utiliser pour fusionner les enregistrements. La sélection de l'option **Clés** ou **Condition** active la moitié inférieure de la boîte de dialogue.

- **Ordre** Fusionne les enregistrements par ordre. Par exemple le *n*ième enregistrement de chaque entrée peut être fusionné pour générer le *n*ième enregistrement de sortie. Lorsqu'un enregistrement n'a plus

d'enregistrement d'entrée correspondant, la création d'enregistrements de sortie s'arrête. Cela signifie que le nombre d'enregistrements qui sont créés correspond au nombre d'enregistrements dans le jeu de données le plus petit.

- **Clés** Utilise un champ-clé, par exemple *ID de transaction*, pour fusionner les enregistrements dont la valeur dans le champ-clé est identique. Cette opération est équivalente à une "équi-jointure" de base de données. S'il existe plusieurs occurrences d'une valeur-clé, toutes les combinaisons possibles sont renvoyées. Par exemple, si des enregistrements avec la même valeur de champ-clé *A* contiennent des valeurs *B*, *C* et *D* dans d'autres champs, les champs fusionnés produisent un enregistrement distinct pour chaque combinaison de *A* avec la valeur *B*, de *A* avec la valeur *C*, et de *A* avec la valeur *D*.

  Remarque : les valeurs nulles ne sont pas considérées comme identiques dans la méthode de fusion par clés et ne sont pas regroupées.
- **Condition** Utilisez cette option afin de spécifier une condition pour la fusion. Pour plus d'informations, voir «Spécification de conditions pour une fusion», à la page 93.
- **Condition classée** Utilisez cette option pour spécifier si chaque ligne appariée dans le jeu de données principal et tous les jeux de données secondaires doit être fusionnée ; utilisez l'expression de classement pour trier plusieurs correspondances par ordre croissant. Pour plus d'informations, voir «Spécification de conditions classées pour une fusion», à la page 93.

Clés possibles Répertorie uniquement les champs dont les noms correspondent exactement dans toutes les sources de données d'entrée. Sélectionnez un champ dans la liste et utilisez la flèche pour l'ajouter en tant que champ-clé pour la fusion des enregistrements. Vous pouvez utiliser plusieurs champs-clés. Vous pouvez renommer les champs d'entrée qui ne correspondent pas à l'aide d'un noeud Filtrer ou de l'onglet Filtrer d'un noeud source.

Clés pour fusion Répertorie tous les champs qui sont utilisés pour fusionner des enregistrements depuis toutes les sources de données d'entrée en fonction des valeurs des champs-clés. Pour supprimer une clé de la liste, sélectionnez-la et utilisez la flèche pour la renvoyer dans la liste Clés possibles. Lorsque plusieurs champs-clés sont sélectionnés, l'option ci-dessous est activée.

Combiner des champs-clés dupliqués Lorsque plusieurs champs-clés sont sélectionnés ci-dessus, cette option garantit qu'il n'existe qu'un seul champ de sortie de ce nom. Elle est activée par défaut, excepté lorsque des flux ont été importés de versions antérieures d'IBM SPSS Modeler. Lorsque cette option est désactivée, les champs-clés en double doivent être renommés ou exclus à l'aide de l'onglet Filtrer de la boîte de dialogue du noeud Fusionner.

Inclure uniquement les enregistrements correspondants (jointure interne) Sélectionnez cette option pour ne fusionner que les enregistrements complets.

Inclure les enregistrements avec et sans correspondance (jointure externe complète) Sélectionnez cette option pour effectuer une "jointure externe complète". Cela signifie que même si les valeurs du champ-clé ne sont pas présentes dans toutes les tables d'entrée, les enregistrements incomplets sont néanmoins conservés. La valeur indéfinie (`$null$`) est ajoutée au champ-clé et incluse dans l'enregistrement de sortie.

Inclure les enregistrements avec correspondance et les enregistrements sans correspondance sélectionnés (jointure externe partielle) Sélectionnez cette option pour effectuer une "jointure externe partielle" des tables que vous sélectionnez dans cette sous-boîte de dialogue. Cliquez sur **Sélectionner** pour indiquer les tables pour lesquelles des enregistrements incomplets seront conservés lors de la fusion.

Inclure les enregistrements dans le premier jeu de données sans correspondance (anti-jointure) Sélectionnez cette option pour effectuer un type d'"anti-jointure", où seuls les enregistrements sans correspondance provenant du premier jeu de données sont transmises en aval. Vous pouvez indiquer l'ordre des jeux de données d'entrée à l'aide des flèches de l'onglet Entrées. Ce type de jointure n'inclut pas les enregistrements complets dans le jeu de données de sortie. Pour plus d'informations, voir «Types de jointure», à la page 89.

Sélection de données pour des jointures partielles

Pour une jointure externe partielle, vous devez sélectionner les tables pour lesquelles des enregistrements incomplets seront conservés. Par exemple, vous pouvez conserver tous les enregistrements d'une table Client et ne conserver que les enregistrements avec correspondance de la table Prêt hypothécaire.

Colonne Jointure externe. Dans la colonne *Jointure externe*, sélectionnez les jeux de données à inclure dans leur intégralité. Pour une jointure partielle, les enregistrements qui se chevauchent seront conservés, de même que les enregistrements incomplets pour les jeux de données sélectionnés à ce niveau. Pour plus d'informations, voir «Types de jointure», à la page 89.

Spécification de conditions pour une fusion

En définissant la méthode de fusion sur **Condition**, vous pouvez spécifier une ou plusieurs conditions à satisfaire qui déterminent si la fusion a lieu ou pas.

Vous pouvez entrer les conditions directement dans le champ Condition ou les créer à l'aide du Générateur de formules en cliquant sur le bouton en forme de calculatrice situé à droite du champ.

Ajouter des balises pour dupliquer les noms de champ afin d'éviter les conflits de fusion Si plusieurs jeux de données à fusionner contiennent les mêmes noms de champ, sélectionnez cette case à cocher afin d'ajouter une balise de préfixe différente au début des en-têtes de colonne de champ. Par exemple, si deux champs s'appellent *Nom*, le résultat de la fusion contiendra *1_Nom* et *2_Nom*. Si la balise a été renommée dans la source de données, le nouveau nom est utilisé à la place de la balise de préfixe représentant un nombre. Si vous ne sélectionnez pas cette case à cocher et qu'il existe des noms en double dans les données, un avertissement s'affiche à droite de la case à cocher.

Spécification de conditions classées pour une fusion

Une fusion de condition classée peut être considérée comme une fusion de jointure externe à gauche par condition ; la partie de gauche de la fusion correspond au jeu de données principal dans lequel chaque enregistrement est un événement. Par exemple, dans un modèle qui est utilisé pour rechercher des motifs dans des données de criminalité, chaque enregistrement dans le jeu de données principal représente un crime et les informations associées (emplacement, type, etc.). Dans cet exemple, la partie de droite peut contenir les jeux de données géospatiales pertinentes.

La fusion utilise une condition de fusion et une expression de classement. La condition de fusion peut utiliser une fonction géospatiale telle que *within* ou *close_to*. Au cours de la fusion, tous les champs de la partie de droite des jeux de données sont ajoutés au jeu de données de gauche, mais plusieurs correspondances génèrent une zone de liste. Par exemple :

- A gauche : données de criminalité
- A droite : jeu de données des départements et jeu de données des routes
- Conditions de fusion : données de criminalité dans (*within*) les départements et près (*close_to*) des routes, avec une définition pour *close_to*.

Dans cet exemple, si un crime est survenu dans le rayon *close_to* requis de trois routes (et que le nombre de correspondances à renvoyer est d'au moins trois), les trois routes sont renvoyées sous forme d'éléments de liste.

En définissant la méthode de fusion **Condition classée**, vous pouvez spécifier une ou plusieurs conditions à remplir pour que la fusion ait lieu.

Jeu de données principal Sélectionnez le jeu de données principal pour la fusion ; les champs de tous les autres jeux de données sont ajoutés au jeu de données que vous sélectionnez. Ce jeu de données peut être considéré comme la partie de gauche d'une fusion de jointure externe.

Lorsque vous sélectionnez un jeu de données principal, tous les autres jeux de données d'entrée qui sont connectés au noeud Fusionner sont répertoriés automatiquement dans la table **Fusions**.

Ajouter des balises pour dupliquer les noms de champ afin d'éviter les conflits de fusion Si plusieurs jeux de données à fusionner contiennent les mêmes noms de champ, sélectionnez cette case à cocher pour ajouter une balise de préfixe différente au début des en-têtes de colonne de champ. Par exemple, s'il existe deux champs appelés *Nom*, le résultat de la fusion contient *1_Nom* et *2_Nom*. Si la balise est renommée dans la source de données, le nouveau nom est utilisé à la place de la balise de préfixe représentant un nombre. Si vous ne sélectionnez pas cette case à cocher et qu'il existe des noms en double dans les données, un avertissement s'affiche à droite de la case à cocher.

Fusions

Jeu de données

Affiche le nom des jeux de données secondaires qui sont connectés en tant qu'entrées au noeud Fusionner. Par défaut, lorsqu'il existe plusieurs jeux de données secondaires, ils sont répertoriés dans l'ordre dans lequel ils ont été connectés au noeud Fusionner.

Condition de fusion

Entrez les conditions uniques pour la fusion de chaque jeu de données dans la table avec le jeu de données principal. Vous pouvez entrer les conditions directement dans la cellule ou les construire à l'aide du générateur de formules en cliquant sur le symbole de la calculatrice à droite de la cellule. Par exemple, vous pouvez utiliser des prédicats géospatiaux pour créer une condition de fusion qui place les données de criminalité d'un jeu de données dans les données de département d'un autre jeu de données. La condition de fusion par défaut dépend du niveau de mesure géospatial, conformément à la liste ci-dessous.

- Point, Chaîne, Multipoint, Multichaîne - condition par défaut de *close_to*.
- Polygone, Multipolygone - condition par défaut de *within*.

Pour plus d'informations sur ces niveaux, voir «Sous-niveaux de mesure géospatiaux», à la page 147.

Si un jeu de données contient plusieurs champs géospatiaux de types différents, la condition par défaut qui est utilisée dépend du premier niveau de mesure trouvé dans les données, dans l'ordre décroissant ci-dessous.

- Point
- Chaîne
- Polygone

Remarque : Les valeurs par défaut ne sont disponibles que s'il existe un champ de données géospatiales dans la base de données secondaire.

Expression de classement

Spécifiez une expression selon laquelle classer la fusion des jeux de données ; cette expression est utilisée pour trier plusieurs correspondances dans un ordre reposant sur les critères de classement. Vous pouvez entrer les conditions directement dans la cellule ou les construire à l'aide du générateur de formules en cliquant sur le symbole de la calculatrice à droite de la cellule.

Les expressions de classement par défaut des distances et des zones sont fournies dans le générateur de formules ; elles vont de faible à élevé, ce qui signifie par exemple que la meilleure correspondance pour la distance est la valeur la plus faible. Par exemple, pour un classement en fonction de la distance, le jeu de données principal contient des crimes et leur emplacement, et chaque autre jeu de données contient des objets et leur emplacement ; dans ce cas, la distance entre les crimes et les objets peut être utilisée comme critère de classement. L'expression de classement par défaut dépend du niveau de mesure géospatial, conformément à la liste ci-dessous.

- Point, Chaîne, Multipoint, Multichaîne - l'expression par défaut est *distance*.
- Polygone, Multipolygone - l'expression par défaut est *area*.

Remarque : Les valeurs par défaut ne sont disponibles que s'il existe un champ de données géospatiales dans la base de données secondaire.

Nombre de correspondances
Spécifiez le nombre de correspondances qui sont renvoyées, en fonction des expressions de condition et de classement. Le nombre de correspondances par défaut dépend du niveau de mesure géospatial dans le jeu de données secondaire, conformément à la liste ci-dessous. Toutefois, vous pouvez cliquer deux fois dans la cellule pour entrer votre propre valeur, jusqu'à 100.

- Point, Chaîne, Multipoint, Multichaîne - la valeur par défaut est 3.
- Polygone, Multipolygone - la valeur par défaut est 1.
- Le jeu de données ne contient pas de champ géospatial - la valeur par défaut est 1.

Par exemple, si vous configurez une fusion qui repose sur une **condition de fusion** *close_to* et une **expression de classement** *distance*, les trois meilleures correspondances (les plus proches) des jeux de données secondaires pour chaque enregistrement dans le jeu de données principal sont renvoyées comme valeurs dans la zone de liste résultante.

Filtrage des champs à partir du noeud Fusionner

Les noeuds Fusionner permettent de filtrer ou de renommer des champs apparaissant en double à la suite de la fusion de plusieurs sources de données. Cliquez sur l'onglet **Filtrer** de la boîte de dialogue pour sélectionner les options de filtrage.

Les options présentées sont quasi-identiques à celles du noeud Filtrer. Toutefois, des options supplémentaires, non présentées ici, sont disponibles dans le menu Filtrer. Pour plus d'informations, voir «Filtrage ou modification du nom des champs», à la page 159.

Champ. Affiche les champs d'entrée des sources de données actuellement connectées.

Marque. Affiche le nom (ou le numéro) de la marque associée au lien de la source de données. Cliquez sur l'onglet **Entrées** pour modifier les liens actifs vers ce noeud Fusionner.

Noeud source. Affiche le noeud source dont les données sont en cours de fusion.

Noeud connecté. Affiche le nom du noeud connecté au noeud Fusionner. Les tâches d'exploration de données complexes nécessitent souvent plusieurs opérations de fusion ou d'ajout qui peuvent inclure le même noeud source. Le nom du noeud connecté permet de les distinguer.

Filtrer. Affiche les connexions actuelles entre le champ d'entrée et le champ de sortie. Les connexions actives affichent une flèche continue. Les connexions affichant un X rouge indiquent des champs filtrés.

Champ. Affiche les champs de sortie après la fusion ou l'ajout. Les champs en double sont affichés en rouge. Cliquez dans le champ Filtrer pour désactiver les champs en double.

Afficher les champs actuels. Sélectionnez cette option pour afficher des informations sur les champs à utiliser en tant que champs-clés.

Afficher les paramètres de champ non utilisés. Sélectionnez cette option pour afficher des informations sur les champs actuellement non utilisés.

Définition de l'ordre d'entrée et du marquage

A l'aide de l'onglet Entrées des boîtes de dialogue des noeuds Fusionner et Ajouter, vous pouvez spécifier l'ordre des sources de données d'entrée et modifier le nom de la marque de chaque source.

Marques et ordre des jeux de données d'entrée. Sélectionnez cette option pour fusionner ou ajouter uniquement les enregistrements complets.

- **Marque.** Affiche les noms de marque actuels pour chaque source de données d'entrée. Les noms de marque, également appelés **marques**, permettent d'identifier de façon unique les liens de données pour les opérations de fusion et d'ajout. Imaginons, par exemple, de l'eau provenant de divers conduits, qui débouche en un point, puis circule à travers un conduit unique. Les données dans IBM SPSS Modeler circulent de façon similaire et le point de fusion est souvent une interaction complexe entre les différentes sources de données. Les marques permettent de gérer les entrées ("conduits") vers un noeud Fusionner ou Ajouter de sorte que si le noeud est enregistré ou déconnecté, les liens ne sont pas supprimés et sont facilement identifiables.

Lorsque vous connectez des sources de données supplémentaires à un noeud Fusionner ou Ajouter, des marques par défaut sont automatiquement créées, à l'aide de nombres, afin de représenter l'ordre de connexion des noeuds. Cet ordre n'est pas lié à l'ordre des champs dans les jeux de données d'entrée ou de sortie. Vous pouvez modifier la marque par défaut en entrant un nouveau nom dans la colonne *Marque*.

- **Noeud source.** Affiche le noeud source dont les données sont en cours de fusion.
- **Noeud connecté.** Affiche le nom du noeud connecté au noeud Fusionner ou Ajouter. Les tâches d'exploration de données complexes nécessitent souvent plusieurs opérations de fusion qui peuvent inclure le même noeud source. Le nom du noeud connecté permet de les distinguer.
- **Champ.** Répertorie le nombre de champs dans chaque source de données.

Afficher les marques actuelles. Sélectionnez cette option pour afficher les marques actuellement utilisées par le noeud Fusionner ou Ajouter. En d'autres termes, les marques actuelles identifient les liens vers le noeud à travers lequel circulent des données. En reprenant la métaphore des conduits, les marques actuelles s'apparentent aux conduits dans lesquels circule de l'eau.

Afficher les paramètres des marques non utilisées. Sélectionnez cette option pour afficher les balises, ou liens, précédemment utilisés pour la connexion au noeud Fusionner ou Ajouter, mais qui ne sont pas actuellement connectées à une source de données. Cette représentation s'apparente aux conduits vides au sein d'un réseau de plomberie. Vous pouvez choisir de connecter ces "conduits" à une nouvelle source ou de les supprimer. Pour supprimer du noeud les marques non utilisées, cliquez sur **Effacer**. Cette action efface en une fois toutes les marques non utilisées.

Paramètres d'optimisation de la fusion

Le système comprend deux options qui permettent une fusion plus efficace des données dans certaines situations. Ainsi, grâce à ces options, vous pouvez optimiser l'opération de fusion lorsqu'un jeu de données d'entrée est nettement plus volumineux que les autres, ou lorsque les données sont déjà triées en fonction de tout ou partie des champs-clés utilisés pour la fusion.

Remarque : Les optimisations dans cet onglet s'appliquent uniquement à l'exécution du noeud natif IBM SPSS Modeler, c'est-à-dire lorsque le noeud Fusionner n'est pas répercuté dans SQL. Les paramètres d'optimisation n'ont pas d'effet sur la génération SQL.

Un jeu de données d'entrée est relativement volumineux. Sélectionnez cette option pour indiquer que l'un des jeux de données d'entrée est bien plus volumineux que les autres. Le système met alors en mémoire cache les jeux de données moins volumineux, puis traite, pour la fusion, le jeu de données le plus volumineux sans le trier ni le mettre en mémoire cache. Vous utilisez généralement ce type de jointure avec des données organisées selon un schéma en étoile (ou selon tout autre schéma semblable), en présence d'une table centrale volumineuse contenant des données partagées (des données

transactionnelles, par exemple). Si vous sélectionnez cette option, cliquez sur **Sélectionner** pour indiquer le jeu de données volumineux. Vous ne pouvez sélectionner qu'*un* jeu de données volumineux. Le tableau suivant reprend chaque type de jointure et indique pour chacun s'il peut être optimisé à l'aide de cette méthode.

Tableau 15. Récapitulatif des optimisations de jointure.

Type de jointure	Peut être optimisé pour un jeu de données d'entrée volumineux ?
Interne	Oui
Partiel	Oui, si le jeu de données volumineux ne contient pas d'enregistrements incomplets.
Complet	Non
Anti-jointure	Oui, si le jeu de données volumineux constitue la première entrée.

Toutes les entrées sont déjà triées par champ(s)-clé(s). Sélectionnez cette option pour indiquer que les données d'entrée sont déjà triées en fonction d'un ou de plusieurs des champs-clés utilisés pour la fusion. Assurez-vous que *tous* les jeux de données d'entrée sont triés.

Spécifier l'ordre de tri existant. Indiquez les champs déjà triés. Dans la boîte de dialogue Sélectionner les champs, ajoutez des champs à la liste. Vous ne pouvez sélectionner que les champs-clés utilisés pour la fusion (tels que définis dans l'onglet Fusionner). Dans la colonne *Ordre*, précisez si chaque champ est trié dans l'ordre croissant ou décroissant. Si vous entrez ici plusieurs champs, veillez à les répertorier dans le bon ordre. Cliquez sur les flèches situées à droite de la liste pour trier les champs dans l'ordre souhaité. Si vous définissez un ordre de tri existant incorrect, un message d'erreur apparaît lors de l'exécution du flux, indiquant le numéro d'enregistrement au niveau duquel le tri n'est pas conforme à ce que vous avez indiqué.

En fonction de la sensibilité à la casse de la méthode de collationnement utilisée par la base de donnée, il se peut que l'optimisation ne fonctionne pas correctement lorsqu'une ou plusieurs entrées sont triées par la base de données. Par exemple, si sur deux entrées, l'une est sensible à la casse et l'autre non, les résultats du tri peuvent être différents. L'optimisation de la fusion entraîne le traitement des enregistrements selon leur ordre de tri. En conséquence, si les entrées sont triées à l'aide de méthodes de collationnement différentes, le noeud Fusion fait état d'une erreur et affiche le numéro de l'enregistrement dont le tri a été incohérent. Lorsque toutes les entrées proviennent d'une source unique, ou sont triées à l'aide de collationnements mutuellement inclusifs, les enregistrements peuvent être fusionnés avec succès.

Remarque : L'activation du traitement parallèle peut profiter à la vitesse de fusion.

Noeud Ajouter

Les noeuds Ajouter permettent de concaténer des ensembles d'enregistrements. Contrairement au noeud Fusionner, qui joint des enregistrements provenant de sources différentes, le noeud Ajouter lit tous les enregistrements d'une source jusqu'au dernier et les inclut dans le flux en aval. Les enregistrements de la source suivante sont ensuite lus, avec la même structure de données (nombre d'enregistrements, nombre de champs, etc.) que la première source, ou source principale. Si la source principale comporte plus de champs qu'une autre source d'entrée, la chaîne manquante par défaut (`$null$`) est utilisée pour les valeurs incomplètes.

Utilisez un noeud Ajouter pour combiner des jeux de données dont les structures sont similaires, mais les données différentes. Par exemple, vous pouvez avoir stocké les données relatives à vos transactions dans des fichiers différents selon la période à laquelle elles se rapportent (un fichier pour mars et un autre pour avril, par exemple). Supposons que ces fichiers sont structurés de la même manière (les mêmes champs dans le même ordre), le noeud Ajouter réunira les données dans un même fichier que vous pourrez alors analyser.

Remarque : pour que les fichiers puissent être ajoutés, les niveaux de mesure de champ doivent être identiques. Par exemple, un champ *Nominal* ne peut être ajouté à un champ dont le niveau de mesure est *Continu*.

Définition des options du noeud Ajouter

Apparier les champs par. Sélectionnez la méthode d'occurrence des champs à ajouter.

- **Position.** Permet d'ajouter des jeux de données, sur la base de la position des champs dans la source de données principale. Lorsque vous utilisez cette méthode, pensez à trier vos données afin de garantir un ajout adéquat.
- **Nom.** Permet d'ajouter des jeux de données, sur la base du nom des champs dans les jeux de données d'entrée. Sélectionnez également **Respecter la casse** pour activer la distinction des majuscules/minuscules lors de la mise en correspondance des noms de champ.

Champ de sortie. Affiche les noeuds source connectés au noeud Ajouter. Le premier noeud de la liste est la source d'entrée principale. Vous pouvez trier les champs en cliquant sur l'en-tête de la colonne. Ce tri n'a pas pour effet de réorganiser les champs dans le jeu de données.

Inclure les champs de. Sélectionnez **Premier jeu de données uniquement** pour générer des champs de sortie sur la base des champs du jeu de données principal. L'jeu de données principal est la première entrée, spécifiée dans l'onglet Entrées. Sélectionnez **Tous les jeux de données** pour générer des champs de sortie pour tous les champs dans tous les jeux de données, qu'un champ correspondant soit présent dans tous les jeux de données d'entrée ou non.

Baliser les enregistrements en incluant le jeu de données dans le champ. Sélectionnez cette option pour ajouter un champ supplémentaire au fichier de sortie, dont les valeurs indiquent le jeu de données source pour chaque enregistrement. Indiquez un nom dans le champ de texte. Le nom par défaut du champ est *Entrée*.

Noeud Distinguer

Les enregistrements en double d'un jeu de données doivent être supprimés avant le début de l'exploration de données. Par exemple, dans une base de données marketing, certaines personnes peuvent apparaître plusieurs fois avec des adresses différentes ou des informations de contact différentes. Vous pouvez utiliser le noeud Distinguer pour rechercher ou supprimer des enregistrements dans vos données ou pour créer un enregistrement composite unique depuis un groupe d'enregistrements dupliqués.

Pour utiliser le noeud Distinguer, vous devez d'abord définir un ensemble de champs-clés qui déterminent si deux enregistrements sont dupliqués.

Si vous ne sélectionnez pas tous vos champs comme champs-clés, il se peut que deux enregistrements considérés comme "dupliqués" ne soit pas exactement identiques car les valeurs figurant dans les autres champs peuvent différer. Dans ce cas, vous pouvez aussi définir un ordre de tri qui est appliqué dans chaque groupe d'enregistrements dupliqués. Cet ordre de tri permet un meilleur contrôle des enregistrements qui sont traités en premier dans un groupe. Sinon, tous les doublons sont considérés comme interchangeables et n'importe quel enregistrement peut être sélectionné. L'ordre entrant des enregistrements n'est pas pris en compte ; par conséquent, l'utilisation d'un noeud Trier en amont n'a pas d'intérêt (voir "Tri des enregistrements dans le noeud Distinguer" ci-après).

Mode. Indiquez si vous souhaitez créer un enregistrement composite, ou bien inclure ou exclure (supprimer) le premier enregistrement.

- **Créer un enregistrement composite pour chaque groupe.** Permet d'agréger des champs non numériques. Si vous sélectionnez cette option, l'onglet Composite dans lequel vous spécifiez le mode de création des enregistrements composites devient disponible. Pour plus d'informations, voir «Distinct - Composite - Onglet Paramètres», à la page 101.

- **Inclure uniquement le premier enregistrement dans chaque groupe.** Sélectionne le premier enregistrement de chaque groupe d'enregistrements dupliqués et supprime le reste. Le *premier* enregistrement est déterminé par l'ordre de tri défini ci-dessous et non par l'ordre entrant des enregistrements.
- **Ignorer uniquement le premier enregistrement dans chaque groupe.** Supprime le premier enregistrement de chaque groupe d'enregistrements dupliqués et sélectionnez le reste à la place. Le *premier* enregistrement est déterminé par l'ordre de tri défini ci-dessous et non par l'ordre entrant des enregistrements. Cette option permet de *détecter* les doublons présents dans les données, afin qu'ils puissent être examinés ultérieurement dans le flux.

Champs-clés pour le regroupement. Répertorie le ou les champs utilisés pour détecter les enregistrements identiques. Vous pouvez :

- Ajouter des champs à cette liste en utilisant le bouton de sélection des champs situé à droite.
- Supprimer des champs de la liste à l'aide du bouton de suppression rouge en forme de X.

A l'intérieur des groupes, trier les enregistrements par. Répertorie les champs utilisés pour déterminer la façon dont les enregistrements sont triés dans chaque groupe de doublons et s'ils sont triés par ordre croissant ou décroissant. Vous pouvez :

- Ajouter des champs à cette liste en utilisant le bouton de sélection des champs situé à droite.
- Supprimer des champs de la liste à l'aide du bouton de suppression rouge en forme de X.
- Déplacer les champs à l'aide des boutons Haut ou Bas, si vous triez en fonction de plusieurs champs.

Vous devez spécifier un ordre de tri si vous avez choisi d'inclure ou d'exclure le premier enregistrement dans chaque groupe et s'il est important pour vous de déterminer quel enregistrement est traité en premier.

Vous pouvez aussi spécifier un ordre de tri, si vous avez choisi de créer un enregistrement composite, pour certaines options dans l'onglet Composite. Pour plus d'informations, voir «Distinct - Composite - Onglet Paramètres», à la page 101.

Ordre de tri par défaut. Spécifiez si, par défaut, les enregistrements sont triés par ordre **Croissant** ou **Décroissant** en fonction des valeurs de clé de tri.

Tri des enregistrements dans le noeud Distinguer

Si l'ordre des enregistrements dans un groupe de doublons est important pour vous, vous devez spécifier l'ordre à l'aide de l'option **A l'intérieur des groupes, trier les enregistrements par** dans le noeud Distinguer. Ne vous appuyez pas sur un noeud Trier en amont. N'oubliez pas que l'ordre entrant des enregistrements n'est pas pris en compte et que seul l'ordre spécifié n'a de valeur.

Si vous ne spécifiez aucun champ de tri (ou si vous spécifiez un nombre insuffisant de champs de tri), les enregistrements dans chaque groupe de doublons ne sont pas triés (ou sont triés partiellement) et les résultats peuvent être imprévisibles.

Par exemple, imaginez que vous disposez d'un ensemble volumineux d'enregistrements de journal appartenant à plusieurs machines. Le journal contient des données telles que les suivantes :

Tableau 16. Données de journal de machine

Horodatage	Machine	Température
17:00:22	Machine A	31
13:11:30	Machine B	26
16:49:59	Machine A	30
18:06:30	Machine X	32

Tableau 16. Données de journal de machine (suite)

Horodatage	Machine	Température
16:17:33	Machine A	29
19:59:04	Machine C	35
19:20:55	Machine Y	34
15:36:14	Machine X	28
12:30:41	Machine Y	25
14:45:49	Machine C	27
19:42:00	Machine B	34
20:51:09	Machine Y	36
19:07:23	Machine X	33

Pour réduire le nombre d'enregistrements et n'afficher que le dernier enregistrement pour chaque machine, utilisez `Machine` comme champ clé et `Horodatage` comme champ de tri (par ordre décroissant). L'ordre d'entrée n'a pas d'impact sur le résultat car la sélection de tri spécifie quelle ligne parmi de nombreuses lignes pour une machine donnée doit être renvoyée. La sortie de données finale est similaire à la suivante :

Tableau 17. Données de journal de machine triées

Horodatage	Machine	Température
17:00:22	Machine A	31
19:42:00	Machine B	34
19:59:04	Machine C	35
19:07:23	Machine X	33
20:51:09	Machine Y	36

Paramètres d'optimisation distincts

Si les données sur lesquelles vous travaillez ne contiennent qu'un petit nombre d'enregistrements ou ont déjà été triées, vous pouvez optimiser la manière dont elles sont traitées pour permettre à IBM SPSS Modeler de traiter les données de manière plus efficace.

Remarque : Que vous sélectionniez l'option **Le jeu de données d'entrée comporte un petit nombre de clés distinctes.** ou utilisiez la génération SQL pour le noeud, toute ligne dans la valeur de clé distincte peut être retournée. Afin de contrôler quelle ligne est retournée pour une clé distincte, vous devez spécifier l'ordre de tri en utilisant les champs **Au sein des groupes, trier les enregistrements par** sur l'onglet Paramètres. Les options d'optimisation n'affectent pas les résultats retournés par le noeud Distinguer tant qu'un ordre de tri est spécifié sur l'onglet Paramètres.

Le jeu de données d'entrée comporte un petit nombre de clés distinctes. Sélectionnez cette option si vous avez un petit nombre d'enregistrements et/ou un petit nombre de valeurs uniques des champs-clés. Ainsi, vous améliorez les performances.

Le jeu de données d'entrée est déjà ordonnée en champs de regroupement et en champs de tri sur l'onglet Paramètres. Sélectionnez cette option uniquement si vos données sont déjà triées en fonction de tous les champs **Au sein des groupes, trier les enregistrements par** répertoriés sur l'onglet Paramètres, et si l'ordre de tri (croissant ou décroissant) est identique pour toutes les données. Ainsi, vous améliorez les performances.

Désactiver la génération SQL. Sélectionnez cette option pour désactiver la génération SQL pour le noeud.

Distinct - Composite - Onglet Paramètres

Si les données sur lesquelles vous travaillez contiennent plusieurs enregistrements, par exemple pour la même personne, vous pouvez optimiser la manière dont elles sont traitées en créant un enregistrement composite unique, ou agrégat.

Remarque : Cet onglet est disponible uniquement lorsque vous sélectionnez **Créer un enregistrement composite pour chaque groupe** sur l'onglet Paramètres.

Définition des options de l'onglet Composite

Champ. Cette colonne affiche tous les champs, à l'exception des champs-clés du modèle de données, dans leur ordre de tri naturel. Si le noeud n'est pas connecté, aucun champ n'est affiché. Pour trier alphabétiquement les lignes en fonction du nom de champ, cliquez sur l'en-tête de colonne. Vous pouvez sélectionner plusieurs lignes en utilisant la combinaison de touches Maj+clic ou Ctrl+clic. En outre, si vous cliquez sur un champ avec le bouton droit de la souris, un menu s'affiche, à partir duquel vous pouvez choisir de sélectionner toutes les lignes, de trier les lignes par valeur ou nom de champ croissant ou décroissant, de sélectionner des champs par type de mesure ou de stockage, ou de sélectionner une valeur pour ajouter automatiquement la même entrée **Remplir avec des valeurs basées sur** à chaque ligne sélectionnée.

Remplir avec des valeurs basées sur. Sélectionnez le type de valeur à utiliser pour l'enregistrement composite pour le **champ**. Les options disponibles dépendent du type de champ.

- Pour les champs d'intervalle numérique, vous pouvez choisir l'une des options suivantes :
  - Premier enregistrement du groupe
  - Dernier enregistrement du groupe
  - Total
  - Moyenne
  - Minimum
  - Maximum
  - Personnalisé
- Pour les champs de date ou d'heure, vous pouvez choisir l'une des options suivantes :
  - Premier enregistrement du groupe
  - Dernier enregistrement du groupe
  - La plus ancienne
  - La plus récente
  - Personnalisé
- Pour les champs de type chaîne ou sans type, vous pouvez choisir l'une des options suivantes :
  - Premier enregistrement du groupe
  - Dernier enregistrement du groupe
  - Premier alphanumérique
  - Dernier alphanumérique
  - Personnalisé

Dans chaque cas, vous pouvez utiliser l'option **Personnalisé** pour mieux contrôler la valeur à utiliser pour renseigner l'enregistrement composite. Pour plus d'informations, voir «Distinct - Onglet Composite - Personnalisé», à la page 102.

Inclure le comptage des enregistrements dans le champ. Sélectionnez cette option pour inclure un champ supplémentaire dans chaque enregistrement de sortie, appelé par défaut `Effectif`. Pour chaque enregistrement de sortie, ce champ indique le nombre d'enregistrements d'entrée qui ont été agrégés. Pour créer un nom personnalisé pour ce champ, saisissez votre entrée dans le champ d'édition.

Distinct - Onglet Composite - Personnalisé

La boîte de dialogue Remplissage personnalisé vous permet de mieux contrôler la valeur utilisée pour le nouvel enregistrement composite. Notez que vous devez d'abord instancier vos données avant d'utiliser cette option si vous ne personnalisez qu'une seule ligne Champ sur l'onglet Composite.

Remarque : Cette boîte de dialogue n'est disponible que lorsque vous sélectionnez la valeur Personnalisé dans la colonne **Remplir avec des valeurs basées sur** de l'onglet Composite.

En fonction du type de champ, vous pouvez choisir l'une des options ci-dessous.

- **Sélectionner par fréquence.** Choisissez une valeur en fonction de la fréquence à laquelle l'opération se produit dans l'enregistrement de données.

  Remarque : Cette option n'est pas disponible pour les champs de type Continu, Sans type ou Date/Heure.
  - **Utiliser.** Sélectionnez Le plus fréquent ou Le moins fréquent.
  - **Ex aequo.** Si deux enregistrements ou plus possèdent la même fréquence d'occurrence, indiquez le mode de sélection de l'enregistrement requis. Vous pouvez sélectionner l'une des quatre options suivantes : Utiliser le premier, Utiliser le dernier, Utiliser le plus bas ou Utiliser le plus élevé.
- **Valeurs d'inclusion (T/F).** Sélectionnez cette option pour convertir un champ en indicateur permettant d'identifier si l'un des enregistrements d'un groupe comporte une valeur donnée. Vous pouvez ensuite sélectionner la **valeur** dans la liste de celles valides pour le champ sélectionné.

  Remarque : Cette option n'est pas disponible si vous sélectionnez plusieurs lignes Champ sur l'onglet Composite.
- **Première correspondance dans la liste.** Sélectionnez cette option pour définir la priorité à affecter à l'enregistrement composite. Vous pouvez ensuite sélectionner l'un des **éléments** dans la liste de ceux valides pour le champ sélectionné.

  Remarque : Cette option n'est pas disponible si vous sélectionnez plusieurs lignes Champ sur l'onglet Composite.
- **Concaténer les valeurs.** Sélectionnez cette option pour conserver toutes les valeurs d'un groupe en les concaténant dans une chaîne. Vous devez spécifier un délimiteur à utiliser entre chaque valeur.

  Remarque : Il s'agit de la seule option disponible si vous sélectionnez une ou plusieurs lignes Champ de type Continu, Sans type ou Date/Heure.
- **Utiliser le délimiteur.** Vous pouvez choisir d'utiliser un **espace** ou une **virgule** comme délimiteur dans la chaîne concaténée. Vous pouvez également saisir votre propre délimiteur dans le champ **Autre**.

  Remarque : Cette option est disponible uniquement si vous sélectionnez l'option **Concaténer les valeurs**.

Noeud Flux TS

Le noeud Flux TS permet de générer et d'évaluer des modèles de série temporelle en une seule opération. Un modèle de série temporelle distinct est créé pour chaque champ cible. Toutefois, les nuggets de modèle ne sont pas ajoutés à la palette de modèles générés et vous ne pouvez pas parcourir les informations de modèle.

Les méthodes de modélisation des séries temporelles nécessitent l'utilisation d'un intervalle uniforme entre chaque mesure, chaque valeur manquante étant signalée par des lignes vides. Si vos données ne répondent pas à ces exigences, vous devrez transformer les valeurs si nécessaire.

Autres points à noter en relation avec les noeuds Séries temporelles :

- Les champs doivent être numériques.
- Les champs de date ne peuvent pas être utilisés comme entrées.
- Les partitions sont ignorées.

Le noeud Flux TS estime les modèles de lissage exponentiel, d'ARIMA (Autoregressive Integrated Moving Average) univariable et d'ARIMA multivariable (ou fonction de transfert) pour les séries temporelles, et il génère des prévisions basées sur les séries temporelles. Vous disposez également d'un Expert Modeler, qui tente d'identifier et d'estimer automatiquement le modèle ARIMA ou de lissage exponentiel le mieux adapté pour un ou plusieurs champs cible.

Pour plus d'informations sur la modélisation des séries temporelles, reportez-vous à la section Modèles de séries temporelles du guide Noeuds de modélisation SPSS Modeler.

Le noeud Flux TS peut être utilisé dans un environnement de déploiement de flux, via IBM SPSS Modeler Solution Publisher, à l'aide du service d'évaluation IBM SPSS Collaboration and Deployment Services.

Noeud Streaming Time Series - Options de champ

Utiliser des rôles prédéfinis Cette option utilise les paramètres de rôle (cibles, prédicteurs, etc.) d'un noeud Type en amont (ou l'onglet Types d'un noeud source en amont).

Utiliser des affectations de champs personnalisés. Pour affecter manuellement des cibles, des prédicteurs et d'autres rôles, sélectionnez cette option.

Remarque : Si vous avez partitionné vos données, les partitions sont prises en compte si vous sélectionnez **Utiliser des rôles prédéfinis**, mais elles ne le sont pas si vous sélectionnez **Utiliser des affectations de champs personnalisés**.

Champs. Utilisez les boutons fléchés pour affecter manuellement des éléments à partir de cette liste aux divers champs de rôle à droite de l'écran. Les icônes indiquent les niveaux de mesure valides pour chaque champ de rôle.

Pour sélectionner tous les champs dans la liste, cliquez sur le bouton **Tous** ou cliquez sur un bouton de niveau de mesure individuelle pour sélectionner tous les champs avec ce niveau de mesure.

Cibles Sélectionnez un ou plusieurs des champs comme cible de la prévision.

Entrées candidates Sélectionnez un ou plusieurs champs comme entrées pour la prévision.

Evénements et interventions Utilisez cette zone pour désigner certaines zones d'entrée comme champs d'événement ou d'intervention. Cette désignation identifie une zone comme contenant des données de série temporelle qui peuvent être affectées par des événements (situations récurrentes prévisibles, telles que des campagnes publicitaires) ou des interventions (incidents occasionnels, tels qu'une panne de courant ou une grève du personnel).

Noeud Streaming Time Series - Options de spécification des données

L'onglet Spécifications des données est celui où vous définissez toutes les options des données à inclure dans votre modèle. A partir du moment où vous spécifiez à la fois un **champ Date/Heure** et un

Intervalle, vous pouvez cliquer sur le bouton **Exécuter** pour créer un modèle avec toutes les options par défaut, mais normalement vous pouvez personnaliser la création selon vos propres objectifs.

L'onglet contient plusieurs sous-fenêtres dans lesquelles vous définissez les personnalisations propres à votre modèle.

Noeud Streaming Time Series - Observations

Utilisez les paramètres de cette sous-fenêtre pour spécifier les champs qui définissent les observations.

Observations spécifiées par un champ de date/heure

Vous pouvez spécifier que les observations sont définies par un champ date, heure ou horodatage. Outre le champ qui définit les observations, sélectionnez l'intervalle de temps approprié qui décrit les observations. En fonction de l'intervalle de temps donné, vous pouvez également spécifier d'autres paramètres, tels que l'intervalle compris entre des observations (incrémentation) ou le nombre de jours par semaine. Les considérations suivantes s'appliquent à l'intervalle de temps :

- Utilisez la valeur **Irrégulier** lorsque les observations sont réparties irrégulièrement dans le temps, par exemple l'heure de traitement d'une commande. Lorsque **Irrégulier** est sélectionné, vous devez spécifier l'intervalle de temps utilisé pour l'analyse, à partir des paramètres **Intervalle de temps** dans l'onglet Spécifications des données.
- Lorsque les observations représentent une date et une heure et que l'intervalle de temps est en heures, minutes, ou secondes, utilisez **Heures par jour**, **Minutes par jour** ou **Secondes par jour**. Lorsque les observations représentent une période (durée) sans référence à une date et que l'intervalle est en heures, minutes, ou secondes, utilisez **Heures (non périodique)**, **Minutes** ou **Secondes (non périodique)**.
- En fonction de l'intervalle de temps sélectionné, la procédure peut détecter des observations manquantes. La détection d'observations manquantes est nécessaire vu que la procédure suppose que toutes les observations sont également espacées dans le temps et qu'aucune observation ne manque. Par exemple, si l'intervalle de temps choisi est Jours et que la date 2015-10-27 est suivie du 2015-10-29, une observation est manquante pour la date du 2015-10-28. Des valeurs sont attribuées en cas d'observations manquantes ; utilisez la zone **Traitement des valeurs manquantes** de l'onglet Spécifications des données pour spécifier les paramètres de gestion des valeurs manquantes.
- L'intervalle de temps indiqué permet à la procédure de détecter plusieurs observations dans le même intervalle de temps lorsqu'elles ont besoin d'être regroupées. Il lui permet aussi d'aligner des observations sur une limite d'intervalle, telle que le premier du mois, pour garantir un espacement égal des observations. Par exemple, si l'intervalle de temps est Mois, plusieurs dates dans le même mois sont agrégées. Ce type d'agrégation est appelé *regroupement*. Par défaut, des observations sont additionnées lorsqu'elles sont regroupées. Vous pouvez spécifier une autre méthode de regroupement, telle que la moyenne des observations, à partir des paramètres **Agrégation et distribution** de l'onglet Spécifications des données.
- Pour certains intervalles de temps, les paramètres supplémentaires peuvent définir des interruptions dans des intervalles qui sont d'habitude espacés équitablement. Par exemple, si l'intervalle de temps est Jours, mais que seuls les jours de la semaine sont valides, vous pouvez spécifier que la semaine ne compte que cinq jours et commence le lundi.

Observations définies comme périodes ou périodes cycliques

Des observations peuvent être définies par un ou plusieurs champs de type entier qui représentent des périodes ou des cycles de répétition de périodes, jusqu'à un nombre arbitraire de niveaux de cycles. Via cette structure, vous pouvez décrire des séries d'observations qui ne se prêtent pas à un intervalle de temps standard. Par exemple, une année fiscale de seulement 10 mois peut être décrite avec un champ de cycle représentant les années et un champ de période représentant les mois, où la longueur d'un cycle est de 10.

Les champs qui spécifient les périodes cycliques définissent une hiérarchie de niveaux périodiques, où le niveau le plus bas est défini par le champ **Période**. Le niveau supérieur suivant est défini par un champ de cycle dont le niveau est 1, suivi d'un champ de cycle dont le niveau est 2, etc. Les valeurs de champ pour chaque niveau, à l'exception du plus élevé, doivent être périodiques par rapport au suivant. Les valeurs du niveau le plus élevé ne peuvent pas être périodiques. Par exemple, dans le cas de l'année fiscale sur 10 mois, les mois sont périodiques dans les années et les années ne sont pas périodiques.

- La longueur d'un cycle à un certain niveau est la périodicité du niveau immédiatement inférieur. Dans l'exemple de l'année fiscale, il existe un seul niveau de cycle et la longueur du cycle est de 10, car le niveau immédiatement inférieur suivant représente les mois et l'année fiscale compte 10 mois.
- Spécifiez la valeur de départ de chaque champ périodique ne commençant pas à 1. Ce paramètre est nécessaire pour détecter les valeurs manquantes. Par exemple, si un champ périodique commence à 2, mais avec une valeur de départ définie sur 1, la procédure suppose qu'il existe une valeur manquante pour la première période dans chaque cycle de ce champ.

Noeud Streaming Time Series - Intervalle de temps pour l'analyse

L'intervalle de temps que vous utilisez pour l'analyse peut être différent de celui des observations. Par exemple, si l'intervalle de temps des observations est spécifié comme Jours, vous pourriez choisir Mois comme intervalle de temps pour l'analyse. Les données sont alors agrégées de données quotidiennes en données mensuelles avant la construction du modèle. Vous pouvez également choisir de distribuer les données d'un intervalle de temps plus long sur un intervalle de temps plus court. Par exemple, si les observations sont trimestrielles, vous pouvez les distribuer sur une base mensuelle.

Utilisez les paramètres de cette sous-fenêtre pour spécifier l'intervalle de temps de l'analyse. La méthode d'agrégation ou de distribution des données est spécifiée depuis les paramètres **Agrégation et distribution** dans l'onglet Spécifications des données.

Les choix disponibles pour l'intervalle de temps sur lequel l'analyse est effectuée dépendent de la façon dont les données sont définies et de leur intervalle de temps. En particulier, lorsque les observations sont définies par des périodes cycliques, seule l'agrégation est prise en charge. Dans ce cas, l'intervalle d'analyse doit être supérieur ou égal à celui des observations.

Noeud Streaming Time Series - Options d'agrégation et de distribution

Utilisez les paramètres de cette sous-fenêtre pour spécifier les paramètres d'agrégation et de distribution des données d'entrée par rapport aux intervalles de temps des observations.

Fonctions d'agrégation
: Lorsque l'intervalle de temps qui est utilisé pour l'analyse est supérieur à l'intervalle des observations, les données d'entrée sont agrégées. Par exemple, l'agrégation s'effectue lorsque l'intervalle de temps des observations est Jours et l'intervalle de temps de l'analyse est Mois. Les fonctions d'agrégation suivantes sont disponibles : moyenne, somme, mode, min ou max.

Fonctions de distribution
: Lorsque l'intervalle de temps qui est utilisé pour l'analyse est inférieur à l'intervalle des observations, les données d'entrée sont distribuées. Par exemple, la distribution s'effectue lorsque l'intervalle de temps des observations est Trimestres et l'intervalle de temps de l'analyse est Mois. Les fonctions de distribution suivantes sont disponibles : moyenne ou somme.

Fonctions de regroupement
: Le regroupement est appliqué lorsque des observations sont définies par date/heures et que plusieurs observations se produisent dans le même intervalle de temps. Par exemple, si l'intervalle de temps des observations est Mois, plusieurs dates dans le même mois sont regroupées et associées au mois dans lequel elles se produisent. Les fonctions de regroupement suivantes sont disponibles : Moyenne, Somme, Mode, Min ou Max. Le regroupement est toujours effectué lorsque les observations sont définies par des dates et des heures et que leur intervalle de temps est spécifié comme étant irrégulier.

Remarque : bien qu'un regroupement soit une forme d'agrégation, il est effectué avant le traitement des valeurs manquantes, tandis que l'agrégation formelle s'effectue une fois que les valeurs manquantes ont été traitées. Lorsque l'intervalle de temps des observations est spécifié en tant que Irrégulier, l'agrégation est effectuée uniquement avec la fonction de regroupement.

Agréger les observations de la journée à la journée précédente
Spécifie si les observations qui franchissent une limite de journée doivent être agrégées aux valeurs de la journée précédente. Par exemple, pour les observations horaires avec un jour couvrant huit heures et commençant à 20h00, ce paramètre spécifie si les observations situées entre minuit et 04h00 doivent être incluses dans les résultats agrégés de la journée précédente. Ce paramètre s'applique uniquement si l'intervalle de temps des observations est Heures par jour, Minutes par jour ou Secondes par jour et si l'intervalle de temps de l'analyse est Jours.

Paramètres personnalisés pour les champs spécifiés
Vous pouvez spécifier les fonctions d'agrégation, de distribution et de regroupement champ par champ. Ces paramètres remplacent les paramètres par défaut des fonctions d'agrégation, de distribution et de regroupement.

Noeud Streaming Time Series - Options de valeurs manquantes

Utilisez les paramètres de cette sous-fenêtre pour spécifier la manière dont les valeurs manquantes des données d'entrée doivent être remplacées par une valeur imputée. Les méthodes de remplacement suivantes sont disponibles :

Interpolation linéaire
Remplace les valeurs manquantes par le biais d'une interpolation linéaire. La dernière valeur valide avant la valeur manquante et la première valeur valide après la valeur manquante sont utilisées pour l'interpolation. Si la première ou la dernière observation de la série a une valeur manquante, les deux valeurs non manquantes les plus proches au début ou à la fin de la série sont utilisées.

Moyenne de la série
Remplace les valeurs manquantes par la moyenne de toute la série.

Moyenne des points voisins
Remplace les valeurs manquantes par la moyenne des valeurs valides qui les entourent. La sphère des points voisins est le nombre de valeurs valides au-dessus et au-dessous de la valeur manquante qui sont utilisées pour calculer la moyenne.

Valeur médiane des points voisins
Remplace les valeurs manquantes par la médiane des valeurs valides qui les entourent. La sphère des points voisins est le nombre de valeurs valides au-dessus et au-dessous de la valeur manquante qui sont utilisées pour calculer la valeur médiane.

Tendance linéaire
Cette option utilise toutes les observations non manquantes dans la série afin d'alimenter un modèle de régression linéaire simple, qui est ensuite utilisé pour imputer les valeurs manquantes.

Autres paramètres :

Score de qualité de données inférieur (%)
Calcule les mesures de qualité des données pour la variable de temps et pour les données d'entrée correspondant à chaque série temporelle. Si le score de qualité des données est inférieur à ce seuil, la série temporelle correspondante sera annulée.

Noeud Streaming Time Series - Période d'estimation

Dans la sous-fenêtre Période d'estimation, vous pouvez préciser l'intervalle d'enregistrements à utiliser dans l'estimation du modèle. Par défaut, la période d'estimation commence au moment de l'observation la plus ancienne et se finit au moment de la dernière observation sur toutes les séries.

Par heure de début et de fin

Vous pouvez spécifier à la fois le début et la fin de la période d'estimation, ou juste son début ou sa fin. Si vous omettez d'indiquer le début ou la fin de la période d'estimation, la valeur par défaut est utilisée.

- Si les observations sont définies par un champ de date/heure, entrez des valeurs de début et de fin dans le même format que celui utilisé pour le champ de date/heure.
- Pour les observations définies par des périodes cycliques, spécifiez une valeur pour chacune des zones de période cyclique. Chaque champ est affiché dans une colonne distincte.

Par intervalle de temps le plus récent ou le plus ancien

Définit la période d'estimation en fonction du nombre d'intervalles de temps spécifié commençant au plus ancien ou se terminant au plus récent dans les données, avec un décalage facultatif. Dans ce contexte, l'intervalle de temps se réfère à celui de l'analyse. Supposons, par exemple, que les observations sont mensuelles mais que l'intervalle de temps de l'analyse couvre des trimestres. Si vous spécifiez **Le plus récent** avec la valeur 24 pour le **Nombre d'intervalles de temps**, ceci signifie que la période d'estimation porte sur les 24 derniers trimestres.

Vous avez la possibilité de spécifier un nombre d'intervalles de temps à exclure. Par exemple, si vous spécifiez les 24 intervalles de temps les plus récents et 1 pour le nombre d'intervalles à exclure, la période d'estimation couvrira les 24 intervalles qui précédent le plus récent.

Noeud Streaming Time Series - Options de création

L'onglet Options de création est celui où vous définissez toutes les options de création de votre modèle. Vous pouvez bien sûr cliquer simplement sur le bouton **Exécuter** pour créer un modèle avec toutes les options par défaut, mais normalement vous pouvez personnaliser la création selon vos propres objectifs.

L'onglet contient plusieurs sous-fenêtres dans lesquelles vous définissez les personnalisations propres à votre modèle.

Noeud Streaming Time Series - Options générales de création

Les options disponibles dans cette sous-fenêtre dépendent du paramètre que vous choisissez parmi les trois de la liste **Méthode** :

- **Expert Modeler.** Sélectionnez cette option pour utiliser Expert Modeler, qui recherche automatiquement le modèle le mieux adapté pour chaque série dépendante.
- **Lissage exponentiel.** Utilisez cette option pour indiquer un modèle de lissage exponentiel personnalisé.
- **ARIMA.** Utilisez cette option pour indiquer un modèle ARIMA personnalisé.

Expert Modeler

Sous **Type de modèle**, sélectionnez le type des modèles à générer :

- **Tous les modèles.** Expert Modeler considère aussi bien les modèles ARIMA que ceux de lissage exponentiel.
- **Modèles de lissage exponentiel uniquement.** Expert Modeler ne prend en compte que les modèles de lissage exponentiel.
- **Modèles ARIMA uniquement.** Expert Modeler ne prend en compte que les modèles ARIMA.

Expert Modeler prend en compte les modèles saisonniers. Cette option n'est activée que si une périodicité est définie pour le jeu de données actif. Lorsque cette option est sélectionnée, Expert Modeler prend en considération les modèles saisonniers et non saisonniers. Si cette option n'est pas sélectionnée, Expert Modeler considère uniquement les modèles non saisonniers.

Expert Modeler prend en compte les modèles de lissage exponentiel. Si cette option est sélectionnée, Expert Modeler effectue une recherche dans 13 modèles de lissage exponentiel (sept d'entre eux existaient

dans le noeud Série temporelle d'origine et six ont été ajoutés dans la version 18.1). Si cette option n'est pas sélectionnée, Expert Modeler n'effectue la recherche que dans les sept modèles de lissage exponentiel d'origine.

Sous **Valeurs extrêmes**, sélectionnez l'une des options suivantes :

Détecter les valeurs extrêmes automatiquement. Par défaut, la détection automatique de valeurs extrêmes n'est pas exécutée. Sélectionnez cette option pour procéder à la détection automatique des valeurs extrêmes, puis sélectionnez les types de valeur extrême qui vous intéressent.

Les champs d'entrée doit avoir un niveau de mesure *Indicateur*, *Nominal*, ou *Ordinal* et doivent être numériques (par exemple, 1/0, pas Vrai/Faux, pour un champ indicateur) avant d'être inclus dans cette liste.

Expert Modeler n'envisage qu'une régression simple et ignore les fonctions de transfert arbitraires pour les entrées identifiées en tant que champs d'événement ou d'intervention dans l'onglet **Champs**.

lissage exponentiel

Type de modèle. Les modèles de lissage exponentiel sont classés comme saisonniers ou non saisonniers.[1] Les modèles saisonniers ne sont disponibles que si la périodicité définies à l'aide de la sous-fenêtre Intervalles de temps de l'onglet Spécifications des données est saisonnière. Les périodicités saisonnières sont les suivantes : périodes cycliques, années, trimestres, mois, jours par semaine, heures par jour, minutes par jour et secondes par jour. Les types de modèle suivants sont disponibles :

- **Simple.** Ce modèle est approprié aux séries dont ne se dégage aucune tendance ou effet saisonnier. Son seul paramètre de lissage pertinent est le niveau. Le lissage exponentiel simple ressemble sensiblement à une ARIMA avec aucun ordre d'autorégression, un ordre de différenciation, un ordre de moyenne mobile et aucune constante.
- **Tendance linéaire de Holt.** Ce modèle est approprié aux séries dont se dégage une tendance linéaire, mais aucun effet saisonnier. Ses paramètres de lissage pertinents sont le niveau et la tendance. Dans ce modèle, la valeur d'un paramètre n'est pas tributaire de celle d'un autre paramètre. Le modèle de Holt est plus général que le modèle de Brown, mais il prend plus de temps pour calculer les estimations pour les grandes séries. Le lissage exponentiel de Holt ressemble sensiblement à une ARIMA avec aucun ordre d'autorégression, deux ordres de différenciation et deux ordres de moyenne mobile.
- **Tendance amortie.** Ce modèle est approprié aux séries dont se dégage une tendance linéaire qui s'éteint, mais aucun effet saisonnier. Ses paramètres de lissage pertinents sont le niveau, la tendance et la tendance avec amortissement. Le lissage exponentiel amorti ressemble sensiblement à une ARIMA avec un ordre d'autorégression, un ordre de différenciation et deux ordres de moyenne mobile.
- **Tendance multiplicative.** Ce modèle est approprié aux séries qui présentent une tendance qui varie en fonction de la valeur des séries et non de la saisonnalité. Ses paramètres de lissage pertinents sont le niveau et la tendance. Le lissage exponentiel de tendance multiplicatif n'est pas similaire à un modèle ARIMA.
- **Tendance linéaire de Brown.** Ce modèle est approprié aux séries dont se dégage une tendance linéaire, mais aucun effet saisonnier. Ses paramètres de lissage pertinents sont le niveau et la tendance, mais, dans ce modèle, ils sont supposés égaux. Par conséquent, le modèle de Brown est un cas particulier du modèle de Holt. Le lissage exponentiel de Brown ressemble sensiblement à une ARIMA avec aucun ordre d'autorégression, deux ordres de différenciation et deux ordres de moyenne mobile, le coefficient du deuxième ordre de moyenne mobile étant égal à la moitié du coefficient du premier ordre élevé au carré.
- **Saisonnier simple.** Ce modèle est approprié aux séries ne présentant pas de tendance et un effet saisonnier constant dans le temps. Ses paramètres de lissage pertinents sont le niveau et la saison. Le lissage exponentiel saisonnier ressemble sensiblement à une ARIMA avec aucun ordre d'autorégression,

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4, 1-28.

un ordre de différenciation, un ordre de différenciation saisonnier, ainsi que 1, *p* et *p*+1 ordres de moyenne mobile, où *p* représente le nombre de périodes dans un intervalle saisonnier. Pour les données mensuelles, *p* a pour valeur 12.

- **Modèle additif de Winters.** Ce modèle est approprié aux séries présentant une tendance linéaire et un effet saisonnier constant dans le temps. Ses paramètres de lissage pertinents sont le niveau, la tendance et la saison. Le lissage exponentiel additif de Winters ressemble sensiblement à une ARIMA avec aucun ordre d'autorégression, un ordre de différenciation, un ordre de différenciation saisonnière et *p*+1 ordres de moyenne mobile, où *p* représente le nombre de périodes dans un intervalle saisonnier. Pour les données mensuelles, *p* a pour valeur 12.
- **Tendance réduite avec saisonnalité additive.** Ce modèle est approprié aux séries présentant une tendance linéaire décroissante et un effet saisonnier constant dans le temps. Ses paramètres de lissage pertinents sont le niveau, la tendance, la tendance avec amortissement et la saison. La tendance amortie et le lissage exponentiel saisonnier additif ne sont pas similaires à un modèle ARIMA.
- **Tendance multiplicative avec saisonnalité additive.** Ce modèle est approprié aux séries qui présentent une tendance qui varie en fonction de la valeur des séries et un effet saisonnier constant dans le temps. Ses paramètres de lissage pertinents sont le niveau, la tendance et la saison. La tendance multiplicative et le lissage exponentiel saisonnier additif ne sont pas similaires à un modèle ARIMA.
- **Saisonnier multiplicatif.** Ce modèle est approprié aux séries qui ne présentent aucune tendance, mais dont l'effet saisonnier varie en fonction de la valeur des séries. Ses paramètres de lissage pertinents sont le niveau et la saison. Le lissage exponentiel saisonnier multiplicatif n'est pas similaire à un modèle ARIMA.
- **Modèle multiplicatif de Winters.** Ce modèle est approprié aux séries qui présentent une tendance linéaire et un effet saisonnier qui varie en fonction de la valeur des séries. Ses paramètres de lissage pertinents sont le niveau, la tendance et la saison. Le lissage exponentiel multiplicatif de Winter n'est pas similaire à un modèle ARIMA.
- **Tendance réduite avec saisonnalité multiplicative.** Ce modèle est approprié aux séries qui présentent une tendance linéaire décroissante et un effet saisonnier qui varie en fonction de la valeur des séries. Ses paramètres de lissage pertinents sont le niveau, la tendance, la tendance avec amortissement et la saison. La tendance avec amortissement et le lissage exponentiel saisonnier multiplicatif ne sont pas similaires à un modèle ARIMA.
- **Tendance multiplicative avec saisonnalité multiplicative.** Ce modèle est approprié aux séries qui présentent une tendance et un effet saisonnier qui varient tous deux en fonction de la valeur des séries. Ses paramètres de lissage pertinents sont le niveau, la tendance et la saison. La tendance multiplicative et le lissage exponentiel saisonnier multiplicatif ne sont pas similaires à un modèle ARIMA.

Transformation cible. Vous pouvez indiquer qu'une transformation doit être effectuée sur chaque variable dépendante avant que celle-ci ne soit modélisée.

- **Aucun.** Aucune transformation n'est effectuée.
- **Racine carrée.** Une transformation par la racine carrée est effectuée.
- **Log naturel.** Une transformation par le logarithme naturel est effectuée.

ARIMA

Spécifiez la structure d'un modèle ARIMA personnalisé.

Ordres ARIMA. Entrez des valeurs pour les divers composants ARIMA de votre modèle dans les cellules correspondantes de la grille. Toutes les valeurs doivent être des entiers non négatifs. Pour les composants autorégressifs et de moyenne mobile, la valeur représente l'ordre maximum. Tous les ordres inférieurs positifs sont inclus dans le modèle. Par exemple, si vous spécifiez 2, le modèle comprend les ordres 2 et 1. Les cellules de la colonne Saisonnier ne sont activées que si une périodicité est définie pour le jeu de données actif.

- **Autorégressif (p).** Le nombre d'ordres autorégressifs dans le modèle. Les ordres autorégressifs indiquent quelles valeurs précédentes de la série seront utilisées pour prévoir les valeurs en cours. Par exemple, un ordre autorégressif de 2 indique que la valeur de la série de deux périodes antérieures est utilisée pour prévoir la valeur en cours.
- **Différence (d).** Spécifie l'ordre de différenciation appliqué à la série avant d'estimer les modèles. La différenciation est nécessaire lorsque les tendances sont présentes (les séries avec tendances sont en général non stationnaires et la modélisation ARIMA suppose la stationnarité) et est utilisée pour supprimer leurs effets. L'ordre de différenciation correspond au degré de tendance de série ; aux comptes de différenciation de premier ordre pour les tendances linéaires, aux comptes de différenciation de second ordre pour les tendances quadratiques, etc.
- **Moyenne mobile (q).** Le nombre d'ordres de moyenne mobile dans le modèle. Les ordres de moyenne mobile indiquent comment les écarts de la moyenne de la série pour les valeurs précédentes sont utilisés pour prévoir les valeurs en cours. Par exemple, des ordres de moyenne mobile de 1 et 2 indiquent que les écarts de la valeur de la moyenne de la série pour chacune des deux dernières périodes doivent être considérés lors de la prévision des valeurs actuelles de la série.

Saisonnier. Les composants autorégressifs, de moyenne mobile et de différenciation saisonniers tiennent le même rôle que leurs équivalents non saisonniers. Cependant, pour les ordres saisonniers, les valeurs en cours de la série sont affectées par les valeurs de série précédentes séparées par une ou plusieurs périodes saisonnières. Par exemple, pour des données mensuelles (période saisonnière de 12), un ordre saisonnier de 1 indique que la valeur de série en cours est affectée par les 12 périodes de la valeur de série précédant celle en cours. Un ordre saisonnier de 1, pour des données mensuelles, est alors le même que lorsqu'on spécifie un ordre non saisonnier de 12.

Détecter les valeurs extrêmes automatiquement. Sélectionnez cette option pour procéder à la détection automatique des valeurs extrêmes et sélectionnez un ou plusieurs types de valeur extrême disponibles.

Type de valeurs extrêmes à détecter. Sélectionnez les types de valeur extrême à détecter. Les types pris en charge sont :

- Additif (sélectionné par défaut)
- Changement de niveau (sélectionné par défaut)
- Innovation
- Transitoire
- Additif saisonnier
- Tendance locale
- Correctif additif

Ordres de fonction de transfert et transformations. Pour spécifier des transformations et définir les fonctions de transfert pour certains ou tous les champs d'entrée de votre modèle ARIMA, cliquez sur **Définir** ; une boîte de dialogue distincte vous permet d'entrer les détails du transfert et de la transformation.

Inclure la constante dans le modèle. L'inclusion d'une constante est normale, à moins que vous ne soyez sûr que la valeur de série de la moyenne générale est 0. L'exclusion de la constante est recommandée lorsque la différenciation s'applique.

Fonctions de transfert et transformation : La boîte de dialogue Ordres de fonction de transfert et transformations permet de spécifier les transformations et de définir des fonctions de transfert pour les champs d'entrée de votre choix dans votre modèle ARIMA.

Transformations cible. Dans cette sous-fenêtre, vous pouvez indiquer qu'une transformation doit être effectuée sur chaque variable cible avant que celle-ci ne soit modélisée.

- **Aucun.** Aucune transformation n'est effectuée.

- **Racine carrée.** Une transformation par la racine carrée est effectuée.
- **Log naturel.** Une transformation par le logarithme naturel est effectuée.

Transformation et fonctions de transfert des entrées candidates. Les fonctions de transfert permettent de spécifier la manière dont les valeurs passées des champs d'entrée sont utilisées pour la prévision des valeurs futures des séries cible. La liste de gauche de la sous-fenêtre affiche tous les champs d'entrée. Les autres informations de cette sous-fenêtre sont spécifiques au champ d'entrée que vous sélectionnez.

Ordres de fonction de transfert. Entrez des valeurs pour les différents composants de la fonction de transfert dans les cellules correspondantes de la grille **Structure**. Toutes les valeurs doivent être des entiers non négatifs. Pour les composants numérateur et dénominateur, la valeur représente l'ordre maximum. Tous les ordres inférieurs positifs sont inclus dans le modèle. En outre, l'ordre 0 est toujours inclus pour les composants numérateur. Par exemple, si vous spécifiez 2 comme numérateur, le modèle comprend les ordres 2, 1 et 0. Si vous indiquez 3 en guise de dénominateur, le modèle comporte les ordres 3, 2 et 1. Les cellules de la colonne Saisonnier ne sont activées que si une périodicité est définie pour le jeu de données actif.

Numérateur. L'ordre du numérateur de la fonction de transfert détermine les valeurs précédentes de la série indépendante sélectionnée (prédicteur) qui sont utilisées pour la prévision des valeurs actuelles de la série dépendante. Par exemple, un ordre de numérateur de 1 indique que la valeur d'une série indépendante d'une période antérieure, en plus de la valeur en cours de la série indépendante, sont utilisées pour prévoir la valeur en cours de chaque série dépendante.

Dénominateur. L'ordre du dénominateur de la fonction de transfert détermine la manière dont les écarts par rapport à la moyenne de la série, pour les valeurs précédentes de la série indépendante sélectionnée (prédicteur), sont utilisés pour la prévision des valeurs actuelles de la série dépendante. Par exemple, un ordre de dénominateur de 1 indique que les écarts de la valeur moyenne d'une série indépendante d'une période antérieure sont pris en compte lors de la prévision de la valeur en cours de chaque série dépendante.

Différence. Spécifie l'ordre de différenciation appliqué à la série indépendante (prédicteur) avant d'estimer les modèles. La différenciation est nécessaire lorsque les tendances sont présentes et est utilisée pour supprimer leur effet.

Saisonnier. Les composants numérateur, dénominateur et de différenciation saisonniers tiennent le même rôle que leurs équivalents non saisonniers. Cependant, pour les ordres saisonniers, les valeurs en cours de la série sont affectées par les valeurs de série précédentes séparées par une ou plusieurs périodes saisonnières. Par exemple, pour des données mensuelles (période saisonnière de 12), un ordre saisonnier de 1 indique que la valeur de série en cours est affectée par les 12 périodes de la valeur de série précédant celle en cours. Un ordre saisonnier de 1, pour des données mensuelles, est alors le même que lorsqu'on spécifie un ordre non saisonnier de 12.

Délai. La définition d'un délai retarde l'influence du champ d'entrée du nombre d'intervalles spécifié. Par exemple, si le délai a pour valeur 5, la valeur du champ d'entrée au moment t n'affectera les prévisions qu'une fois les cinq périodes écoulées ($t + 5$).

Transformation. La spécification d'une fonction de transfert pour un ensemble de variables indépendantes comprend une transformation facultative à exécuter sur ces variables.
- **Aucun.** Aucune transformation n'est effectuée.
- **Racine carrée.** Une transformation par la racine carrée est effectuée.
- **Log naturel.** Une transformation par le logarithme naturel est effectuée.

Noeud Streaming Time Series - Options de modèle

Limite de confiance (%). Les intervalles de confiance sont calculés pour les prévisions et les autocorrélations résiduelles du modèle. Vous pouvez spécifier n'importe quelle valeur positive inférieure à 100. Par défaut, un intervalle de confiance de 95 % est utilisé.

L'option **Etendre les enregistrements dans le futur** définit le nombre d'intervalles de temps sur lesquels effectuer des prévisions au-delà de la période d'estimation. L'intervalle de temps dans ce cas est celui de l'analyse, que vous spécifiez dans l'onglet Spécifications des données. Lorsque des prévisions sont demandées, des modèles autorégressifs sont construits automatiquement pour chaque série d'entrée ne constituant pas également une cible. Ces modèles sont ensuite utilisés pour générer des valeurs pour ces séries d'entrée sur la période de prévision. Il n'existe pas de limite maximale pour ce paramètre.

Valeurs futures à utiliser dans la prévision

- **Calculer les valeurs futures des entrées** Si vous sélectionnez cette option, les valeurs de prévision des prédicteurs, des prévisions de bruit, de l'estimation de la variance et des valeurs des moments ultérieurs sont calculées. Lorsque des prévisions sont demandées, des modèles autorégressifs sont construits automatiquement pour chaque série d'entrée ne constituant pas également une cible. Ces modèles sont ensuite utilisés pour générer des valeurs pour ces séries d'entrée sur la période de prévision.
- **Sélectionnez les champs dont vous souhaitez ajouter les valeurs aux données.** Pour chaque enregistrement à prévoir (à l'exclusion des éléments restants), si vous utilisez des champs prédicteurs (dont le rôle est défini à `Input`), vous pouvez indiquer les valeurs estimées pour la période de prévision pour chaque prédicteur. Vous pouvez indiquer les valeurs manuellement ou les choisir dans une liste.
  - **Champ.** Cliquez sur le bouton de sélection de champ et choisissez les champs à utiliser en tant que prédicteurs. Notez que les champs sélectionnés ici ne sont pas nécessairement utilisés dans la modélisation. Pour qu'un champ soit réellement utilisé en tant que prédicteur, vous devez le sélectionner dans un noeud de modélisation en aval. Cette boîte de dialogue est cependant pratique pour définir les valeurs futures, pour qu'elles puissent être partagées par plusieurs noeuds de modélisation en aval, sans les définir individuellement dans chaque noeud. Par ailleurs, la liste des champs disponibles peut être soumise à des contraintes dépendantes des sélections effectuées dans l'onglet Options de génération.

    Notez que si des valeurs futures sont indiquées pour un champ qui n'est plus disponible dans le flux (car il a été supprimé ou parce que de nouvelles sélections ont été faites dans l'onglet Créer), le champ apparaît en rouge.
  - **Valeurs.** Pour chaque champ, vous pouvez faire un choix dans une liste de fonctions ou cliquer sur **Spécifier** pour entrer des valeurs manuellement ou faire un choix dans la liste des valeurs prédéfinies. Si les champs prédicteurs correspondent à des éléments que vous contrôlez ou qui peuvent être déterminés par avance pour d'autres raisons, saisissez ces valeurs manuellement. Par exemple, si vous prévoyez les bénéfices du mois prochain d'un hôtel sur la base du nombre de réservations, vous pouvez indiquer le nombre de réservations réel pour cette période. A l'inverse, si un champ prédicteur est en rapport avec un élément hors de votre contrôle, tel que le cours d'une valeur mobilière, vous pouvez utiliser une fonction telle que la valeur la plus récente ou la moyenne des points récents.

    Les fonctions disponibles dépendent du niveau de mesure du champ.

Tableau 18. Fonctions disponibles pour les niveaux de mesure

Niveau de mesure	Fonctions
Champ continu ou nominal	Vide Moyenne des points récents Valeur la plus récente Spécifier

Tableau 18. Fonctions disponibles pour les niveaux de mesure (suite)

Niveau de mesure	Fonctions
Champ booléen	Vide Valeur la plus récente Vrai Faux Spécifier

Moyenne des points récents calcule la valeur future à partir de la moyenne des trois derniers points de données.

Valeur la plus récente. Définit la valeur future en fonction du point de données le plus récent.

Vrai/Faux. Définit la valeur future d'un champ indicateur comme Vrai ou Faux, selon le cas.

Spécifier. Ouvre une boîte de dialogue qui permet de définir manuellement les valeurs futures ou de les choisir dans une liste prédéfinie.

Rendre disponible pour l'évaluation

Vous pouvez définir ici les valeurs par défaut des options de scoring qui apparaissent dans la boîte de dialogue du nugget de modèle.

- **Calculer les limites de confiance supérieure et inférieure.** Si cette option est sélectionnée, elle permet de créer des champs (dotés des préfixes par défaut $TSLCI- et $TSUCI-) pour les valeurs inférieure et supérieure de l'intervalle de confiance, et ce, pour chaque champ cible.
- **Calculer les résidus de bruit.** Si cette option est sélectionnée, elle permet de créer un champ (doté du préfixe par défaut $TSResidual-) pour les résidus de modèle de chaque champ cible et de calculer le total de ces valeurs.

Paramètres de modèle

Nombre maximal de modèles à afficher dans la sortie. Spécifiez le nombre maximal de modèles à inclure dans la sortie. Notez que si le nombre de modèles générés dépasse ce seuil, les modèles ne sont pas affichés dans la sortie mais restent disponibles pour l'évaluation. La valeur par défaut est 10. L'affichage d'un trop grand nombre de modèles peut avoir comme conséquence des performances médiocres ou l'instabilité du système.

Noeud SMOTE

Le noeud Synthetic Minority Over-sampling Technique (SMOTE) fournit un algorithme de suréchantillonnage permettant de traiter les jeux de données déséquilibrés. Il offre une méthode avancée d'équilibrage des données. Le noeud de traitement SMOTE est implémenté en Python et requiert la bibliothèque Python `imbalanced-learn`. Pour plus d'informations sur la bibliothèque imbalanced-learn, voir http://contrib.scikit-learn.org/imbalanced-learn/about.html[1].

L'onglet Python de la palette de noeuds contient le noeud SMOTE et d'autres noeuds Python.

[1]Lemaître, Nogueira, Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research*, vol. 18, no. 17, 2017, pp. 1-5. (http://jmlr.org/papers/v18/16-365.html)

Paramètres du noeud SMOTE

Définissez les paramètres suivants sous l'onglet **Paramètres** du noeud SMOTE.

Paramètre cible

Champ cible. Sélectionnez le champ cible. Tous les types de mesure discret, nominal, ordinal ou indicateur sont pris en charge. Si l'option **Utiliser des données partitionnées** est sélectionnée dans la section Partition, seules les données d'apprentissage seront suréchantillonnées.

Rapport de suréchantillonnage

Sélectionnez **Automatique** pour sélectionner automatiquement un ratio de suréchantillonnage ou sélectionnez **Définir le rapport (minorité sur majorité)** pour définir un ratio personnalisé. Le ratio correspond au nombre d'échantillons dans la classe minoritaire sur le nombre d'échantillons dans la classe majoritaire. La valeur doit être supérieure à **0** et inférieure ou égale à **1**.

Valeur de départ aléatoire

Définir une valeur de départ aléatoire. Sélectionnez cette option et cliquez sur **Générer** pour générer la valeur de départ utilisée par le générateur de nombres aléatoires.

Méthodes

Type d'algorithme. Sélectionnez le type d'algorithme SMOTE que vous souhaitez utiliser.

Règles d'échantillonnage

K_Neighbours. Spécifiez le nombre de voisins les plus proches à utiliser pour la construction d'échantillons synthétiques

M_Neighbours. Spécifiez le nombre de voisins les plus proches à utiliser pour déterminer si un échantillon minoritaire est en danger. Cette règle n'est utilisée que si le type d'algorithme SMOTE **Borderline1** ou **Borderline1** est sélectionné.

Partition

Utiliser les données partitionnées. Sélectionnez cette option si vous souhaitez que seules les données d'apprentissage soient suréchantillonnées.

Le noeud SMOTE requiert la bibliothèque Python `imbalanced-learn`. Le tableau suivant présente la relation entre les paramètres de la boîte de dialogue du noeud SMOTE SPSS Modeler et l'algorithme Python.

Tableau 19. Relation entre les propriétés du noeud et les paramètres de la bibliothèque Python

Paramètre SPSS Modeler	Nom du script (nom de la propriété)	Nom du paramètre de l'API Python
Rapport de suréchantillonnage (contrôle d'entrée des nombres)	`sample_ratio_value`	`ratio`
Valeur de départ aléatoire	`random_seed`	`random_state`
K_Neighbours	`k_neighbours`	`k`
M_Neighbours	`m_neighbours`	`m`
Type d'algorithme	`algorithm_kind`	`kind`

noeud Extension Transform

Avec le noeud de transformation d'extension, vous pouvez prendre les données d'un flux IBM SPSS Modeler et appliquer les transformations aux données à l'aide de script R ou Python for Spark. Une fois modifiées, les données sont renvoyées dans le flux à des fins de traitement, de génération de modèle et de scoring de modèle. Le noeud de transformation d'extension rend possible la transformation des données à l'aide d'algorithmes écrits en R ou Python for Spark et permet à l'utilisateur de développer des méthodes de transformation de données adaptées à une situation particulière.

Pour utiliser ce noeud avec R, IBM SPSS Modeler - Essentials for R doit être installé. Pour les instructions d'installation et des informations sur la compatibilité, voir *IBM SPSS Modeler - Essentials for R: Installation Instructions*. Vous devez également disposer d'une version compatible de R sur votre ordinateur.

Noeud de transformation d'extension - Onglet Syntaxe

Sélectionnez le type de syntaxe – **R** ou **Python for Spark**. Consultez les sections suivantes pour plus d'informations. Lorsque votre syntaxe est prête, vous pouvez cliquer sur **Exécuter** pour exécuter le noeud de transformation d'extension.

Syntaxe R

Syntaxe R. Permet d'entrer ou de coller la syntaxe de script R personnalisé en vue de l'analyse des données dans ce champ.

Convertir les champs indicateurs. Indique comment sont traités les champs indicateurs. Deux options sont disponibles : **Chaînes en facteur, Entiers et Réels en double** et **Valeurs logiques (True, False)**. Si vous sélectionnez **Valeurs logiques (True, False)**, les valeurs originales des champs indicateurs sont perdues. Par exemple, si un champ a les valeurs Mâle et Femelle, elles sont remplacées par True et False.

Convertir les valeurs manquantes en valeur R 'non disponible' (NA). Lorsque cette option est sélectionnée, toute valeur manquante est convertie en valeur R `NA`. La valeur `NA` est utilisée par R pour identifier les valeurs manquantes. Des fonctions R que vous utilisez peuvent comporter un argument par le biais duquel il est possible de contrôler le comportement des fonctions lorsque les données contiennent `NA`. Par exemple, la fonction peut vous permettre de choisir d'exclure automatiquement les enregistrements qui contiennent `NA`. Si cette option n'est pas sélectionnée, les valeurs manquantes sont transmises à R en l'état et peuvent entraîner des erreurs lors de l'exécution du script R.

Convertir les champs date/heure en classes R avec contrôle spécial pour les fuseaux horaires. Lorsque cette option est sélectionnée, les variables de format de date et de date/heure sont converties en objets R date/heure. Vous devez sélectionner l'une des options suivantes :

- **R POSIXct.** Les variables de format de date ou de date/heure sont converties en objets R `POSIXct`.
- **R POSIXlt (liste).** Les variables de format de date ou de date/heure sont converties en objets R `POSIXlt`.

Remarque : Les formats POSIX sont des options avancées. Utilisez-les uniquement si le script R spécifie que les champs date/heure sont traités de telle manière que ces formats sont requis. Les formats POSIX ne s'appliquent pas aux variables de format horaire.

Syntaxe Python

Syntaxe Python. Permet d'entrer ou de coller la syntaxe de script Python personnalisé en vue de l'analyse des données dans ce champ. Pour plus d'informations relatives à Python for Spark, voir Python for Spark et Scriptage avec Python for Spark.

Noeud de transformation d'extension - Sortie de la console

L'onglet **Sortie de la console** contient les sorties reçues lorsque le script R ou le script Python for Spark de l'onglet Syntaxe est exécuté (par exemple, si un script R est utilisé, il affiche la sortie reçue de la console R lorsque le script R du champ **Syntaxe R** de l'onglet **Syntaxe** est exécuté). La sortie peut contenir des messages d'erreur ou d'avertissement R ou Python générés lors de l'exécution du script R ou Python. Cette sortie permet essentiellement de déboguer le script. L'onglet **Sortie de la console** contient également le script du champ **Syntaxe R** ou **Syntaxe Python**.

A chaque exécution du script de transformation d'extension, le contenu de l'onglet **Sortie de la console** est écrasé par la sortie reçue de la console R ou Python for Spark. La sortie ne peut pas être éditée.

Noeud Boîtes espace-temps

Les noeuds Boîtes espace-temps constituent une extension des emplacements spatiaux avec Geohash. Un noeud Boîtes espace-temps est en particulier une chaîne alphanumérique qui représente une zone d'espace et de temps de forme régulière.

Par exemple, le noeud Boîtes espace-temps **dr5ru7 | 2013-01-01 00:00:00 | 2013-01-01 00:15:00** se compose des trois parties suivantes :

- Le geohash **dr5ru7**
- L'horodatage de début **2013-01-01 00:00:00**
- L'horodatage de fin **2013-01-01 00:15:00**

Par exemple, vous pouvez utiliser les informations d'espace et de temps pour vous assurer que deux entités sont identiques car elles sont virtuellement au même endroit au même moment. Vous pouvez également améliorer l'exactitude de l'identification de la relation en montrant que deux entités sont liées en raison de leur proximité dans l'espace et le temps.

Vous pouvez choisir le mode **Enregistrements individuels** ou **Affluences** en fonction de vos besoins. Ces deux modes requièrent les mêmes informations de base, comme suit :

Champ de latitude. Sélectionnez le champ qui identifie la latitude (dans le système de coordonnées WGS84).

Champ de longitude. Sélectionnez le champ qui identifie la longitude (dans le système de coordonnées WGS84).

Champ d'horodatage. Sélectionnez le champ qui identifie la date ou l'heure.

Options du mode Enregistrements individuels

Utilisez ce mode pour ajouter un champ supplémentaire à un enregistrement afin d'identifier son emplacement à un moment donné.

Calculer. Sélectionnez une ou plusieurs densités d'espace et de temps à partir desquelles dériver le nouveau champ. Consultez «Définition de la densité Space-Time-Box», à la page 118 pour plus d'informations.

Extension nom de champ. Entrez l'extension à ajouter aux nouveaux noms de champ. Vous pouvez choisir d'ajouter cette extension sous forme de **suffixe** ou de **préfixe**.

Options du mode Affluences

Une affluence peut être un emplacement et/ou un moment où une entité est détectée continuellement ou de façon répétée. Par exemple, il peut être utilisé pour identifier un véhicule qui effectue un transport régulier et pour identifier les écarts-types par rapport à la norme.

Le détecteur d'affluence contrôle le déplacement des entités et signale les conditions dans lesquelles il constate qu'une entité "afflue" dans la zone. Il affecte automatiquement chaque affluence signalée à une ou plusieurs boîtes espace-temps et utilise le suivi des entités en mémoire et des événements pour détecter les affluences avec une efficacité optimale.

Densité STB. Sélectionnez la densité d'espace et de temps à partir de laquelle dériver le nouveau champ. Par exemple, la valeur **STB_GH4_10MINS** correspond à une boîte geohash de quatre caractères dont la taille est de 20 km par 20 km et dont la fenêtre de temps est de 10 minutes. Voir «Définition de la densité Space-Time-Box», à la page 118 pour plus d'informations.

Champ ID d'entité. Sélectionnez l'entité à utiliser comme identificateur d'affluence. Le champ ID identifie l'événement.

Nombre minimal d'événements. Un événement est une ligne dans les données. Sélectionnez le nombre minimal d'occurrences d'un événement pour l'entité à prendre en compte comme bloquée. Une affluence peut aussi être qualifiée en fonction du champ **La durée de résidence est au moins**.

La durée de résidence est au moins. Indiquez la durée minimale pendant laquelle une entité doit résider dans le même emplacement. Par exemple, cela peut empêcher une voiture qui attend à un feu d'être considérée comme bloquée. Une affluence peut aussi être qualifiée en fonction du champ **Nombre minimal d'événements** précédent.

Vous trouverez ci-après des détails supplémentaires sur la qualification d'une affluence :

$e_1, \ldots, e_n$ indique tous les événements triés par heure qui sont reçus d'un ID d'entité donné au cours d'une période (t_1, t_n). Ces événements sont considérés comme une affluence si :

- n >= *nombre minimal d'événements*
- $t_n - t_1$ >= *durée de résidence minimale*
- Tous les événements $e_1, \ldots, e_n$ se produisent dans la même boîte espace-temps

Autoriser les affluences à étendre les limites STB. Si cette option est sélectionnée, la définition d'une affluence est moins stricte et peut inclure, par exemple, une entité qui est bloquée dans plusieurs noeuds Boîtes espace-temps. Par exemple, si vos boîtes espace-temps sont définies comme des heures pleines, la sélection de cette option permet de reconnaître comme valide une entité qui reste bloquée pendant une heure, même si l'heure est composée des 30 minutes avant minuit et des 30 minutes après minuit. Si cette option n'est pas sélectionnée, 100 % du temps d'affluence doit se trouver dans une boîte espace-temps unique.

Proportion min. d'événements dans la durée qualifiante (%). Cette option est disponible uniquement si **Autoriser les affluences à étendre les limites STB** est sélectionné. Utilisez ce bouton pour contrôler la mesure dans laquelle une affluence signalée dans un noeud STB peut chevaucher une autre affluence. Sélectionnez la proportion minimale d'événements qui doivent se produire dans un noeud STB unique pour identifier une affluence. Si elle est définie sur 25 % et que la proportion d'événements est de 26 %, il s'agit d'une affluence.

Par exemple, supposez que vous configurez le détecteur d'affluence pour qu'il requière au moins deux événements (nombre minimal d'événements = 2) et un temps d'existence contiguë d'au moins 2 minutes dans un espace geohash de 4 octets et une fenêtre de temps de 10 minutes (STB_NAME = STB_GH4_10MINS). Lorsqu'une affluence est détectée, imaginez que l'entité existe dans le même espace

geohash de 4 octets alors que les trois événements qualifiants surviennent dans une fenêtre de temps de 10 minutes entre 16h57 et 15h07 à 16h58, 17h01 et 17h03. La valeur en pourcentage de la fenêtre de temps qualifiante spécifie les boîtes espace-temps prises en compte pour l'affluence, comme suit :

- **100 %.** L'affluence est signalée dans la fenêtre de temps 17h00 à 17h10 et non dans la fenêtre de temps 16h50 à 17h00 (les événements 17h01 et 17h03 remplissent toutes les conditions requises pour une affluence qualifiante et 100 % de ces événements se sont produits dans la fenêtre de temps 17h00 à 17h10).
- **50 %.** Les affluences dans les deux fenêtres de temps sont signalées (les événements 17h01 et 17h03 remplissent toutes les conditions requises pour une affluence qualifiante, au moins 50 % de ces événements se sont produits dans la fenêtre de temps 16h50 à 17h00 et au moins 50 % de ces événements se sont produits dans la fenêtre de temps 17h00 à 17h10).
- **0 %.** Les affluences dans les deux fenêtres de temps sont signalées.

Lorsque 0 % est spécifié, les rapports sur les affluences incluent les boîtes espace-temps représentant chaque fenêtre de temps affectée par la durée qualifiante. La durée qualifiante doit être inférieure ou égale à la durée correspondante de la fenêtre de temps dans la boîte espace-temps. En d'autres termes, il ne doit jamais y avoir une configuration selon laquelle une boîte espace-temps de 10 minutes est configurée en tandem avec une durée qualifiante de 20 minutes.

Une affluence est signalée dès que les conditions qualifiantes sont remplies et n'est pas signalée plus d'une fois par boîte espace-temps. Supposez que trois événements satisfont les conditions d'affluence et que 10 événements au total surviennent au cours d'une durée qualifiante, dans la même boîte espace-temps. Dans ce cas, l'affluence est signalée lorsque le troisième événement qualifiant se produit. Aucun des sept autres événements ne déclenche de rapport d'affluence.

Remarque :

- Les données d'événement en mémoire du détecteur d'affluence ne sont pas partagées entre les processus. Par conséquent, une entité particulière possède une affinité avec un noeud de détecteur d'affluence spécifique. En d'autres termes, les données de mouvement entrantes pour une entité doivent toujours être transmises de façon cohérente au noeud de détecteur d'affluence qui effectue le suivi de cette entité, et qui est généralement le même noeud tout au long de l'exécution.
- Les données d'événement en mémoire du détecteur d'affluence sont volatiles. A chaque fois que vous quittez le détecteur d'affluence, les affluences détectées sont perdues. Cela signifie que si vous arrêtez et redémarrez le processus, le système peut ne pas signaler certaines affluences réelles. Pour remédier à ce problème, vous pouvez relire certaines des données de mouvement historiques (par exemple, vous pouvez revenir 48 heures en arrière et relire les enregistrements de mouvement applicables à un noeud qui a été redémarré).
- Vous devez fournir les données au détecteur d'affluence dans un ordre séquentiel.

Définition de la densité Space-Time-Box

Choisissez la taille (densité) de vos noeuds Space-Time-Boxes (STB) en spécifiant la surface physique et le temps écoulé à inclure dans chacun d'eux.

Densité géographique. Sélectionnez la taille de la zone à inclure dans chaque noeud STB.

Intervalle de temps. Sélectionnez le nombre d'heures à inclure dans chaque noeud STB.

Nom du champ. Il contient le préfixe STB et est automatiquement complété en fonction des sélections que vous avez effectuées dans les deux champs précédents.

Noeud Streaming TCM

Le noeud Streaming TCM permet de créer et d'évaluer des modèles causals temporels en une seule opération.

Pour plus d'informations sur la modélisation causale temporelle, voir la rubrique Modèles causals temporel dans la section Modèles de séries temporelles du guide Noeuds de modélisation SPSS Modeler.

Noeud Streaming TCM - Options Série temporelle

Dans l'onglet Champs, utilisez les paramètres **Série temporelle** pour préciser la série à inclure dans le système modèle.

Sélectionnez l'option de structure de données qui s'applique à vos données. Pour les données multidimensionnelles, cliquez sur **Sélectionner les dimensions** pour spécifier les champs de dimension. L'ordre des champs de dimension indiqués définit l'ordre dans lequel ils apparaissent dans tous les boîtes de dialogue et sorties ultérieures. Utilisez les boutons de flèche vers le haut et le bas dans la sous-boîte de dialogue Sélectionnez les dimensions pour réorganiser les champs de dimension.

Pour les données basées sur des colonnes, le terme *série* a le même sens que le terme *champ*. Pour les données multidimensionnelles, les champs contenant des séries temporelles sont appelés des champs *métriques*. Une série temporelle, pour les données multidimensionnelles, est définie par un champ métrique et une valeur pour chacun des champs de dimension. Les considérations suivantes s'appliquent aux données basées sur les colonnes et aux données multidimensionnelles.

- Les séries qui sont spécifiées en tant qu'entrées candidates ou à la fois comme séries cible et en entrée sont prises en compte pour être incluses dans le modèle de chaque cible. Le modèle de chaque cible inclut toujours des valeurs en attente de la cible elle-même.
- Les séries qui sont spécifiées comme entrées forcées sont toujours incluses dans le modèle de chaque cible.
- Au moins une série doit être spécifiée comme cible ou à la fois comme cible et entrée.
- Lorsque **Utiliser des rôles prédéfinis** est sélectionné, les champs prédéfinis qui ont un rôle d'entrée sont définis comme entrées candidates. Aucun rôle prédéfini n'est mappé avec une entrée forcée.

Données multidimensionnelles

Pour les données multidimensionnelles, vous spécifiez des champs métriques et leurs rôles associés dans une grille, où chaque ligne de la grille indique une seule métrique et un seul rôle. Par défaut, le système modèle comprend une série pour toutes les combinaisons des champs de dimension pour chaque ligne de la grille. Par exemple, s'il existe des dimensions pour *région* et *marque*, par défaut, le fait de spécifier la métrique *ventes* comme cible signifie qu'il existe une série de cibles de ventes distinctes pour chaque combinaison de *région* et de *marque*.

Pour chaque ligne de la grille, vous pouvez personnaliser l'ensemble des valeurs des champs de dimension en cliquant sur le bouton de points de suspension d'une dimension. Cette action ouvre la sous-boîte de dialogue Sélectionnez des valeurs pour Dimension. Vous pouvez également ajouter, supprimer ou copier des lignes de la grille.

La colonne **Comptage des séries** affiche le nombre des ensembles de valeurs de dimension actuellement spécifiées pour la métrique associée. La valeur affichée peut être supérieure au nombre réel de séries (une série par ensemble). Cette condition se produit lorsque certaines des combinaisons spécifiées de valeurs de dimension ne correspondent pas aux séries contenues dans la métrique associée.

Noeud Streaming TCM - Sélectionnez une valeur pour Dimension

Pour les données multidimensionnelles, vous pouvez personnaliser l'analyse en spécifiant les valeurs de dimension qui s'appliquent à un champ métrique particulier avec un rôle particulier. Par exemple, si *ventes* est un champ de métrique et *canal*, une dimension avec les valeurs 'détail' et 'Web,' vous pouvez indiquer que les ventes 'Web' correspondent à une entrée et les ventes 'détail', à une cible. Vous pouvez également spécifier des sous-ensembles de dimension qui s'appliquent à tous les champs de métrique utilisés dans l'analyse. Par exemple, si *région* est un champ de dimension qui indique une région géographique, vous pouvez limiter l'analyse à des régions particulières.

Toutes les valeurs
Indique que toutes les valeurs du champ de dimension en cours sont incluses. Il s'agit de l'option par défaut.

Sélectionnez les valeurs à inclure ou à exclure
Cette option permet de spécifier l'ensemble des valeurs du champ de dimension en cours. Quand **Inclure** est sélectionné pour **Mode**, seules les valeurs spécifiées dans la liste **Valeurs sélectionnées** sont incluses. Quand **Exclure** est sélectionné pour **Mode**, toutes les valeurs autres que celles spécifiées dans la liste **Valeurs sélectionnées** sont incluses.

Vous pouvez filtrer l'ensemble des valeurs dans lesquelles choisir. Les valeurs conformes à la condition de filtrage apparaissent dans l'onglet **Correspondance** et celles qui n'y correspondent pas figurent dans l'onglet **Sans correspondance** de la liste **Valeurs non sélectionnées**. L'onglet **Toutes** recense toutes les valeurs non sélectionnées, sans considération de condition de filtrage quelconque.

- Vous pouvez utiliser des astérisques (*) comme caractères génériques lorsque vous spécifiez un filtre.
- Pour effacer le filtre en cours, spécifiez une valeur vide pour le terme de recherche dans la boîte de dialogue Filtrer les valeurs affichées.

Noeud Streaming TCM - Options Observations

Dans l'onglet Champs, utilisez les paramètres **Observations** pour préciser les champs qui définissent les observations.

Observations définies par date/heures

Vous pouvez spécifier que les observations sont définies par un champ de date, d'heure, ou de date/heure. Outre le champ qui définit les observations, sélectionnez l'intervalle de temps approprié qui décrit les observations. En fonction de l'intervalle de temps donné, vous pouvez également spécifier d'autres paramètres, tels que l'intervalle compris entre des observations (incrémentation) ou le nombre de jours par semaine. Les considérations suivantes s'appliquent à l'intervalle de temps :

- Utilisez la valeur **Irrégulier** lorsque les observations sont réparties irrégulièrement dans le temps, par exemple l'heure de traitement d'une commande. Lorsque **Irrégulier** est sélectionné, vous devez spécifier l'intervalle de temps utilisé pour l'analyse, à partir des paramètres **Intervalle de temps** dans l'onglet Spécifications des données.
- Lorsque les observations représentent une date et une heure et que l'intervalle de temps est en heures, minutes, ou secondes, utilisez **Heures par jour**, **Minutes par jour** ou **Secondes par jour**. Lorsque les observations représentent une période (durée) sans référence à une date et que l'intervalle est en heures, minutes, ou secondes, utilisez **Heures (non périodique)**, **Minutes** ou **Secondes (non périodique)**.
- En fonction de l'intervalle de temps sélectionné, la procédure peut détecter des observations manquantes. La détection des observations manquantes est nécessaire, car la procédure considère que toutes les observations sont équitablement espacées dans le temps et qu'il n'existe aucune observation manquante. Par exemple, si l'intervalle de temps est Jours et si la date 2014-10-27 est suivie de 2014-10-29, il existe une observation manquante pour la date du 2014-10-28. Des valeurs sont imputées à toutes les observations manquantes. Les paramètres de gestion des valeurs manquantes peuvent être spécifiés dans l'onglet Spécifications des données.
- L'intervalle de temps indiqué permet à la procédure de détecter plusieurs observations dans le même intervalle de temps lorsqu'elles ont besoin d'être regroupées. Il lui permet aussi d'aligner des observations sur une limite d'intervalle, telle que le premier du mois, pour garantir un espacement égal des observations. Par exemple, si l'intervalle de temps est Mois, plusieurs dates dans le même mois sont agrégées. Ce type d'agrégation est appelé *regroupement*. Par défaut, des observations sont additionnées lorsqu'elles sont regroupées. Vous pouvez spécifier une autre méthode de regroupement, telle que la moyenne des observations, à partir des paramètres **Agrégation et distribution** de l'onglet Spécifications des données.

- Pour certains intervalles de temps, les paramètres supplémentaires peuvent définir des interruptions dans des intervalles qui sont d'habitude espacés équitablement. Par exemple, si l'intervalle de temps est Jours, mais que seuls les jours de la semaine sont valides, vous pouvez spécifier que la semaine ne compte que cinq jours et commence le lundi.

Observations définies par des périodes ou des périodes cycliques

Des observations peuvent être définies par un ou plusieurs champs de type entier qui représentent des périodes ou des cycles de répétition de périodes, jusqu'à un nombre arbitraire de niveaux de cycles. Avec cette structure, vous pouvez décrire des séries d'observations qui ne correspondent pas à l'un des intervalles de temps standard. Par exemple, une année fiscale de seulement 10 mois peut être décrite avec un champ de cycle représentant les années et un champ de période représentant les mois, où la longueur d'un cycle est de 10.

Les champs qui spécifient les périodes cycliques définissent une hiérarchie de niveaux périodiques, où le niveau le plus bas est défini par le champ **Période**. Le niveau supérieur suivant est défini par un champ de cycle dont le niveau est 1, suivi d'un champ de cycle dont le niveau est 2, etc. Les valeurs de zone de chaque niveau, sauf la plus élevée, doivent être périodiques par rapport au niveau supérieur suivant. Les valeurs du niveau le plus élevé ne peuvent pas être périodiques. Par exemple, dans le cas de l'année fiscale sur 10 mois, les mois sont périodiques dans les années et les années ne sont pas périodiques.

- La longueur d'un cycle à un certain niveau est la périodicité du niveau immédiatement inférieur. Dans l'exemple de l'année fiscale, il existe un seul niveau de cycle et la longueur du cycle est de 10, car le niveau immédiatement inférieur suivant représente les mois et l'année fiscale compte 10 mois.
- Spécifiez la valeur de départ de tout champ périodique qui ne commence pas à 1. Ce paramètre est nécessaire pour détecter les valeurs manquantes. Par exemple, si un champ périodique commence à 2, mais avec une valeur de départ définie sur 1, la procédure suppose qu'il existe une valeur manquante pour la première période dans chaque cycle de ce champ.

Noeud Streaming TCM - Options Intervalle de temps

L'intervalle de temps utilisé pour l'analyse peut être différent de celui des observations. Par exemple, si l'intervalle de temps des observations est spécifié comme Jours, vous pourriez choisir Mois comme intervalle de temps pour l'analyse. Les données sont alors agrégées de données quotidiennes en données mensuelles avant la construction du modèle. Vous pouvez également choisir de distribuer les données d'un intervalle de temps plus long sur un intervalle de temps plus court. Par exemple, si les observations sont trimestrielles, vous pouvez les distribuer sur une base mensuelle.

Les choix disponibles pour l'intervalle de temps sur lequel l'analyse est effectuée dépendent de la façon dont les données sont définies et de leur intervalle de temps. En particulier, lorsque les observations sont définies par des périodes cycliques, seule l'agrégation est prise en charge. Dans ce cas, l'intervalle d'analyse doit être supérieur ou égal à celui des observations.

L'intervalle de temps pour l'analyse est spécifié dans les paramètres **Intervalle de temps** sur l'onglet Spécifications des données. La méthode d'agrégation ou de distribution des données est spécifiée depuis les paramètres **Agrégation et distribution** dans l'onglet Spécifications des données.

Noeud Streaming TCM - Options Agrégation et Distribution

Fonctions d'agrégation

Lorsque l'intervalle de temps qui est utilisé pour l'analyse est supérieur à l'intervalle des observations, les données d'entrée sont agrégées. Par exemple, l'agrégation s'effectue lorsque l'intervalle de temps des observations est Jours et l'intervalle de temps de l'analyse est Mois. Les fonctions d'agrégation suivantes sont disponibles : moyenne, somme, mode, min ou max.

Fonctions de distribution

Lorsque l'intervalle de temps qui est utilisé pour l'analyse est inférieur à l'intervalle des

observations, les données d'entrée sont distribuées. Par exemple, la distribution s'effectue lorsque l'intervalle de temps des observations est Trimestres et l'intervalle de temps de l'analyse est Mois. Les fonctions de distribution suivantes sont disponibles : moyenne ou somme.

Fonctions de regroupement
Le regroupement est appliqué lorsque des observations sont définies par date/heures et que plusieurs observations se produisent dans le même intervalle de temps. Par exemple, si l'intervalle de temps des observations est Mois, plusieurs dates dans le même mois sont regroupées et associées au mois dans lequel elles se produisent. Les fonctions de regroupement suivantes sont disponibles : Moyenne, Somme, Mode, Min ou Max. Le regroupement est toujours effectué lorsque les observations sont définies par des dates et des heures et que leur intervalle de temps est spécifié comme étant irrégulier.

Remarque : bien qu'un regroupement soit une forme d'agrégation, il est effectué avant le traitement des valeurs manquantes, tandis que l'agrégation formelle s'effectue une fois que les valeurs manquantes ont été traitées. Lorsque l'intervalle de temps des observations est spécifié en tant que Irrégulier, l'agrégation est effectuée uniquement avec la fonction de regroupement.

Agréger les observations de la journée à la journée précédente
Indique si des observations dont les heures chevauchent deux journées sont agrégées avec les valeurs de la journée précédente. Par exemple, pour des observations horaires dans une journée de huit heures qui commence à 20:00, ce paramètre indique si des observations intervenant entre 00:00 et 04:00 sont incluses. Ce paramètre s'applique uniquement si l'intervalle de temps des observations est Heures par jour, Minutes par jour ou Secondes par jour et si l'intervalle de temps de l'analyse est Jours.

Paramètres personnalisés pour les champs spécifiés
Vous pouvez spécifier les fonctions d'agrégation, de distribution et de regroupement champ par champ. Ces paramètres remplacent les paramètres par défaut des fonctions d'agrégation, de distribution et de regroupement.

Noeud Streaming TCM - Options Valeur manquante

Les valeurs manquantes dans les données d'entrée sont remplacées par une valeur saisie. Les méthodes de remplacement suivantes sont disponibles :

Interpolation linéaire
Remplace les valeurs manquantes par le biais d'une interpolation linéaire. La dernière valeur valide avant la valeur manquante et la première valeur valide après la valeur manquante sont utilisées pour l'interpolation. Si la première ou la dernière observation de la série a une valeur manquante, les deux valeurs non manquantes les plus proches au début ou à la fin de la série sont utilisées.

Moyenne de la série
Remplace les valeurs manquantes par la moyenne de toute la série.

Moyenne des points voisins
Remplace les valeurs manquantes par la moyenne des valeurs valides qui les entourent. La sphère des points voisins est le nombre de valeurs valides au-dessus et au-dessous de la valeur manquante qui sont utilisées pour calculer la moyenne.

Valeur médiane des points voisins
Remplace les valeurs manquantes par la médiane des valeurs valides qui les entourent. La sphère des points voisins est le nombre de valeurs valides au-dessus et au-dessous de la valeur manquante qui sont utilisées pour calculer la valeur médiane.

Tendance linéaire
Cette option utilise toutes les observations non manquantes dans la série afin d'alimenter un modèle de régression linéaire simple, qui est ensuite utilisé pour imputer les valeurs manquantes.

Autres paramètres :

Pourcentage maximal de valeurs manquantes (%)
Spécifie le pourcentage maximal de valeurs manquantes autorisé pour une série. Les séries avec un nombre de valeurs manquantes supérieur à ce maximum sont exclues de l'analyse.

Noeud Streaming TCM - Options générales pour les données

Nombre maximal de valeurs distinctes par champ de dimension
Ce paramètre s'applique aux données multidimensionnelles et spécifie le nombre maximal de valeurs distinctes permises pour un champ de dimension quelconque. Par défaut, cette limite est fixée à 10000, mais vous pouvez l'augmenter jusqu'au nombre de votre choix.

Noeud Streaming TCM - Options de création générales

Largeur de l'intervalle de confiance (%)
Ce paramètre contrôle les intervalles de confiance pour les prévisions et pour les paramètres du modèle. Vous pouvez spécifier n'importe quelle valeur positive inférieure à 100. Par défaut, un intervalle de confiance de 95 % est utilisé.

Nombre maximal d'entrées pour chaque cible
Ce paramètre spécifie le nombre maximal d'entrées autorisé dans le modèle pour chaque cible. Vous pouvez spécifier un entier sur la plage 1 à 20. Le modèle pour chaque cible inclut toujours ses propres valeurs décalées, et donc si vous spécifiez la valeur 1, la seule entrée est la cible elle-même.

Tolérance du modèle
Ce paramètre spécifie le processus itératif utilisé pour déterminer le jeu d'entrées optimal pour chaque cible. Vous pouvez spécifier n'importe quelle valeur supérieure à zéro. Valeur par défaut : 0.001. La tolérance du modèle est un critère d'arrêt pour la sélection de prédicteur. Elle peut affecter le nombre de prédicteurs inclus dans le modèle final. Mais même si une cible peut se prévoir très bien elle-même, d'autres prédicteurs peuvent ne pas être inclus dans le modèle final. Plusieurs tâtonnements peuvent être nécessaires (par exemple, si vous définissez une valeur trop élevée, vous pourriez la réduire pour voir si d'autres prédicteurs sont inclus ou non).

Seuil valeurs extrêmes (%)
Une observation est marquée comme étant une valeur extrême si la probabilité, calculée depuis le modèle, qu'il s'agisse d'une valeur extrême dépasse ce seuil. Vous pouvez spécifier une valeur dans la plage de 50 à 100.

Nombre de décalages pour chaque entrée
Ce paramètre spécifie le nombre de termes de décalage pour chaque entrée pour chaque cible dans le modèle. Par défaut, le nombre de termes de décalage est déterminé automatiquement à partir de l'intervalle de temps utilisé pour l'analyse. Par exemple, si l'intervalle de temps est défini à Mois (avec un incrément d'un mois), le nombre de décalages est alors de 12. Vous avez la possibilité de spécifier explicitement le nombre de décalages. La valeur spécifiée doit être un entier sur la plage 1 à 20.

Poursuivre l'estimation en utilisant les modèles existants
Si vous avez déjà généré un modèle de causalité temporaire, sélectionnez cette option pour réutiliser les paramètres de critère spécifiés pour ce modèle au lieu d'en construire un nouveau. De la sorte, vous pouvez gagner du temps en générant une nouvelle prévision basée sur les mêmes paramètres de modèle qu'auparavant, mais faisant appel à des données plus récentes.

Noeud Streaming TCM - Options Période d'estimation

Par défaut, la période d'estimation commence au moment de l'observation la plus ancienne et se finit au moment de la dernière observation sur toutes les séries.

Par heure de début et de fin

Vous pouvez spécifier à la fois le début et la fin de la période d'estimation, ou juste son début ou sa fin. Si vous omettez d'indiquer le début ou la fin de la période d'estimation, la valeur par défaut est utilisée.

- Si les observations sont définies par un champ de date/heure, entrez des valeurs de début et de fin dans le même format que celui utilisé pour le champ de date/heure.
- Pour les observations définies par des périodes cycliques, spécifiez une valeur pour chacune des zones de période cyclique. Chaque champ est affiché dans une colonne distincte.

Par intervalle de temps le plus récent ou le plus ancien

Définit la période d'estimation en fonction du nombre d'intervalles de temps spécifié commençant au plus ancien ou se terminant au plus récent dans les données, avec un décalage facultatif. Dans ce contexte, l'intervalle de temps se réfère à celui de l'analyse. Supposons, par exemple, que les observations sont mensuelles mais que l'intervalle de temps de l'analyse couvre des trimestres. Si vous spécifiez **Le plus récent** avec la valeur 24 pour le **Nombre d'intervalles de temps**, ceci signifie que la période d'estimation porte sur les 24 derniers trimestres.

Vous avez la possibilité de spécifier un nombre d'intervalles de temps à exclure. Par exemple, si vous spécifiez les 24 intervalles de temps les plus récents et 1 pour le nombre d'intervalles à exclure, la période d'estimation couvrira les 24 intervalles qui précédent le plus récent.

Noeud Streaming TCM - Options Modèle

Nom du modèle

Vous pouvez spécifier un nom personnalisé pour le modèle ou accepter celui généré automatiquement, à savoir *TCM*.

Prévision

L'option **Etendre les enregistrements dans le futur** définit le nombre d'intervalles de temps sur lesquels effectuer des prévisions au-delà de la période d'estimation. L'intervalle de temps dans ce cas est celui de l'analyse, tel que spécifié dans l'onglet Spécifications des données. Lorsque des prévisions sont demandées, des modèles autorégressifs sont construits automatiquement pour chaque série d'entrée ne constituant pas également une cible. Ces modèles sont ensuite utilisés pour générer des valeurs pour ces séries d'entrée sur la période de prévision. Il n'existe pas de limite maximale pour ce paramètre.

Noeud Optimisation CPLEX

Le noeud Optimisation CPLEX permet d'utiliser l'optimisation mathématique complexe (CPLEX) via un fichier de modèle OPL (Optimization Programming Language). Cette fonctionnalité est disponible dans le produit IBM Analytical Decision Management, mais désormais, vous pouvez également utiliser le noeud CPLEX dans SPSS Modeler sans IBM Analytical Decision Management.

Pour plus d'informations sur l'optimisation CPLEX et sur OPL, voir la documentation d'IBM ILOG CPLEX Optimization Studio.

Le noeud Optimisation CPLEX prend en charge plusieurs sources de données, ou plusieurs données entrantes dimensionnelles. Vous pouvez connecter plusieurs noeuds au noeud Optimisation CPLEX, et chaque noeud antérieur peut être utilisé pour fournir des données au calcul du modèle OPL (déini comme des ensembles de tuples individuels contenant des mappages de champs individuels).

Lors de la sortie des données générées par le noeud Optimisation CPLEX, les données originales venant des sources de données peuvent donner un résultat sous la forme d'un seul index ou de plusieurs index dimensionnels.

Remarque : Lorsque vous exécutez un flux contenant un noeud Optimisation CPLEX dans **IBM SPSS Modeler Server**, la bibliothèque CPLEX intégrée de l'édition Community est utilisée par défaut. Elle

comporte une limitation de 1000 variables et de 1000 contraintes. Si vous installez l'édition complète d'IBM ILOG CPLEX et préférez le moteur CPLEX de l'édition complète, qui n'est pas soumis à ces limitations, exécutez l'étape suivante pour votre plateforme.

- Dans Windows, éditez `options.cfg` et ajoutez le chemin d'accès à la bibliothèque OPL. Par exemple :
  ```
  cplex_opl_lib_path="<chemin_CPLEX>\opl\bin\<rép_Plateforme>"
  ```
  où `<chemin_CPLEX>` est le répertoire d'installation CPLEX tel que `C:\Program Files\IBM\ILOG\CPLEX_Studio127`, et `<rép_Plateforme>` est le répertoire de la plateforme, par exemple `x64_win64`.
- Sous Linux, éditez `modelersrv.sh` et ajoutez le chemin d'accès à la bibliothèque OPL. Par exemple :
  ```
  CPLEX_OPL_LIB_PATH= <chemin_CPLEX> /opl/bin/ <rép_Plateforme>
  ```
  où `<chemin_CPLEX>` est le répertoire d'installation CPLEX tel que `/root/Libs_127_FullEdition/Linux_x86_64`, et `<rép_Plateforme>` est le répertoire de la plateforme, par exemple `x86-64_linux`.

Remarque : Lorsque vous exécutez un flux contenant un noeud d'optimisation de CPLEX dans **SPSS Modeler Solution Publisher**, la bibliothèque CPLEX intégrée de l'édition Community est utilisée par défaut. Elle comporte une limitation de 1000 variables et de 1000 contraintes. Si vous installez l'édition complète d'IBM ILOG CPLEX et préférez le moteur CPLEX de l'édition complète, qui n'est pas soumis à ces limitations, exécutez l'étape suivante pour votre plateforme.

- Sous Windows, ajoutez le chemin d'accès à la bibliothèque sous forme d'argument de ligne de commande pour `modelerrun.exe`. Par exemple :
  ```
  -o cplex_opl_lib_path="<chemin_CPLEX>\opl\bin\<rép_Plateforme>"
  ```
  où `<chemin_CPLEX>` est le répertoire d'installation CPLEX tel que `C:\Program Files\IBM\ILOG\CPLEX_Studio127`, et `<rép_Plateforme>` est le répertoire de la plateforme, par exemple `x64_win64`.
- Sous Linux, éditez `modelerrun` et ajoutez le chemin d'accès à la bibliothèque OPL. Par exemple :
  ```
  CPLEX_OPL_LIB_PATH= <chemin_CPLEX> /opl/bin/ <rép_Plateforme>
  ```
  où `<chemin_CPLEX>` est le répertoire d'installation CPLEX tel que `/root/Libs_127_FullEdition/Linux_x86_64`, et `<rép_Plateforme>` est le répertoire de la plateforme, par exemple `x86-64_linux`.

Définition des options du noeud Optimisation CPLEX

L'onglet Options du noeud Optimisation CPLEX contient les champs ci-après.

Fichier de modèle OPL. Sélectionnez un fichier de modèle OPL (Optimization Programming Language).

Modèle OPL. Une fois que vous avez sélectionné un modèle OPL, son contenu est affiché ici.

Données d'entrée

Dans l'onglet Données d'entrée, le menu déroulant **Source de données** répertorie toutes les sources de données (noeuds précédents) connectées au noeud Optimisation CPLEX actuel. La sélection d'une source de données dans le menu déroulant actualise la section **Input Mappings** ci-dessous. Cliquez sur **Apply All Fields** afin de générer automatiquement tous les mappages de fichiers pour la source de données sélectionnée. La table **Input Mappings** sera remplie automatiquement.

Entrez le nom de l'ensemble de tuples dans le modèle OPL correspondant aux données entrantes. Ensuite, si nécessaire, vérifiez que tous les champs de tuple sont mappés aux champs d'entrée des données selon leur ordre dans la définition du tuple.

Après avoir défini le mappage d'entrée pour une source de données, vous pouvez sélectionner une autre source de données dans le menu déroulant et répéter l'opération. Les précédents mappages de source de données seront enregistrés automatiquement. Cliquez sur **Appliquer** ou **OK** lorsque vous avez terminé.

Autres données

Dans l'onglet Autres données, utilisez la section **Données OPL** si vous devez spécifier d'autres données pour l'optimisation.

Sortie

Lorsque la sortie est une variable de décision, elle doit utiliser les sources de données précédentes (données entrantes) en tant qu'index, et les index doivent être prédéfinis dans la section **Input Mappings** de l'onglet Données d'entrée. Actuellement, aucun autre type de variables de décision n'est pris en charge. La variable de décision peut avoir un seul index ou plusieurs index. SPSS Modeler génère les résultats CPLEX avec une partie ou l'ensemble des données originales entrantes, ce qui est conforme aux autres noeuds SPSS Modeler. Les index correspondants indiqués doivent être spécifiés dans le champ **Tuple de sortie** décrit ci-dessous.

Dans l'onglet Sortie, choisissez le mode de sortie (**Sortie brute** ou **Variable de décision**) et spécifiez d'autres options si nécessaire. L'option Sortie brute génère directement la valeur de la fonction d'objectif, quel que soit le nom.

Nom de la variable de valeur de la fonction d'objectif dans l'OPL. Ce champ est activé si vous avez sélectionné le mode de sortie **Variable de décision**. Entrez le nom de la variable de valeur de la fonction d'objectif dans le modèle OPL.

Nom de la zone de valeur de la fonction d'objectif pour la sortie. Entrez le nom de champ à utiliser dans la sortie. La valeur par défaut est `_OBJECTIVE`.

Tuple de sortie. Entrez le nom du tuple prédéfini à partir des données entrantes. Il agit comme les index de la variable de décision, et est supposé être généré avec les sorties de variables. Le tuple de sortie doit être conforme à la définition de la variable de décision du modèle OPL. S'il existe plusieurs index, les noms des tuples doivent être séparés par une virgule (`,`).

Sorties de variables. Ajoutez une ou plusieurs variables à inclure dans la sortie.

Chapitre 4. Noeuds d'opérations sur les champs

Présentation des opérations sur les champs

Après une première exploration des données, vous devrez peut-être sélectionner, nettoyer ou élaborer des données en vue d'une préparation à l'analyse. La palette Opérations sur les champs contient de nombreux noeuds utiles aux opérations de transformation et de préparation.

Par exemple, à l'aide d'un noeud Calculer, vous pouvez créer un attribut qui n'est pas actuellement représenté dans les données. Vous pouvez également utiliser un noeud Discrétiser pour recoder automatiquement les valeurs de champ de l'analyse cible. Vous aurez certainement recours fréquemment au noeud type ; en effet, il permet d'attribuer un niveau de mesure, des valeurs et un rôle de modélisation à chaque champ du jeu de données. Ces opérations sont utiles pour la gestion des valeurs manquantes et la modélisation en aval.

La palette Opérations sur les champs contient les noeuds suivants :

Le noeud de préparation automatisée de données (ADP) peut analyser vos données, identifier des corrections et filtrer des champs qui sont problématiques ou qui sont peu susceptibles d'être utiles. Il peut aussi créer de nouveaux attributs le cas échéant et améliorer la performance au moyen de techniques de filtrage et d'échantillonnage intelligentes. Vous pouvez utiliser le noeud de manière totalement automatisée, en laissant le noeud choisir et appliquer les corrections, ou vous pouvez prévisualiser les modifications avant qu'elles ne soient mises en place et les accepter, les rejeter ou les modifier selon les besoins.

Le noeud Typer définit les propriétés et métadonnées de champ. Par exemple, vous pouvez indiquer un niveau de mesure (continu, nominal, ordinal ou indicateur) pour chaque champ, définir des options pour la gestion des valeurs manquantes et des valeurs système nulles, spécifier le rôle d'un champ en vue de la modélisation, définir des libellés de champ et de valeur, et indiquer les valeurs d'un champ.

Le noeud Filtrer filtre (supprime) les champs, les renomme et les mappe entre un noeud source et un autre.

Le noeud Calculer modifie les valeurs de données ou crée des nouveaux champs à partir d'un ou de plusieurs champs existants. Il crée des champs de type formule, indicateur, ensemble, nominal, statistiques, comptage et conditionnel.

Le noeud Ensemble combine deux ou plusieurs nuggets de modèles pour obtenir des prévisions plus précises que celles acquises à partir d'un modèle quelconque.

Le noeud Remplacer permet de remplacer les valeurs de champ et de modifier le type de stockage. Vous pouvez décider de remplacer les valeurs reposant sur une condition CLEM, telle que `@BLANK(@FIELD)`. Vous pouvez également choisir de remplacer tous les blancs ou toutes les valeurs nulles par une valeur précise. Un noeud Remplacer est souvent associé à un noeud Typer pour remplacer les valeurs manquantes.

Le noeud Anonymiser transforme la façon dont les noms et les valeurs des champs sont représentés en aval, masquant ainsi les données d'origine. Cela peut s'avérer utile si vous souhaitez permettre à d'autres utilisateurs de générer des modèles utilisant des données confidentielles, par exemple des noms de clients ou autre.

Le noeud Recoder permet de transformer un ensemble de valeurs catégorielles en un autre. La recodification est utile pour réduire des catégories ou regrouper des données à analyser.

Le noeud Discrétiser crée automatiquement des champs nominaux (ensemble) sur la base des valeurs d'un ou de plusieurs champs continus (intervalle numérique) existants. Par exemple, vous pouvez transformer un champ continu de revenus en un nouveau champ catégoriel contenant des groupes de revenus comme écarts par rapport à la moyenne. Une fois les intervalles du nouveau champ créés, vous pouvez générer un noeud Calculer à partir des points de césure.

Le noeud Analyse RFM (Récence, Effectif, Monétaire) permet de déterminer de façon quantitative les clients susceptibles d'être les meilleurs par l'étude de leur dernier achat (récence), l'effectif de leurs achats (effectif), et la somme dépensée lors de toutes les transactions (monétaire).

Le noeud Partitionner génère un champ de partition qui répartit les données dans des sous-ensembles distincts pour les étapes d'apprentissage, de test et de validation de la création d'un modèle.

Le noeud Binariser calcule plusieurs champs indicateurs en fonction des valeurs catégorielles définies pour un ou plusieurs champs nominaux.

Le noeud Restructurer convertit un champ nominal ou un champ indicateur en un groupe de champs renseignés à partir des valeurs d'un autre champ. Par exemple, si l'on considère un champ nommé *type de paiement*, qui comporte les valeurs *crédit*, *liquide* et *débit*, trois champs sont alors créés (*crédit*, *liquide*, *débit*), chacun contenant la valeur du paiement réel effectué.

Le noeud Transposer fait passer les données des lignes vers les colonnes (et réciproquement) de sorte que les enregistrements deviennent des champs et les champs des enregistrements.

Utilisez le noeud Intervalles de temps pour spécifier des intervalles et calculer un nouveau champ Heure pour l'estimation ou la prévision. Un ensemble complet d'intervalles de temps, allant des secondes aux années, est pris en charge.

Le noeud Historiser crée des champs contenant des données provenant de champs d'enregistrements antérieurs. Les noeuds Historiser sont souvent utilisés pour les données séquentielles, telles que les séries temporelles. Avant d'utiliser un noeud Historiser, vous pouvez trier les données à l'aide d'un noeud Trier.

Le noeud Re-trier définit l'ordre naturel utilisé pour afficher les champs situés en aval. Cet ordre a une incidence sur l'affichage des champs en différents endroits : tableaux, listes et sélecteur de champs. Cette opération est utile lorsque vous utilisez des jeux de données volumineux pour rendre plus visibles les champs intéressants.

Dans SPSS Modeler, les éléments tels que les fonctions spatiales du générateur de formules, le noeud de prévision spatio-temporelle (STP) et le noeud Visualisation de carte utilisent le système de coordonnées projetées. Utilisez le noeud Reprojeter pour changer le système de coordonnées des données que vous importez et qui utilisent un système de coordonnées géographiques.

Certains de ces noeuds peuvent être générés directement à partir du rapport d'audit créé par un noeud Audit données. Pour plus d'informations, voir «Génération d'autres noeuds en vue d'une préparation de données», à la page 340.

Préparation automatique des données

La préparation des données pour l'analyse est une des étapes les plus importantes des projets et généralement, l'une de celles qui prend le plus de temps. La préparation automatique des données (ADP) s'occupe de cette tâche à votre place, analyse vos données, identifie les corrections, supprime les champs problématiques ou inutiles, dérive de nouveaux attributs si nécessaire et améliore les performances grâce à des techniques de balayage intelligentes. Vous pouvez utiliser l'algorithme en mode complètement **automatique**, le laissant choisir et appliquer les corrections ou vous pouvez utiliser son mode **interactif** qui prévoit les modifications avant qu'elles ne soient effectuées vous laissant libre de les accepter ou de les refuser.

L'utilisation de l'ADP vous permet de préparer facilement et rapidement vos données pour la génération de modèle, sans qu'il soit nécessaire de maîtriser les concepts de statistiques utilisés. Les modèles seront alors créés et les scores déterminés plus rapidement ; de plus, l'utilisation de l'ADP améliore la robustesse des processus de modélisation automatique , tels que les actualisations de modèles et les champion / challenger.

Remarque : lorsque la préparation automatique des données prépare un champ pour l'analyse, elle crée un nouveau champ contenant les ajustements ou les transformations, au lieu de remplacer les valeurs et les propriétés existantes de l'ancien champ. L'ancien champ n'est pas utilisé pour l'analyse, son rôle est défini sur Aucun.

Exemple. Une compagnie d'assurances disposant de ressources restreintes pour enquêter sur les demandes de remboursement des propriétaires de biens immobiliers, souhaite construire un modèle pour

signaler des réclamations suspectes et potentiellement frauduleuses. Avant de construire le modèle, il est nécessaire de préparer les données à l'aide de la préparation automatique des données. La compagnie souhaitant être capable de consulter et modifier les transformations avant de les appliquer, elle utilise la préparation automatique des données de manière interactive.

Un groupe automobile suit les ventes de véhicules automobiles personnels divers. Afin d'être en mesure d'identifier les modèles dont les ventes sont très satisfaisantes et ceux pour lesquels elles le sont moins, des responsables du groupe souhaitent établir une relation entre les ventes de véhicules et les caractéristiques des véhicules. Ils utilisent la préparation automatique des données pour cette analyse afin de construire des modèles à l'aide des données " avant" et " après " la préparation et de pouvoir en comparer les résultats.

Quel est votre objectif ? La préparation automatique des données recommande des étapes de préparation de données qui amélioreront la vitesse de création de modèles par les autres algorithmes et le pouvoir prédictif de ces modèles. Cela peut comprendre la transformation, la construction et la sélection de fonctions. La cible peut également être transformée. Vous pouvez spécifier les priorités de création de modèle sur lesquelles le processus de préparation des données doit se concentrer.

- **Equilibrer la vitesse et la précision.** Cette option prépare les données à accorder la même importance à la vitesse à laquelle les données sont traitées par les algorithmes de création de modèle et à la précision des prévisions.
- **Optimiser la vitesse.** Cette option prépare les données à accorder la priorité à la vitesse à laquelle les données sont traitées par les algorithmes de création de modèle. Lorsque vous travaillez avec de très grands jeux de données ou que vous recherchez une réponse rapide, sélectionnez cette option.
- **Optimiser l'exactitude.** Cette option prépare les données à accorder la priorité à la précision des prédictions produites par les algorithmes de création de modèle.
- **Analyse personnalisée.** Lorsque vous souhaitez modifier manuellement l'algorithme dans l'onglet Paramètres, sélectionnez cette option. Veuillez noter que ce paramètre est automatiquement sélectionné si vous modifiez ensuite des options dans l'onglet Paramètres qui ne sont pas compatibles avec l'un des autres objectifs.

Formation du noeud

Le noeud ADP est mis en oeuvre en tant que noeud de processus et fonctionne de la même façon qu'un noeud Type ; la **formation** du noeud ADP correspond à l'instanciation du noeud Type. Lorsque l'analyse est terminée, les transformations spécifiées sont appliquées aux données sans analyse supplémentaire tant que le modèle des données en amont ne change pas. Tout comme les noeuds Type et Filtre, si le noeud ADP est déconnecté, il se souvient du modèle de données et des transformations et n'a pas besoin d'être de nouveau formé lorsqu'il est reconnecté. Cela vous permet de le former sur un sous-jeu de données standard puis de le déployer ou de le copier pour l'utiliser avec des données en direct aussi souvent que possible.

Utilisation de la barre d'outils

La barre d'outils permet d'exécuter et de mettre à jour l'affichage de l'analyse des données et de générer des noeuds pouvant être utilisés en conjonction avec les données d'origine.

- **Générer** Depuis ce menu, vous pouvez générer un noeud Filtre ou un noeud Calculer. Veuillez noter que ce menu est uniquement disponible lorsqu'une analyse apparaît dans l'onglet Analyse.

  Le noeud Filtre supprime les champs d'entrée transformés. Si vous configurez le noeud ADP pour qu'il laisse les champs d'entrée d'origine dans l'ensemble de donnés, cela restaure l'ensemble d'entrées d'origine et vous permet d'interpréter le champ des scores en terme d'entrées. Par exemple, cela peut être utile si vous souhaitez produire un graphique du champ de scores par rapport aux différentes entrées.

  Le noeud Calculer peut restaurer le jeu de données d'origine et les unités cibles. Vous ne pouvez générer un noeud Calculer que lorsque le noeud ADP contient une analyse qui rééchelonne une cible

plage (c'est-à-dire que le rééchelonnement de Box-Cox est sélectionné dans le panneau Préparer les entrées & la cible). Vous ne pouvez pas générer de noeud Calculer si la cible n'est pas une plage, ou si le rééchelonnement de Box-Cox n'est pas sélectionné. Pour plus d'informations, voir «Génération d'un noeud Calculer», à la page 143.

- **Afficher** Contient des options qui contrôlent ce qui apparaît dans l'onglet Analyse. Cela comprend les contrôles de modification des graphiques et les sélections d'affichage à la fois pour le panneau principal et les vues liées.
- **Aperçu** Affiche un échantillon des transformations qui seront appliquées aux données d'entrée.
- **Analyser les données** Démarre une analyse avec les paramètres actuels et affiche les scores dans l'onglet Analyse.
- **Effacer l'analyse** Supprime l'analyse existante (disponible uniquement lorsqu'une analyse en cours existe).

Statut du noeud

Le statut du noeud ADP sur le canevas IBM SPSS Modeler est indiqué par une flèche ou par une graduation sur l'icône qui indique si l'analyse a eu lieu ou pas.

Pour plus d'informations sur les calculs effectués avec le noeud Préparation automatique de données, reportez-vous à la section relatives aux *algorithmes de préparation automatique des données*, dans le document *Guide des algorithmes d'IBM SPSS Modeler*. Ce document est disponible au format PDF, dans le répertoire `\Documentation` du disque d'installation, suite au téléchargement de votre produit, ou sur le Web.

Onglet Champs

Avant de pouvoir construire un modèle, vous devez spécifier les champs que vous souhaitez utiliser comme cible et comme entrée. A quelques exceptions près, tous les noeuds de modélisation utilisent les informations de champ à partir d'un noeud Type en amont. Si vous utilisez un noeud Typer pour sélectionner les champs d'entrée et les champs cible, vous n'avez aucun changement à apporter dans cet onglet.

Utiliser des paramètres de noeud type. Cette option indique au noeud d'utiliser les informations du champ à partir d'un noeud Typer en amont. Il s'agit de la valeur par défaut.

Utiliser des paramètres personnalisés. Cette option indique au noeud d'utiliser les informations du champ spécifiées ici, au lieu de celles fournies par un ou des noeuds Type en amont. Après avoir sélectionné cette option, spécifiez les champs ci-dessous selon les besoins.

Cible. Pour les modèles qui nécessitent un ou plusieurs champs cibles, sélectionnez le ou les champs cibles. Il s'agit de la même action que lorsqu'on définit le rôle du champ sur *Cible* dans un noeud Type.

Entrées. Sélectionnez le ou les champs d'entrée. Il s'agit de la même action que lorsque l'on définit le rôle du champ sur *Entrée* dans un noeud Type.

Onglet Paramètres

L'onglet Paramètres contient plusieurs groupes de paramètres différents que vous pouvez modifier pour affiner le traitement des données par l'algorithme. Si vous modifiez les paramètres par défaut et que ces modifications sont incompatibles avec les autres objectifs, l'onglet Objectif est automatiquement mis à jour pour sélectionner l'option **Personnaliser l'analyse**.

Paramètres des champs

Utiliser le champ de fréquence. Cette option permet de sélectionner un champ en tant que pondération de fréquence. Utilisez cette option si les enregistrements de vos données d'apprentissage représentent

chacun plus d'une seule unité, par exemple si vous utilisez des données agrégées. Les valeurs des champs doivent être égales au nombre d'unités représentées par chaque enregistrement.

Utiliser le champ de pondération. Cette option permet de sélectionner un champ en tant que pondération d'observation. Les pondérations d'observation sont utilisées pour représenter les différences de variance dans les niveaux du champ de sortie.

Comment traiter les champs exclus de la modélisation. Spécifiez la manière de traiter les champs exclus. Vous pouvez choisir de les filtrer des données ou de simplement définir leur *Rôle* sur **Aucun**.

Remarque : Cette action va également être appliquée à la cible si elle transformée. Par exemple, si la nouvelle version dérivée de la cible est utilisée comme champ de **Cible**, la cible d'origine est filtrée ou définie sur **Aucune**.

Si les champs d'entrée ne correspondent pas à l'analyse existante. Spécifiez ce qui se passe si un ou plusieurs champs d'entrée requis sont manquants de le jeu de données entrant, lors de l'exécution d'un noeud ADP d'apprentissage.

- **Arrêter l'exécution et conserver l'analyse existante.** Cette option interrompt l'exécution, conserve les informations de l'analyse en cours et affiche une erreur.
- **Effacer l'analyse existante et analyser les nouvelles données.** Cette option efface l'analyse existante, analyse les données d'entrée et applique les transformations recommandées aux données.

Préparer les dates et les heures

De nombreux algorithmes de modélisation ne peuvent pas traiter directement les informations sur la date et l'heure. Ces paramètres vous permettent de calculer de nouvelles données de durée qui peuvent être utilisées comme entrées de modèle à partir des dates et des heures de vos données existantes. Les champs contenant les dates et les heures doivent être prédéfinis à l'aide des types de stockage de dates et d'heures. Il n'est pas recommandé de définir les champs de date et d'heure d'origine comme entrées de modèle après la préparation automatique des données.

Préparer les dates et les heures pour la modélisation En désélectionnant cette option, vous désactivez tous les autres contrôles Préparer les dates et les heures, tout en conservant les sélections.

Calculer la durée écoulée jusqu'à la date de référence Cette option génère le nombre d'années/mois/jours depuis une date de référence pour chaque variable qui contient des dates.

- **Date de référence.** Spécifier la date à partir de laquelle la durée sera calculée en fonction des informations sur la date dans les données d'entrée. Sélectionner **Date d'aujourd'hui** signifie que la date du système actuelle est toujours utilisée lorsque l'ADP est exécuté. Pour utiliser une date spécifique, sélectionnez **Date fixe** et saisissez la date désirée. La date actuelle est automatiquement saisie dans le champ **Date fixe** lors de la création initiale du noeud.
- **Unités de durée de la date.** Spécifier si l'ADP doit décider automatiquement de l'unité de la durée Date ou choisir dans les **unités fixes** des Années, Mois ou Jours.

Calculer la durée écoulée jusqu'à l'heure de référence Cette option génère le nombre d'heures/minutes/secondes depuis une heure de référence pour chaque variable qui contient des heures.

- **Heure de référence.** Spécifier l'heure à partir de laquelle la durée sera calculée en fonction des informations sur l'heure dans les données d'entrée. Sélectionner **Heure actuelle** signifie que l'heure du système actuelle est toujours utilisée lorsque l'ADP est exécuté. Pour utiliser une heure spécifique, sélectionnez **Heure fixe** et saisissez l'heure désirée. L'heure actuelle est automatiquement saisie dans le champ **Heure fixe** lors de la création initiale du noeud.
- **Unités de durée de l'heure.** Spécifier si l'ADP doit décider automatiquement de l'unité de la durée Heure ou choisir dans les **unités fixes** des Heures, Minutes ou Secondes.

Extraire les éléments de temps cycliques Utilisez ces paramètres pour scinder un champ de date ou d'heure en un ou plusieurs autres champs. Par exemple, si vous sélectionnez les trois cases de date, le champ de date d'entrée "1954-05-23" est divisé en trois champs : 1954, 5 et 23, chacun utilisant le suffixe défini dans le panneau **Noms des champ** et le champ de date d'origine est ignoré.

- **Extraire à partir des dates** Pour chaque entrée de date, spécifiez si vous souhaitez extraire des années, des mois, des jours ou une des combinaisons possibles.
- **Extraire à partir des heures.** Pour chaque entrée de date, spécifiez si vous souhaitez extraire des heures, des minutes ou des secondes ou une des combinaisons possibles.

Exclure des champs

Les données de mauvaise qualité peuvent affecter la précision de vos prédictions. Par conséquent, vous pouvez spécifier le niveau de qualité acceptable des fonctions d'entrée. Tous les champs constants ou avec 100% de valeurs manquantes sont automatiquement exclus.

Exclure les champs d'entrée de mauvaise qualité. En déselectionnant cette option, vous désactivez tous les autres contrôles Exclure les champs, tout en conservant les sélections.

Exclure les champs avec trop de valeurs manquantes Les champs ayant plus que le pourcentage spécifié de valeurs manquantes sont supprimés de l'analyse. Définissez une valeur supérieure ou égale à 0, ce qui revient à déselectionner cette option, et inférieure ou égale à 100, puisque les champs qui ne contiennent que des valeurs manquantes sont exclus automatiquement. Valeur par défaut : 50

Exclure les champs nominaux avec trop de catégories uniques Les champs nominaux ayant plus que le nombre spécifié de catégories sont supprimés de l'analyse. Spécifiez un nombre entier positif. Par défaut, la valeur est 100. Cette option est utile pour supprimer automatiquement de la modélisation les champs contenant des informations d'enregistrement unique, tels que l'ID, l'adresse ou le nom.

Exclure les champs catégoriels avec trop de valeurs dans une même catégorie Les champs ordinaux et nominaux avec une catégorie contenant plus que le pourcentage spécifié d'enregistrements sont supprimés de l'analyse. Définissez une valeur supérieure ou égale à 0, ce qui revient à déselectionner cette option, et inférieure ou égale à 100, puisque les champs constants sont exclus automatiquement. La valeur par défaut est 95.

Préparation des entrées et des cibles

Aucune donnée n'étant jamais dans un parfait état avant le traitement, vous voudrez sans doute ajuster certains paramètres avant d'exécuter une analyse. Par exemple, vous voudrez supprimer les valeurs extrêmes, spécifier la manière de traiter les valeurs manquantes ou encore ajuster le type.

Remarque : Si vous modifiez les valeurs de ce panneau, l'onglet **Objectifs** est automatiquement mis à jour pour sélectionner l'option **Analyse personnalisée**.

Préparer les champs d'entrée et cible pour la modélisation. Active ou désactive tous les champs du volet.

Ajuster le type et améliorer la qualité des données. Pour les entrées et la cible, il est possible de spécifier plusieurs transformations de données de manière séparée, si vous ne souhaitez pas modifier les valeurs de la cible. Par exemple, une prévision de revenu en dollars est plus significative qu'une prévision mesurée en log(dollars). En outre, si la cible contient des valeurs manquantes, il n'existe pas de gain prévu pour remplacer les valeurs manquantes, alors que le remplacement des valeurs manquantes en entrée peut permettre à certains algorithmes de traiter des informations qui auraient été perdues.

Des paramètres supplémentaires pour ces transformations, par exemple la valeur de césure des valeurs extrêmes, sont communs aux entrées et à la cible.

Vous pouvez sélectionner les paramètres suivants pour les entrées ou la cible ou pour les deux à la fois :

- **Ajuster le type des champs numériques (ordinal et continu).** Sélectionnez cette option pour déterminer si les champs numériques avec un niveau de mesure *Ordinal* peuvent être convertis en champs *Continus*, et vice versa. Vous pouvez spécifier les valeurs minimale et maximale du seuil pour contrôler la conversion.
- **Réorganiser les champs nominaux.** Sélectionnez cette option pour trier les champs nominaux (ensemble) de la plus petite catégorie à la plus grande.
- **Remplacer les valeurs éloignées dans les champs continus..** Spécifiez si vous souhaitez remplacer les valeurs extrêmes. Utilisez cette option en conjonction avec les options **Méthode de remplacement des valeurs extrêmes** ci-dessous.
- **Champs continus : remplacer les valeurs manquantes par la moyenne.** Sélectionnez cette option pour remplacer les valeurs manquantes des fonctions continues (plage).
- **Champs nominaux : remplacer les valeurs manquantes par le mode** Sélectionnez cette option pour remplacer les valeurs manquantes des fonctions nominales (ensemble).
- **Champs ordinaux : remplacer les valeurs manquantes par la médiane.** Sélectionnez cette option pour remplacer les valeurs manquantes des fonctions ordinales (ensemble ordonné).

Nombre maximal de valeurs pour les champs ordinaux Spécifiez le seuil pour la redéfinition des champs ordinaux (ensemble ordonné) en champs continus (plage). La valeur par défaut est 10. Si un champ ordinal contient plus de 10 catégories, il est redéfini en champ continu (plage).

Nombre minimal de valeurs pour les champs continus. Spécifiez le seuil pour la redéfinition des champs continus ou d'échelle (plage) en champs ordinaux (ensemble ordonné). La valeur par défaut est 5. Si un champ continu contient plus de 5 valeurs, il est redéfini en champ ordinal (ensemble ordonné).

Valeur de césure de valeur éloignée. Spécifiez la limite des valeurs extrêmes, mesurée dans les écarts-types. La valeur par défaut est 3.

Méthode de remplacement des valeurs éloignées. Choisissez si les valeurs extrêmes doivent être remplacées (en les tronquant de force) par la valeur de césure ou supprimées et définies comme valeurs manquantes. Les valeurs extrêmes définies comme valeurs manquantes suivent les paramètres de traitement des valeurs manquantes sélectionnés ci-dessus.

Attribuer la même échelle à tous les champs d'entrée continus. Pour normaliser les champs d'entrée continus, cochez cette case et choisissez la méthode de normalisation. La méthode par défaut est **transformation en score z**, pour laquelle vous pouvez spécifier la **moyenne finale**, dont la valeur par défaut est 0, et l'**écart -type final**, dont la valeur par défaut est 1. Sinon, vous pouvez choisir d'utiliser l'option **Transformation min/max** et spécifier les valeurs minimum et maximum, dont les valeurs par défaut respectives sont 0 et 100.

Ce champ est particulièrement utile lorsque vous sélectionnez l'option **Exécuter la construction des caractéristiques** dans le volet Construire et Sélectionner les caractéristiques.

Redimensionner une cible continue avec la transformation de Box-Cox Pour normaliser un champ cible continu (d'échelle ou de plage), cochez cette case. La transformation de Box-Cox possède une valeur par défaut de 0 pour la **moyenne finale** et de 1 pour l'**écart-type final**.

Remarque : si vous choisissez de normaliser la cible, sa dimension sera transformée. Dans ce cas, vous pourriez avoir besoin de générer un noeud Calculer pour appliquer une transformation inverse et redonner un format reconnaissable aux unités transformées pour un traitement ultérieur. Pour plus d'informations, voir «Génération d'un noeud Calculer», à la page 143.

Construction et sélection des fonctions

Pour améliorer le pouvoir prédictif de vos données, vous pouvez transformer les champs d'entrées ou en construire de nouveaux basés sur les champs existants.

Remarque : Si vous modifiez les valeurs de ce panneau, l'onglet **Objectifs** est automatiquement mis à jour pour sélectionner l'option **Analyse personnalisée**.

Transformation, construction et sélection de champs d'entrée pour améliorer la puissance de prédiction. Active ou désactive tous les champs du volet.

Fusionner les catégories éparpillées pour optimiser l'association avec une cible. Sélectionner cette option pour créer un modèle plus petit en réduisant le nombre de variables à traiter en association avec la cible. Si nécessaire, modifiez la valeur de probabilité dont la valeur par défaut est de 0,05.

Remarque : si toutes les catégories sont fusionnées en une seule, les versions d'origine et dérivées du champ sont exclues car elles n'ont pas de valeur de prédicteur.

Lorsqu'il n'existe aucune cible, fusionner les modalités éparpillées en fonction de leur nombre. Si le jeu de données n'a pas de cible, vous pouvez choisir de fusionner les catégories éparpillées des fonctions ordinales (ensemble ordonné) ou nominales (ensemble) ou des deux à la fois. Spécifiez le pourcentage minimum d'observations, ou d'enregistrements dans les données, qui identifie les catégories à fusionner. La valeur par défaut est 10.

Les catégories sont fusionnées en utilisant les règles suivantes :

- La fusion n'est pas réalisée sur les champs binaires.
- Lorsqu'il n'y a que deux catégories à fusionner, la fusion est interrompue.
- La fusion est interrompue s'il n'existe pas de catégorie d'origine, ni de catégorie créée durant la fusion, avec un pourcentage d'observations inférieur au pourcentage minimum spécifié.

Regrouper les champs continus tout en conservant le pouvoir prédictif.. Si le jeu de données comprend une cible qualitative, vous pouvez regrouper les entrées continues ayant de fortes associations pour améliorer les performances du traitement. Si nécessaire, modifiez la valeur de probabilité des sous-ensembles homogènes dont la valeur par défaut est de 0,05.

Si l'opération de regroupement génère un regroupement unique pour un champ spécifique, les versions d'origine et regroupées du champ sont exclues car elles n'ont pas de valeur de prédicteur.

Remarque : Le regroupement dans l'ADP est différent du regroupement optimal utilisé dans les autres parties d'IBM SPSS Modeler. Le regroupement optimal utilise des informations d'entropie pour convertir une variable continue en une variable qualitative ; il doit trier les données et les stocker dans la mémoire. L'ADP utilise des sous-ensembles homogènes pour regrouper une variable continue. Cela signifie que le regroupement ADP n'a pas besoin de trier les données et ne stocke pas toutes les données dans une mémoire. L'utilisation de la méthode des sous-ensembles homogènes pour regrouper une variable continue signifie que le nombre de catégories après le regroupement est toujours inférieur ou égal au nombre de catégories de la cible.

Effectuer la sélection de fonctions. Sélectionnez cette option pour supprimer les fonctions dont le coefficient de corrélation est faible. Si nécessaire, modifiez la valeur de probabilité dont la valeur par défaut est de 0,05.

Cette option s'applique uniquement aux fonctions d'entrée continues où la cible est continue et aux fonctions d'entrée qualitatives.

Effectuer la construction des fonctions Sélectionner cette option pour dériver de nouvelles fonctions d'une combinaison de plusieurs fonctions existantes (qui sont ensuite supprimées de la modélisation).

Cette option s'applique uniquement aux fonctions d'entrée continues où la cible est continue ou lorsqu'il n'y a pas de cible.

Noms de champ

Pour identifier facilement les fonctions nouvelles et transformées, l'ADP crée et applique de nouveaux noms, préfixes ou suffixes de base. Vous pouvez modifier ces noms pour qu'ils soient plus adaptés à vos propres besoins et données. Si vous souhaitez spécifier d'autres libellés, vous devez le faire dans le noeud Type en aval.

Champs transformés et construits. Spécifiez les extensions de nom à appliquer aux champs cibles et d'entrées transformés.

Veuillez noter que le noeud ADP définissant les champs de chaîne pour qu'ils soient vides, peut provoquer une erreur en fonction du traitement accordé aux champs non utilisés. Si **Comment traiter les champs exclus de la modélisation** est défini sur **Eliminer les champs non utilisés** dans le panneau Paramètres des champs de l'onglet Paramètres, les extensions de nom des entrées et de la cible peuvent être définies sur rien. Les champs d'origine sont éliminés et les champs transformés sont enregistrés à leur place. Dans ce cas, les nouveaux champs transformés auront le même nom que vos champs d'origine.

Cependant, si vous choisissez de paramétrer **Définir la direction des champs non utilisés sur Aucune**, les extensions de nom de la cible et des entrées nulles ou vides provoqueront une erreur car vous essaierez de créer des noms de champ en doublon.

En outre, spécifiez le nom du préfixe à appliquer aux fonctions construites à l'aide des paramètres Sélectionner et Construire. Le nouveau nom est créé en ajoutant un suffixe numérique à ce nom de racine du préfixe. Le format du nombre dépend du nombre de nouvelles fonctions dérivées, par exemple :

- si 1 à 9 caractéristiques sont construites, elles seront nommées : caractéristique1 à caractéristique9.
- si 10 à 99 caractéristiques sont construites, elles seront nommées : caractéristique10 à caractéristique99.
- si 100 à 999 caractéristiques sont construites, elles seront nommées : caractéristique001 à caractéristique999.

Cela permet que les fonctions construites soient triées dans un ordre cohérent quel que soit leur nombre.

Durée calculée à partir des dates et des heures. Spécifier les extensions de nom à appliquer aux durées calculées à partir des dates et des heures.

Eléments cycliques extraits de dates et des heures. Spécifier les extensions de nom à appliquer aux éléments cycliques extraits des dates et des heures.

Onglet Analyse

1. Lorsque les paramètres d'ADP vous conviennent, y compris les modifications effectuées dans les onglets Objectif, Champs et Paramètres, cliquez sur **Analyser les données**. L'algorithme applique les paramètres aux entrées de données et affiche les résultats dans l'onglet Analyse.

L'onglet Analyse contient à la fois des sorties en tableaux et des sorties graphiques qui résument le traitement de vos données et affichent les recommandations sur la façon de modifier ou d'améliorer les données pour l'évaluation. Vous pouvez ensuite revoir puis accepter ou refuser ces recommandations.

L'onglet Analyse est composé de deux panneaux, la vue principale à gauche et la vue liée, ou auxiliaire, à droite. Il existe trois vues principales :

- Récapitulatif de traitement des champs (par défaut). Pour plus d'informations, voir «Récapitulatif de traitement des champs», à la page 137.
- Champs. Pour plus d'informations, voir «Champs», à la page 137.
- Récapitulatif des actions. Pour plus d'informations, voir «Récapitulatif des actions», à la page 138.

Il existe quatre vues liées/auxiliaires :

- Pouvoir prédictif (par défaut). Pour plus d'informations, voir «Pouvoir prédictif», à la page 139.
- Tableau des champs. Pour plus d'informations, voir «Tableau des champs», à la page 139.
- Détails des champs. Pour plus d'informations, voir «Détails des champs», à la page 139.
- Détails des actions. Pour plus d'informations, voir «Détails des actions», à la page 140.

Liens entre les vues

Dans la vue principale, le texte souligné dans les tableaux contrôle ce qui apparaît dans la vue liée. Si vous cliquez sur ces parties de texte, vous obtenez des détails sur un champ, un ensemble de champs ou une étape de traitement spécifique. Le lien que vous avez sélectionné en dernier apparaît en une couleur plus foncée qui permet d'identifier la connexion entre les contenus des deux panneaux de la vue.

Réinitialisation des vues

Pour afficher de nouveau les recommandations d'analyse d'origine et abandonner les modifications effectuées sur les vues Analyse, cliquez sur **Réinitialiser** au bas du panneau de la vue principale.

Récapitulatif de traitement des champs

La table récapitulative de traitement des champs fournit un instantané de l'impact du traitement général projeté, y compris les modifications de l'état des fonctions et le nombre de fonctions construites.

Veuillez noter que le modèle est bien construit, et que par conséquent il n'y a pas de mesure ou de graphique de la modification du pouvoir prédictif général avant et après la préparation des données. Par contre, vous pouvez afficher les graphiques du pouvoir prédictif des prédicteurs individuels recommandés.

Le tableau affiche les informations suivantes :

- le nombre de champs cibles.
- Le nombre de prédicteurs (d'entrée) d'origine.
- Les prédicteurs recommandés pour l'analyse et la modélisation. Cela comprend le nombre total de champs recommandés ; le nombre de champs d'origine non transformés recommandés ; le nombre de champs transformés recommandés (sans les versions intermédiaires des champs, champs dérivés des prédicteurs de date/heure et prédicteurs construits) ; le nombre de champs dérivés recommandés des champs date/heure ; et le nombre de prédicteurs construits.
- Le nombre de prédicteurs d'entrée non recommandés quelle que soit leur forme, que ce soit sous leur forme d'origine, comme champ dérivé, ou comme entrée d'un prédicteur construit.

Lorsque des informations sur les **champs** sont soulignées, cliquez pour afficher plus de détails dans une vue liée. Les détails de la **Cible**, des **Fonctions d'entrée**, et des **Fonctions d'entrée non utilisées** apparaissent dans la vue liée Tableau des champs. Pour plus d'informations, voir «Tableau des champs», à la page 139. **Les Caractéristiques recommandées pour l'analyse** apparaissent dans la vue liée Puissance de prédiction. Pour plus d'informations, voir «Pouvoir prédictif», à la page 139.

Champs

La vue principale Champs affiche les champs traités et si l'ADP recommande de les utiliser dans les modèles en aval. Vous pouvez ignorer les recommandations pour n'importe quel champ ; par exemple, exclure les fonctions construites ou inclure les fonctions que l'ADP recommande d'exclure. Si un champ a été transformé, vous pouvez décider d'accepter ou non la transformation suggérée ou d'utiliser ou non la version d'origine.

La vue Champs est composée de deux tableaux, un pour la cible et un pour les prédicteurs qui ont été traités ou créés.

Tableau Cible

Le tableau **Cible** n'apparaît que si une cible est définie dans les données.

Ce tableau contient deux colonnes :

- **Nom.** C'est le nom ou le libellé du champ cible ; le nom d'origine est toujours utilisé, même si le champ a été transformé.
- **Niveau de mesure.** Ceci affiche l'icône représentant le niveau de mesure. Placez la souris sur l'icône pour afficher un libellé (continu, ordinal, nominal, etc.) qui décrit les données.

  Si la cible a été transformée, la colonne **Niveau de mesure** reflète la version transformée finale. *Remarque* : vous ne pouvez pas désactiver les transformations pour la cible.

Tableau des prédicteurs

Le tableau **Prédicteurs** est affiché en permanence. Chaque ligne du tableau représente un champ. Les lignes sont triées par défaut dans l'ordre décroissant du pouvoir prédictif.

Pour les fonctions ordinaires, le nom d'origine est toujours utilisé comme nom de ligne. Les versions d'origine et dérivée des champs date/heure apparaissent dans le tableau (dans des lignes séparées) ; le tableau contient également les prédicteurs construits.

Veuillez noter que les versions transformées des champs apparaissant dans le tableau représentent toujours les versions finales.

Par défaut, seuls les champs recommandés sont affichés dans le tableau des prédicteurs. Pour afficher les champs restants, sélectionnez la boîte de dialogue **Inclure les champs non recommandés dans le tableau** au-dessus du tableau ; ces champs sont ensuite affichés au bas du tableau.

Le tableau contient les colonnes suivantes :

- **Version à utiliser.** Affiche une liste déroulante qui contrôle l'utilisation d'un champ en aval et s'il faut utiliser les transformations recommandées. Par défaut, la liste déroulante reflète les recommandations.

  Pour les prédicteurs ordinaires qui ont été transformés, la liste déroulante contient trois choix : **Transformée**, **Originale** et **Originale**.

  Pour les prédicteurs non transformés ordinaires, les choix sont : **Originale** et **Ne pas utiliser**.

  Pour les champs dérivés date/heure et les prédicteurs construits, les choix sont : **Transformée** et **Ne pas utiliser**.

  Pour les champs de date d'origine, la liste déroulante est désactivée et définie sur **Ne pas utiliser**.

  Remarque : Pour les prédicteurs contenant à la fois les versions d'origine et transformés, passer des versions **d'origine** aux versions **transformées** met automatiquement à jour les paramètres **Niveau de mesure** et **Puissance de prédiction** pour ces caractéristiques.
- **Nom.** Chaque nom de champ est un lien. Cliquez sur un nom pour afficher plus d'informations sur le champ dans la vue liée. Pour plus d'informations, voir «Détails des champs», à la page 139.
- **Niveau de mesure.** Affiche l'icône représentant le type de données ; passez la souris sur l'icône pour afficher un libellé (continu, ordinal, nominal, etc.) qui décrit les données.
- **Puissance de prédiction.** Le pouvoir prédictif est affiché uniquement pour les champs recommandés par l'ADP. Cette colonne n'apparaît pas si aucune cible n'est définie. Le pouvoir prédictif est compris entre 0 et 1, les valeurs les plus élevées, indiquant des prédicteurs de "meilleur" qualité. En général, le pouvoir prédictif est utile pour comparer les prédicteurs dans une analyse ADP, mais les valeurs du pouvoir prédictif ne peuvent être comparées entre des analyses différentes.

Récapitulatif des actions

Pour chaque action effectuée par la préparation automatique des données, les prédicteurs d'entrée sont transformées et/ou supprimées ; les champs qui survivent à une action sont utilisés à la suivante. Les champs qui survivent jusqu'à la dernière étape sont ensuite recommandés pour la modélisation, alors que les entrées des prédicteurs transformés et construits sont supprimés.

Le Récapitulatif des actions est un simple tableau qui répertorie les actions effectuées par l'ADP. Lorsqu'une **Action** est soulignée, vous pouvez cliquer dessus pour afficher plus de détails sur les actions effectuées dans une vue liée. Pour plus d'informations, voir «Détails des actions», à la page 140.

Remarque : Seules les versions d'origine et transformées finales de chaque champ sont affichées, et pas les versions intermédiaires utilisées pendant l'analyse.

Pouvoir prédictif

Affichée par défaut au début de l'analyse ou lorsque vous sélectionnez **Prédicteurs recommandés pour l'analyse** dans la vue principale Récapitulatif du traitement des champs, le graphique affiche le pouvoir prédictif des prédicteurs recommandés. Les champs sont triés par pouvoir prédictif, avec le champ ayant la plus haute valeur apparaissant en premier.

Pour les versions transformées des prédicteurs ordinaires, le nom des champs reflète votre choix de suffixe dans le panneau Noms de champ de l'onglet Paramètres ; par exemple : *_transformed*.

Les icônes de niveau de mesure sont affichées après les noms de champ individuels.

La puissance de prédiction de chaque prédicteur recommandé est calculée à partir d'une régression linéaire ou d'un modèle de Nave Bayes selon que la cible est continue ou qualitative.

Tableau des champs

La vue Tableau des champs est un simple tableau qui répertorie les fonctions importantes et qui apparaît lorsque vous cliquez sur **Cible**, **Prédicteurs**, ou **Prédicteurs non utilisés** dans la vue principale Récapitulatif du traitement des champs.

Ce tableau contient deux colonnes :

- **Nom.** Nom de la valeur prédite.

  Pour les cibles, le libellé ou le nom d'origine du champ est utilisé, même si la cible a été transformée.

  Pour les versions transformées des prédicteurs ordinaires, le nom reflète votre choix de suffixe dans le panneau Noms de champ de l'onglet Paramètres ; par exemple : *_transformed*.

  Pour les champs dérivés des dates et des heures, le nom de la version transformée finale est utilisé ; par exemple : *bdate_years*.

  Pour les prédicteurs construits, le nom du prédicteur construit est utilisé ; par exemple : *Predictor1*.
- **Niveau de mesure.** Affiche l'icône représentant le type de données.

  Pour la cible, le **Niveau de mesure** reflète toujours la version transformée (si la cible a été transformée), par exemple, changée d'ordinale (ensemble ordonné) à continue (plage, échelle) et vice versa.

Détails des champs

La vue Détails des champs contient les graphiques de distribution, des valeurs manquantes et du pouvoir prédictif (le cas échéant) pour le champ sélectionné et s'affiche lorsque vous cliquez sur un **Nom** de la vue principale Champs. De plus, l'historique du traitement pour le champ et le nom du champ transformé apparaissent également (le cas échéant).

Pour chaque ensemble de graphiques, deux versions apparaissent côte à côte pour comparer le champ avec et sans transformations appliquées ; si aucune version transformée du champ n'existe, un graphique apparaît pour la version d'origine uniquement. Pour les champs de date ou d'heure dérivés et les prédicteurs construits, les graphiques n'apparaissent que pour le nouveau prédicteur.

Remarque : Si un champ est exclu parce qu'il contient trop de modalités, seul l'historique de traitement apparaît.

Graphique de distribution

La distribution des champs continus apparaît dans un histogramme, avec une courbe normale superposée et une ligne de référence verticale pour la valeur moyenne ; les champs catégoriels apparaissent sous forme de graphique à barres.

Les histogrammes sont libellés pour montrer l'écart-type et l'asymétrie, toutefois l'asymétrie n'apparaît pas si le nombre des valeurs est inférieur ou égal à 2 ou si la variance du champ d'origine est inférieure à 10-20.

Passez la souris sur le graphique pour afficher la moyenne des histogrammes ou le nombre et le pourcentage du nombre total d'enregistrements des catégories dans les graphiques à barres.

Graphique des valeurs manquantes

Les graphiques circulaires comparent le pourcentage des valeurs manquantes avec et sans transformations appliquées ; les libellés de graphique indiquent le pourcentage.

Si l'ADP traite les valeurs manquantes, le graphique circulaire après la transformation comprend la valeur de remplacement comme libellé, c'est-à-dire la valeur utilisée à la place des valeurs manquantes.

Passez la souris sur le graphique pour afficher le nombre des valeurs manquantes et le pourcentage du nombre total d'enregistrements.

Graphique de pouvoir prédictif

Pour les champs recommandés, les graphiques à barres affichent le pouvoir prédictif avant et après la transformation. Si la cible a été transformée, le pouvoir prédictif calculé tient compte de la cible transformée.

Remarque : Les graphiques de puissance de prédiction ne sont pas affichés si aucune cible n'est définie, ou si la cible est atteinte depuis le panneau de la vue principale.

Passez la souris sur le graphique pour afficher la valeur du pouvoir prédictif.

Tableau des historiques du traitement

Ce tableau indique la façon dont la version transformée d'un champ a été dérivée. Les actions entreprises par l'ADP sont répertoriées dans l'ordre dans lequel elles ont été exécutées ; mais, pour certaines étapes, plusieurs actions ont pu être exécutées pour un champ particulier.

Remarque : ce tableau n'apparaît pas pour les champs qui n'ont pas été transformés.Ce tableau n'apparaît pas pour les champs qui n'ont pas été transformés.

Les informations du tableau sont divisées en deux ou trois colonnes :

- **Action.** Le nom de l'action. Par exemple, Prédicteurs continus. Pour plus d'informations, voir «Détails des actions».
- **Détails.** La liste des traitements effectués. Par exemple, Transformer en unités standard.
- **Fonction.** Apparaît uniquement pour les prédicteurs construits et affiche la combinaison linéaire de champs d'entrée, par exemple, 0,06*âge + 1,21*hauteur.

Détails des actions

La vue liée Détails des actions apparaît lorsque vous cliquez sur **Action** dans la vue principale Récapitulatif des actions.La vue liée Détails des actions affiche des informations relatives aux actions et des informations communes pour chaque étape de traitement effectuée. Les détails relatifs à chaque action spécifique apparaissent d'abord.

La description de chaque action est utilisée comme titre en haut de la vue liée. Les détails relatifs à chaque action sont affichés sous le titre, et peuvent contenir des détails sur le nombre de prédicteurs dérivés, de champs reconvertis, de transformations de cible, de catégories fusionnées ou réorganisées et de prédicteurs construits ou exclus.

Au cours du traitement des actions, le nombre de prédicteurs utilisés pour le traitement peut varier, par exemple lorsque des prédicteurs sont exclus ou fusionnés.

Remarque : Si une action est désactivée ou qu'aucune cible n'est spécifiée, un message d'erreur apparaît à la place des détails de l'action lorsque vous cliquez dessus dans la vue principale Récapitulatif des actions.

Il existe neuf actions possibles, toutefois, toutes ne sont pas nécessairement actives pour chaque analyse.

tableau Champs de texte

Ce tableau affiche le nombre :
- Blancs à droite tronqués.
- Prédicteurs exclus de l'analyse.

Tableau Prédicteurs de date et d'heure

Ce tableau affiche le nombre :
- Durées dérivées des prédicteurs de date et d'heure.
- d'éléments Date et heure.
- Prédicteurs de date et d'heure dérivées, au total.

La date ou heure de référence est affichée comme note de bas de page si des durées de date ont été calculées.

Tableau Balayage des prédicteurs

Ce tableau affiche le nombre des prédicteurs suivants exclus du traitement :
- constantes.
- Prédicteurs avec trop de valeurs manquantes.
- Prédicteurs avec trop d'observations dans une seule catégorie.
- Champs nominaux (ensembles) avec trop de catégories.
- Prédicteurs supprimés, au total.

Tableau Vérifier le niveau de mesure

Ce tableau affiche le nombre de champs reconvertis, répartis selon les catégories suivantes :
- Champs ordinaux (ensembles ordonnés) reconvertis en champs continus.
- Champs continus reconvertis en champs ordinaux.
- Nombre total des champs reconvertis.

Si aucun champ d'entrée (cible ou de prédicteurs) n'est un ensemble continu ou ordinal, cela apparaît en note de bas de page.

Tableau Valeurs extrêmes

Ce tableau affiche le nombre de valeurs extrêmes traitées.

- soit le nombre de champs continus pour lesquels des valeurs extrêmes ont été recherchées et tronquées, ou le nombre de champs continus pour lesquels les valeurs extrêmes ont été recherchées et définies sur manquantes, en fonction de vos paramètres dans le panneau Préparer les entrées & la cible dans l'onglet Paramètres.
- le nombre de champs continus exclus parce qu'ils étaient constants après le traitement des valeurs extrêmes.

Une note de bas de page indique la valeur de césure des valeurs extrêmes et une autre note de bas de page apparaît si aucun champ d'entrée (cible ou de prédicteurs) n'est continu.

Tableau Valeurs manquantes

Ce tableau affiche le nombre de champs qui contenaient des valeurs manquantes remplacées, selon les catégories suivantes :

- Cible. Cette ligne n'apparaît pas si aucune cible n'est spécifiée.
- Valeurs prédites. Elles sont divisées en nombre de champs nominaux (ensemble), ordinaux (ensemble ordonné) et continus.
- Le nombre total de valeurs manquantes remplacées.

Tableau Cible

Ce tableau indique si la cible a été transformée :

- transformation de Box-Cox en normalité. Cette catégorie est elle-même divisée en colonnes qui indiquent le critère spécifié (moyenne et écart-type) et le Lambda.
- catégories cibles réorganisées pour améliorer la stabilité.

Tableau prédicteurs catégoriels

Ce tableau affiche le nombre de prédicteurs catégoriels:

- dont les catégories ont été réorganisées de la plus faible la plus élevée pour améliorer la stabilité.
- dont les catégories ont été fusionnées pour optimiser l'association avec la cible.
- dont les catégories ont été fusionnées pour traiter les catégories éparpillées.
- exclues en raison d'une faible association avec la cible.
- exclues parce qu'elles étaient constantes après la fusion.

Une note de bas de page apparaît si aucun prédicteur catégoriel n'existe.

Tableau Prédicteurs continus

Il existe deux tableaux. Le premier affiche une des transformations suivantes :

- les valeurs des prédicteurs transformées en unités standard. De plus, il indique le nombre de prédicteurs transformés, la moyenne spécifiée et l'écart-type.
- Les valeurs des prédicteurs mappées sur une plage commune. De plus, il indique le nombre de prédicteurs transformés utilisant une transformation min-max, ainsi que les valeurs minimum et maximum spécifiées.
- les valeurs des prédicteurs et le nombre de prédicteurs regroupés.

Le deuxième tableau affiche les détails de construction de l'espace des prédicteurs, sous la forme du nombre de prédicteurs :

- construites.
- exclues en raison d'une faible association avec la cible.
- exclues parce qu'elles étaient constantes après le regroupement.

- exclues parce qu'elles étaient constantes après la construction.

Une note de bas de page apparaît si aucun prédicteur continu n'a été saisi.

Génération d'un noeud Calculer

Lorsque vous générez un noeud Calculer, il applique la transformation inversée de la cible au champ de scores. Par défaut, le noeud entre le nom du champ de scores qui serait produit par un logiciel de modélisation automatique (comme Auto Classifier ou Auto Numeric) ou par le noeud Ensemble. Si une cible d'échelle (plage) a été transformée, le champ de scores apparaît en unités transformées ; par exemple, log($) à la place de $. Afin d'interpréter et d'utiliser les résultats, vous devez reconvertir la valeur observée dans son échelle d'origine.

Remarque : Vous ne pouvez générer un noeud Calculer que lorsque le noeud ADP contient une analyse qui rééchelonne une cible plage (c'est-à-dire que le rééchelonnement de Box-Cox est sélectionné dans le panneau Préparer les entrées & la cible). Vous ne pouvez pas générer de noeud Calculer si la cible n'est pas une plage, ou si le rééchelonnement de Box-Cox n'est pas sélectionné.

Le noeud Calculer est créé en mode Multiple et utilise @FIELD dans l'expression pour que vous puissiez ajouter la cible transformée si nécessaire. Par exemple, en utilisant les informations suivantes :

- Nom de champ cible : response
- Nom de champ cible transformé : response_transformed
- Nom du champ de scores : $XR-response_transformed

Le noeud Calculer créerait un nouveau champ : $XR-response_transformed_inverse.

Remarque : Si vous n'utilisez pas de logiciel de modélisation automatique ou de noeud Ensemble, vous devrez modifier le noeud Calculer pour transformer le bon champ de scores pour votre modèle.

Cibles continues normalisées

Par défaut, si vous sélectionnez la case **Rééchelonner une cible continue avec la transformation de Box-Cox** dans le panneau Préparer les entrées & la cible, cela transforme la cible et vous créez un nouveau champ qui sera la cible pour la génération de votre modèle. Par exemple, si votre cible d'origine était *response*, la nouvelle cible sera *response_transformed*; les modèles en aval du noeud ADP choisiront automatiquement cette nouvelle cible.

Mais, cela peut provoquer des problèmes, en fonction de la cible d'origine. Par exemple, si la cible était *Age*, les valeurs de la nouvelle cible ne seront pas *Années*, mais une version transformée de *Années*. Cela signifie que vous ne pouvez pas consulter les scores et les interpréter car ils ne sont pas présentés en unités reconnaissables. Dans ce cas, vous pouvez appliquer une transformation inverse qui reconvertira vos unités transformées en ce qu'elles devaient être. Pour ce faire :

1. Après avoir cliqué sur **Analyser les données** pour effectuer l'analyse ADP, sélectionnez le *noeud Calculer* dans le menu *Générer*.
2. Placez le noeud Calculer après votre nugget sur le canevas des modèles.

Le noeud Calculer restaurera le champ de scores aux dimensions d'origine afin que la prédiction soit effectuée en des valeurs *Années* d'origine.

Par défaut, le noeud Calculer transforme le champ de scores généré par un logiciel de modélisation automatique ou un modèle combiné. Si vous construisez un modèle individuel, vous devez modifier le noeud Calculer pour dériver à partir de votre champ de scores actuel. Si vous souhaitez évaluer votre modèle, vous devez ajouter la cible transformée au champ **Calculer à partir de** dans le noeud Calculer.

Cela applique la même transformation inverse à la cible et les noeuds en aval Evaluation ou Analyse utiliseront les données transformées correctement tant que vous modifiez ces noeuds pour qu'ils utilisent des noms de champs à la place des métadonnées.

Si vous voulez également restaurer le nom d'origine, vous pouvez utiliser un noeud Filtre pour supprimer le champ cible d'origine s'il existe encore et renommer la cible et les champs de scores.

Noeud Typer

Les propriétés de champ peuvent être indiquées dans un noeud source ou dans un noeud Typer distinct. Les fonctionnalités sont similaires dans les deux noeuds. Les propriétés suivantes sont disponibles :

- **Champ** Cliquez deux fois sur un nom de champ pour spécifier des libellés de valeur et de champ pour les données dans IBM SPSS Modeler. Par exemple, vous pouvez consulter ou modifier ici les métadonnées de champ importées à partir de IBM SPSS Statistics. De même, vous pouvez créer des libellés pour les champs et leurs valeurs. La présence des libellés indiqués ici dans IBM SPSS Modeler dépend des sélections effectuées dans la boîte de dialogue Propriétés du flux.
- **Mesure** Il s'agit du niveau de mesure utilisé pour décrire les caractéristiques des données dans un champ précis. Si tous les détails d'un champ sont connus, il est dit **complètement instancié**. Pour plus d'informations, voir «Niveaux de mesure», à la page 145.

  Remarque : Le niveau de mesure d'un champ est différent de son type de stockage, qui indique si les données sont stockées sous forme de chaînes, d'entiers, de nombres réels, de dates, d'heures, d'horodatages ou de listes.
- **Valeurs** Cette colonne vous permet de spécifier des options pour la lecture des valeurs de données depuis le jeu de données ou d'utiliser l'option **Spécifier** afin de spécifier des niveaux de mesure et des valeurs dans une boîte de dialogue distincte. Vous pouvez également choisir de transférer les champs sans lire leurs valeurs. Pour plus d'informations, voir «Valeurs de données», à la page 150.

  Remarque : Vous ne pouvez pas modifier la cellule dans cette colonne si l'entrée **Champ** correspondante contient une liste.
- **Manquantes** Utilisé pour spécifier la façon dont les valeurs manquantes pour le champ sont traitées. Pour plus d'informations, voir «Définition de valeurs manquantes», à la page 155.

  Remarque : Vous ne pouvez pas modifier la cellule dans cette colonne si l'entrée **Champ** correspondante contient une liste.
- **Vérifier** Dans cette colonne, vous pouvez définir des options pour vous assurer que les valeurs de champ sont conformes aux valeurs ou plages spécifiées. Pour plus d'informations, voir «Vérification des valeurs de type», à la page 155.

  Remarque : Vous ne pouvez pas modifier la cellule dans cette colonne si l'entrée **Champ** correspondante contient une liste.
- **Rôle** Utilisé pour indiquer aux noeuds de modélisation si les champs sont des champs d'**entrée** (champs prédicteurs) ou **cible** (champs prédits) pour un processus d'apprentissage automatique). Sont également disponibles les rôles **Les deux** et **Aucun**, et l'option **Partition**. Cette dernière signale les champs utilisés pour partitionner les enregistrements en échantillons distincts à des fins d'apprentissage, de test et de validation. La valeur **Diviser** spécifie que des modèles séparés seront construits pour chaque valeur possible du champ. Pour plus d'informations, voir «Définition du rôle de champ», à la page 155.

Plusieurs autres options peuvent également être spécifiées dans la fenêtre du noeud Typer :

- Les options du menu Outils permettent d'**ignorer les champs uniques** une fois le noeud typeinstancié (via vos spécifications, la lecture des valeurs ou l'exécution du flux). Si vous choisissez d'ignorer les champs uniques, les champs comportant une seule valeur sont automatiquement ignorés.

- Les options du menu Outils permettent d'**ignorer les grands ensembles** une fois le noeud Typer instancié. Si vous choisissez d'ignorer les grands ensembles, les ensembles dont le nombre de membres est élevé sont automatiquement ignorés.
- Les options du menu Outils vous permettent de choisir de **Convertir les entiers continus en ordinaux** une fois le noeud Typer instancié. Pour plus d'informations, voir «Conversion de données continues», à la page 149.
- Les options du menu Outils permettent de créer un noeud Filtrer pour exclure les champs sélectionnés.
- A l'aide des boutons bascule représentant des lunettes de soleil, vous pouvez définir le paramètre par défaut Lire ou Transférer pour tous les champs. L'onglet Types du noeud source transmet les champs par défaut, alors que le noeud Typer lit les valeurs par défaut.
- A l'aide du bouton **Effacer les valeurs**, vous pouvez supprimer les changements apportés aux valeurs de champ de ce noeud (valeurs non héritées) et relire les valeurs des opérations effectuées en amont. Cette option est utile pour réinitialiser les changements que vous avez apportés à certains champs en amont.
- A l'aide du bouton **Effacer toutes les valeurs**, vous pouvez réinitialiser les valeurs de **tous** les champs lus dans le noeud. Cette option paramètre la colonne **Valeurs** de tous les champs sur **`Lire`**. Cette option est utile pour réinitialiser les valeurs de tous les champs, et relire les valeurs et les types des opérations effectuées en amont.
- Dans le menu contextuel, vous pouvez choisir de **copier** les attributs d'un champ à l'autre. Pour plus d'informations, voir «Copie d'attributs de type», à la page 156.
- A l'aide de l'option **Afficher les paramètres de champ non utilisés**, vous pouvez afficher les paramètres de type des champs qui ne figurent plus dans les données ou qui étaient auparavant connectés à ce noeud Typer. Cette option est utile lorsque vous réutilisez un noeud Typer pour des jeux de données qui ont été modifiés.

Niveaux de mesure

Le niveau de mesure (auparavant appelé « type de données » ou « type d'utilisation ») décrit l'utilisation des champs de données dans IBM SPSS Modeler. Le niveau de mesure peut être spécifié sur l'onglet Types d'une source ou d'un noeud type. Par exemple, vous pouvez définir le niveau de mesure de nombre entier comportant les valeurs 1 et 0 comme étant un champ *indicateur*. En général, 1 correspond à la valeur *True (vrai)* et 0 à la valeur *False (faux)*.

Stockage et mesure. Le type niveau de mesure d'un champ est différent de son type de stockage, lequel indique si les données sont stockées sous la forme d'une chaîne, d'un entier, d'un nombre réel, d'une date, d'une heure ou d'un horodatage. Si les types de données peuvent être modifiés en tout point d'un flux à l'aide d'un noeud Typer, le stockage doit, quant à lui, être déterminé au niveau de la source dans IBM SPSS Modeler (il peut cependant être modifié ultérieurement à l'aide d'une fonction de conversion). Pour plus d'informations, voir «Définition du stockage et du formatage des champs», à la page 9.

Certains noeuds de modélisation indiquent les types de niveau de mesure autorisés pour leurs champs d'entrée et de sortie à l'aide d'icônes sur leur onglet Champs.

Icones de niveau de mesure

Tableau 20. Icones de niveau de mesure

Icône	Niveau de mesure
	Par défaut
	Continu

Tableau 20. Icones de niveau de mesure (suite)

Icône	Niveau de mesure
	Catégorielle
	Indicateur
	Nominal
	Ordinal
	Sans type
	Résumé
	Géospatial

Les niveaux de mesure suivants sont disponibles :

- **Par défaut** Les données dont le type de stockage et les valeurs sont inconnus (car ces informations n'ont pas encore été lues par exemple) sont affichées avec la chaîne **<Par défaut>**.
- **Continu** Utilisé pour décrire des valeurs numériques, comme une plage de 0 à 100 ou de 0,75 à 1,25. Une valeur continue peut être un entier, un nombre réel ou une date/heure.
- **Catégorielle** Utilisé pour les valeurs de chaîne lorsqu'un nombre exact de valeurs distinctes est inconnu. Il s'agit d'un type de données **non instancié**, ce qui signifie que toutes les informations possibles sur le stockage et l'utilisation des données ne sont pas encore connues. Une fois les données lues, le niveau de mesure sera *Indicateur*, *Nominal* ou *Sans type*, selon le nombre maximal de membres spécifié pour les champs nominaux indiqués dans la boîte de dialogue Propriétés du flux.
- **Indicateur** Utilisé pour les données associées à deux valeurs distinctes qui indiquent la présence ou l'absence d'une caractéristique, comme `true` et `false`, `Oui` et `Non` ou `0` et `1`. Les valeurs utilisées peuvent varier, mais une d'elles doit toujours être désignée comme valeur « vrai » et l'autre comme valeur « faux ». Vous pouvez représenter les données sous forme de texte, d'entier, de nombre réel, de date, d'heure ou d'horodatage.
- **Nominal** Utilisé pour décrire des données associées à plusieurs valeurs distinctes, chacune traitée en tant que membre d'un ensemble, par exemple `petit/moyen/grand`. Les données nominales peuvent bénéficier de n'importe quel stockage numérique, chaînes ou date/heure. Le fait de définir le niveau de mesure *Nominal* n'a pas pour effet de convertir automatiquement les valeurs en stockage de chaîne.
- **Ordinal** Utilisé pour décrire des données associées à plusieurs valeurs distinctes qui possèdent un ordre inhérent. Par exemple, les catégories de salaire ou l'indice de satisfaction peuvent avoir le type données ordinales. L'ordre est défini par l'ordre de tri naturel des éléments des données. Par exemple, `1, 3, 5` est l'ordre de tri par défaut d'un ensemble d'entiers, alors que `ELEVE, FAIBLE, NORMAL` (tri alphabétique croissant) est l'ordre d'un ensemble de chaînes. Le niveau de mesure ordinal vous permet de définir un ensemble de données catégorielles comme des données ordinales, pour la visualisation, la création de modèles et l'exportation vers d'autres applications (telles que IBM SPSS Statistics) qui reconnaissent les données ordinales comme un type distinct. Vous pouvez utiliser le champ ordinal

partout où un champ nominal peut être utilisé. De plus, les champs de n'importe quel type de stockage (réel, entier, chaîne, date, heure, etc.) peuvent être définis comme ordinal.

- **Sans type** Utilisé pour les données qui ne sont pas conformes à l'un des types ci-dessus, pour les champs à valeur unique, ou pour les données nominales pour lesquelles l'ensemble compte plus de membres que le nombre maximal défini. Ce type s'avère pratique dans les cas où autrement le niveau de mesure serait un ensemble avec de nombreux membres (par exemple, un numéro de compte). Lorsque vous sélectionnez **Sans type** pour un champ, le rôle est automatiquement paramétré sur **Aucun**, avec **ID d'enregistrement** comme seule alternative. La taille maximale par défaut des ensembles est de 250 valeurs uniques. Ce nombre peut être ajusté ou désactivé dans l'onglet Options de la boîte de dialogue Propriétés du flux, à laquelle vous accédez à partir du menu Outils.
- **Collection** Utilisé pour identifier des données non géospatiales qui sont enregistrées dans une liste. Une collection est une zone de liste de profondeur zéro, dans laquelle les éléments de la liste sont associés à l'un des autres niveaux de mesure.

  Pour des informations sur les listes, voir la rubrique Stockage de liste et niveaux de mesure associés dans la section Noeuds source du guide Noeuds SPSS Modeler source, d'exécution et de sortie.
- **Géospatial** Utilisé avec le type de stockage Liste pour identifier des données géospatiales. Une liste peut être un champ Liste de nombres entiers ou Liste de nombres réels avec une profondeur de liste comprise entre zéro et deux inclus.

  Pour plus d'informations, voir la rubrique Sous-niveaux de mesure géospatiaux dans la section Noeud Typer du guide Noeuds SPSS Modeler source, d'exécution et de sortie.

Vous pouvez indiquer manuellement des niveaux de mesure, ou laisser le logiciel lire les données et déterminer le niveau de mesure en fonction des valeurs lues.

Vous pouvez aussi sélectionner une option, si vous avez plusieurs champs de données continues qui doivent être traitées comme des données catégorielles, afin de les convertir. Pour plus d'informations, voir «Conversion de données continues», à la page 149.

Utiliser la saisir automatique

1. Dans un noeud Typer ou dans l'onglet Types d'un noeud source, définissez la colonne **Valeurs** sur **<Lire>** pour les champs voulus. Les métadonnées sont ainsi disponibles pour tous les noeuds situés en aval. Vous pouvez rapidement paramétrer tous les champs sur **<Lire>** ou **<Transférer>** à l'aide des boutons représentant des lunettes de soleil dans la boîte de dialogue.
2. Cliquez sur **Lire les valeurs** pour lire les valeurs directement à partir de la source de données.

Définir manuellement le niveau de mesure d'un champ

1. Sélectionnez un champ dans le tableau.
2. Dans la liste déroulante de la colonne **Mesure**, sélectionnez le niveau de mesure du champ.
3. Vous pouvez également utiliser la combinaison Ctrl+A ou Ctrl+clic pour sélectionner plusieurs champs avant de choisir un niveau de mesure dans la liste déroulante.

Sous-niveaux de mesure géospatiaux

Le niveau de mesure Géospatial, qui est utilisé avec le type de stockage Liste, possède six sous-niveaux qui sont utilisés pour identifier différents types de données géospatiales.

- **Point** - Identifie un emplacement spécifique, par exemple le centre d'une ville.
- **Polygone** - Série de points qui identifie la frontière d'une région et son emplacement, par exemple un département.
- **Chaîne** - Aussi appelée ligne polygonale ou juste ligne, une chaîne est une série de points qui identifie la route d'une ligne. Par exemple, une chaîne peut être un élément fixe, comme une route, une rivière ou une voie de chemin de fer. Il peut aussi s'agir du trajet d'un élément qui se déplace, par exemple du trajet d'un avion ou d'un bateau.

- **Multipoint** - Utilisé lorsque chaque ligne de vos données contient plusieurs points par région. Par exemple, si chaque ligne représente une rue, plusieurs points pour chaque rue peuvent être utilisés afin d'identifier chaque réverbère.
- **Multipolygone** - Utilisé lorsque chaque ligne de vos données contient plusieurs polygones. Par exemple, si chaque ligne représente le contour d'un pays, les Etats-Unis peuvent être représentés par plusieurs polygones afin d'identifier les différentes zones, telles le continent, l'Alaska et Hawaï.
- **Multichaîne** - Utilisé lorsque chaque ligne de vos données contient plusieurs lignes. Etant donné que les lignes ne peuvent pas être divisées en branches, vous pouvez utiliser une multichaîne pour identifier un groupe de lignes, par exemple pour les données telles que les voies navigables ou le réseau ferroviaire dans chaque pays.

Ces sous-niveaux de mesure sont utilisés avec le type de stockage Liste. Pour plus d'informations, voir «Stockage de liste et niveaux de mesure associés», à la page 12.

Restrictions

Vous devez tenir compte de certaines restrictions lorsque vous utilisez des données géospatiales.

- Le système de coordonnées peut avoir un impact sur le format des données. Par exemple, un système de coordonnées projetées utilise les valeurs de coordonnées x, y et (si requis) z, alors qu'un système de coordonnées géographiques utilise les valeurs de coordonnées longitude, latitude et (si requis) une valeur pour l'altitude ou la profondeur.

  Pour plus d'informations sur les systèmes de coordonnées, voir la rubrique sur la définition des options géospatiales pour les flux de la section Utilisation des flux dans le guide d'utilisation de SPSS Modeler.
- Une chaîne ne peut pas se croiser elle-même.
- Un polygone ne se ferme pas lui-même ; pour chaque polygone, vous devez vous assurer que le premier point et le dernier point sont les mêmes.
- Le sens des données dans un multipolygone est important ; le sens des aiguilles d'une montre indique une forme solide et le sens inverse des aiguilles d'une montre indique un trou. Par exemple, si vous enregistrez une zone d'un pays qui contient des lacs, le continent peut être enregistré avec le sens des aiguilles d'une montre et la forme de chaque lac avec le sens inverse des aiguilles d'une montre.
- Un polygone ne peut pas présenter d'intersection avec lui-même. Par exemple, vous pourriez avoir une intersection si vous tentiez de représenter la frontière du polygone comme une ligne continue sous la forme illustrée par la figure 8.
- Les multipolygones ne peuvent pas se chevaucher.
- Pour les champs géospatiaux, les seuls types de stockage pertinents sont **Réel** et **Entier** (le paramètre par défaut est **Réel**).

Icônes des sous-niveaux de mesure géospatiaux

Tableau 21. Icônes des sous-niveaux de mesure géospatiaux

Icône	Niveau de mesure
	Point
	Polygone
	Chaîne

Tableau 21. Icônes des sous-niveaux de mesure géospatiaux (suite)

Icône	Niveau de mesure
	Multipoint
	Multipolygone
	multichaîne

Conversion de données continues

Le traitement de données catégorielles en tant que données continues peut avoir des effets importants sur la qualité d'un modèle, surtout s'il s'agit du champ cible, comme par exemple, la création d'un modèle de régression plutôt que d'un modèle binaire. Pour éviter cet inconvénient, vous pouvez convertir des intervalles d'entiers en des types catégoriels tels que *Ordinal* ou *Indicateur*.

1. Dans le bouton du menu Opérations et Générer (comportant le symbole d'outil), sélectionnez **Convertir des entiers continus en ordinaux**. La boîte de dialogue des valeurs de conversion s'affiche.
2. Spécifiez la taille de l'intervalle qui sera automatiquement converti ; cela s'applique à n'importe quel intervalle jusqu'à la taille (incluse) que vous avez saisie.
3. Cliquez sur **OK**. Les intervalles concernés sont convertis soit en *Indicateur* ou en *Ordinal* et sont affichés dans l'onglet Types du noeud type.

Résultats de la conversion

- Lorsqu'un champ *Continu* avec stockage d'entiers est modifié en *Ordinal*, les valeurs inférieures et supérieures sont étendues afin d'inclure toutes les valeurs entières, de la plus basse à la plus élevée. Par exemple, si l'intervalle est 1, 5, l'ensemble des valeurs est 1, 2, 3, 4, 5.
- Lorsque le champ *Continu* est converti en un champ *Indicateur*, les valeurs inférieures et supérieures deviennent des valeurs false (faux) et true (vrai) du champ indicateur.

Qu'est-ce que l'instanciation ?

L'**instanciation** est le processus qui consiste à lire ou à spécifier des informations, telles que le type de stockage ou les valeurs d'un champ de données. Afin d'optimiser les ressources système, l'instanciation est gérée par l'utilisateur Celui-ci demande au logiciel de lire les valeurs en spécifiant des options dans l'onglet Types d'un noeud source ou en exécutant des données via un noeud type.

- Les données dont le type est inconnu sont par ailleurs désignées comme **non instanciées**. Les données dont les valeurs et le type de stockage sont inconnus figurent dans la colonne *Mesure* de l'onglet Types sous la forme **<Par défaut>**.
- Lorsque vous disposez d'informations sur le stockage d'un champ (valeur numérique ou chaîne, par exemple), les données sont dites **partiellement instanciées**. **Les types Catégoriel** et **Continu** sont des mesures de niveau partiellement instanciés. Par exemple, le type **Catégoriel** indique que le champ est symbolique, mais vous ne savez pas s'il s'agit du type nominal, ordinal ou indicateur.
- Lorsque tous les détails sur un type sont connus, y compris les valeurs, le niveau de mesure **entièrement instancié** nominal, ordinal, indicateur ou continu est affiché dans cette colonne. *Remarque* : Le type *continu* est utilisé aussi bien pour les champs de données partiellement instanciés que pour ceux entièrement instanciés. Les données continues peuvent être des entiers ou des nombres réels.

Pendant l'exécution d'un flux de données avec un noeud Typer, les types non instanciés deviennent immédiatement partiellement instanciés, en fonction des valeurs de données initiales. Une fois que toutes les données sont passées dans le noeud, elles deviennent complètement instanciées sauf si les valeurs ont

été définies sur **<Transférer>**. Si l'exécution est interrompue, les données demeurent partiellement instanciées. Une fois l'onglet Types instancié, les valeurs des champs sont statiques à cet endroit du flux. Autrement dit, tout changement intervenant en amont n'affectera pas les valeurs d'un champ particulier, même si vous exécutez de nouveau le flux. Pour modifier ou mettre à jour les valeurs en fonction de nouvelles données ou de manipulations supplémentaires, vous devez les éditer directement dans l'onglet Types ou paramétrer la valeur des champs sur **<Lire>** ou **<Lire +>**.

Moment d'instanciation

En général, si votre jeu de données n'est pas trop volumineux et si vous ne prévoyez pas d'ajouter des champs au flux par la suite, l'instanciation au niveau du noeud source est la méthode la plus pratique. Cependant, l'instanciation dans un autre noeud Typer est utile lorsque :

- L'jeu de données est volumineux et le flux filtre un sous-ensemble avant le noeud Typer.
- Des données ont été filtrées dans le flux.
- Des données ont été fusionnées ou ajoutées dans le flux.
- De nouveaux champs de données sont calculés au cours du traitement.

Valeurs de données

La colonne **Valeurs** de l'onglet Types vous permet de lire automatiquement des valeurs à partir des données, ou de spécifier des niveaux de mesure et des valeurs dans une boîte de dialogue distincte.

Les options disponibles dans la liste déroulante Valeurs fournissent des instructions pour la définition automatique du type, comme illustré dans le tableau ci-après.

Tableau 22. Instructions pour la définition automatique du type

Option	Fonction
<Lire>	Les données sont lues lors de l'exécution du noeud.
<Lire +>	Les données sont lues et ajoutées aux données actuelles (le cas échéant).
<Passage>	Aucune lecture de données.
<En cours>	Les valeurs des données actuelles sont conservées.
Spécifier...	Une boîte de dialogue distincte s'ouvre pour que vous puissiez spécifier des valeurs et des options de niveau de mesure.

Si vous exécutez un noeud Typer ou que vous cliquez sur **Lire les valeurs**, le type est défini automatiquement et les valeurs sont lues à partir de votre source de données, en fonction de votre sélection. Vous pouvez également spécifier ces valeurs manuellement en utilisant l'option Spécifier ou en double-cliquant sur une cellule de la colonne **Champ**.

Une fois les champs du noeud Typer modifiés, vous pouvez réinitialiser les informations concernant les valeurs à l'aide des boutons suivants de la barre d'outils de la boîte de dialogue :

- A l'aide du bouton **Effacer les valeurs**, vous pouvez supprimer les changements apportés aux valeurs de champ de ce noeud (valeurs non héritées) et relire les valeurs des opérations effectuées en amont. Cette option est utile pour réinitialiser les changements que vous avez apportés à certains champs en amont.
- A l'aide du bouton **Effacer toutes les valeurs**, vous pouvez réinitialiser les valeurs de **tous** les champs lus dans le noeud. Cette option paramètre la colonne *Valeurs* de tous les champs sur **Lire**. Cette option est utile pour réinitialiser les valeurs de tous les champs et relire les valeurs et les niveaux de mesure des opérations effectuées en amont.

Texte en gris dans la colonne Valeurs

Dans un noeud Typer ou Source, si les données dans la colonne **Valeurs** apparaissent en noir, cela signifie que les valeurs de cette zone ont été lues et sont stockés dans ce noeud. Si elles n'apparaissent pas en noir, cela signifie que les valeurs de cette zone n'ont pas été lues et qu'elles sont déterminées plus loin en amont.

Les données apparaissent alors en gris. C'est le cas lorsque SPSS Modeler peut identifier ou déduire les valeurs valides d'un champ sans lire et stocker les données. En général, cette situation se produit lorsque vous utilisez l'un des noeuds suivants :

- Noeud Utilisateur. Comme les données sont définies dans le noeud, la plage de valeurs pour un champ est toujours connue, même si les valeurs n'ont pas été stockées dans le noeud.
- Noeud source Statistics. S'il existe des métadonnées pour les types de données, SPSS Modeler peut déduire la plage possible de valeurs sans lire ou stocker les données.

Dans les deux types de noeud, les valeurs sont affichées en gris jusqu'à ce que vous cliquiez sur **Lire les valeurs**.

Remarque : Si vous n'instanciez pas les données dans votre flux et que vos valeurs de données apparaissent en gris, les vérifications de valeurs de type que vous définissez dans la colonne **Vérification** ne sont pas appliquées.

Utilisation de la boîte de dialogue Valeurs

Cliquez sur la colonne **Valeurs** ou **Manquantes** de l'onglet Types affiche une liste déroulante des valeurs prédéfinies. Cliquez sur **Spécifier...** dans cette liste pour ouvrir une boîte de dialogue distincte dans laquelle vous pouvez définir des options pour la lecture, la spécification, l'étiquetage et le traitement des valeurs pour le champ sélectionné.

La majeure partie des contrôles sont communs à tous les types de données. Ces contrôles communs sont abordés ici.

Mesure Affiche le niveau de mesure sélectionné. Vous pouvez modifier le paramètre pour indiquer la façon dont vous souhaitez utiliser les données. Par exemple, si un champ appelé `jour_de_la_semaine` contient des nombres qui représentent des jours, vous pouvez changer ces informations en données nominales afin de créer un noeud de distribution qui examine chaque catégorie individuellement.

Stockage Affiche le type de stockage s'il est connu. Les types de stockage ne sont pas affectés par le niveau de mesure que vous choisissez. Pour modifier le type de stockage, vous pouvez utiliser l'onglet Données des noeuds source Fixe et Délimité, ou la fonction de conversion d'un noeud Remplacer.

Champ de modèle Pour les champs générés suite à l'évaluation d'un nugget de modèle, les détails du champ de modèle peuvent également être affichés. Il s'agit du nom du champ cible ainsi que du rôle du champ dans la modélisation (valeur prévue, probabilité, propension, etc.).

Valeurs Sélectionnez une méthode afin de déterminer les valeurs pour le champ sélectionné. Ces sélections annulent celles faites précédemment à partir de la colonne **Valeurs** de la boîte de dialogue du noeud Typer. Les choix pour la lecture des valeurs sont les suivants :

- **Lire à partir des données** Sélectionnez cette option pour lire les valeurs lorsque le noeud est exécuté. Elle est identique à **<Lire>**.
- **Passage** Sélectionnez cette option afin de ne pas lire les données pour le champ en cours. Elle est identique à **<Transférer>**.
- **Indiquer les valeurs et libellés** Ces options sont utilisées pour spécifier des valeurs et des libellés pour le champ sélectionné. Utilisées avec la vérification des valeurs, elles permettent de spécifier des valeurs

en fonction de votre connaissance du champ en cours. Elle active des contrôles propres à chaque type de champ. Les options des valeurs et des libellés sont abordées une par une dans les rubriques suivantes.

Remarque : Vous ne pouvez pas spécifier de valeurs ni de libellés pour un champ dont le niveau de mesure est Sans type ou <Par défaut>.

- **Etendre les valeurs à partir des données** Sélectionnez cette option pour ajouter aux données en cours les valeurs que vous entrez ici. Par exemple, si l'intervalle du `champ_1` est compris entre 0 et 10 (0,10), et que vous saisissez l'intervalle de valeurs (8,16), l'intervalle est augmenté via l'ajout de la valeur 16, sans que la valeur minimale d'origine soit supprimée. Le nouvel intervalle est (0,16). Si vous choisissez cette option, l'option de définition automatique du type est automatiquement paramétrée sur **<Lire +>**.

Longueur max de la liste Disponible seulement pour les données dont le niveau de mesure est Géospatial ou Collection. Définissez la longueur maximale de la liste en spécifiant le nombre d'éléments que la liste peut contenir.

Longueur de chaîne maximale Disponible uniquement pour les données sans type ; utilisez cette zone si vous générez un code SQL pour créer une table. Entrez la valeur de la plus grande chaîne de vos données ; cette opération génère une colonne dans la table qui est assez grande pour la chaîne. Si la valeur de la longueur de chaîne n'est pas disponible, une taille de chaîne par défaut non appropriée pour les données est utilisée (par exemple, si la valeur est trop faible, des erreurs peuvent se produire lors de l'écriture des données dans la table ; une valeur trop élevée peut avoir un impact négatif sur les performances).

Vérifier les valeurs Sélectionnez une méthode de conversion forcée des valeurs pour qu'elles soient conformes aux valeurs continues, indicateurs ou nominales spécifiées. Cette option correspond à la colonne **Vérifier** de la boîte de dialogue du noeud Typer ; les paramétrages effectués ici annulent ceux de la boîte de dialogue. Utilisée avec l'option **Indiquer les valeurs et libellés**, la vérification des valeurs permet de rendre les valeurs dans les données conformes aux valeurs attendues. Par exemple, si vous indiquez les valeurs 1, 0, l'option **Supprimer** vous permet de supprimer tous les enregistrements contenant des valeurs autres que 1 ou 0.

Définir les blancs Sélectionnez cette option pour activer les contrôles ci-dessous que vous utilisez pour déclarer des valeurs manquantes ou des blancs dans vos données.

- **Valeurs manquantes** Utilisez cette table pour définir des valeurs spécifiques (comme 99 ou 0) comme blancs. La valeur doit être adaptée au type de stockage du champ.
- **Plage** Utilisée pour spécifier une plage de valeurs manquantes, par exemple les âges compris entre 1 et 17 ou supérieurs à 65. Si une valeur de limite n'est pas indiquée, la plage est illimitée. Par exemple, si la limite inférieure de 100 est spécifiée sans limite supérieure, toutes les valeurs supérieures ou égales à 100 sont définies comme manquantes. Les valeurs de limite sont inclusives ; par exemple, une plage dont la limite inférieure est 5 et la limite supérieure est 10 inclut 5 et 10 dans la définition de la plage. Une plage de valeurs manquantes peut être définie pour tous les types de stockage, y compris pour les dates/heures et les chaînes (auquel cas l'ordre de tri alphabétique est utilisé pour déterminer si une valeur se trouve dans la plage).
- **Nul/Blanc** Vous pouvez également spécifier des valeurs null système (affichées dans les données sous la forme `$null$`) et des espaces (valeurs de chaîne sans caractère visible) sous forme de blancs.

  Remarque : Le noeud Typer traite également les chaînes vides comme des espaces pour l'analyse, bien qu'ils soient stockés différemment en interne et qu'ils puissent être traités différemment dans certains cas.

Remarque : Pour coder les blancs comme des valeurs non définies ou `$null$`, utilisez le noeud Remplacer.

Description Utilisez cette zone de saisie pour spécifier un libellé de champ. Ces libellés apparaissent à divers emplacements, par exemple dans des graphiques, des tableaux, la sortie et des navigateurs de modèle, selon vos sélections dans la boîte de dialogue Propriétés du flux.

Spécification de valeurs et d'libellés pour des données continues

Le niveau de mesure *Continu* permet de mesurer des champs numériques. Il existe trois types de stockages pour des données continues :

- Réel
- Entier
- Date/Heure

La même boîte de dialogue permet d'éditer tous les champs continus. Le type de stockage est affiché uniquement à titre de référence.

Spécification de valeurs

Les commandes suivantes sont propres aux champs continus et sont utilisées pour indiquer un intervalle de valeurs :

Inférieur. Indiquez la limite inférieure de l'intervalle des valeurs.

Supérieur. Indiquez la limite supérieure de l'intervalle des valeurs.

Spécification de libellés

Vous pouvez spécifier des libellés pour toutes les valeurs d'un champ d'intervalle. Cliquez sur le bouton **Libellés** pour ouvrir une boîte de dialogue distincte permettant de spécifier les libellés de valeur.

Sous-boîte de dialogue Valeurs et libellés : Cliquez sur **Libellés** dans la boîte de dialogue Valeurs d'un champ d'intervalle pour ouvrir une nouvelle boîte de dialogue permettant de spécifier des libellés pour les valeurs de votre choix dans l'intervalle.

Vous pouvez utiliser les colonnes *Valeurs* et *Libellés* de ce tableau pour définir des paires de valeurs et de libellés. Les paires actuellement définies sont indiquées ici. Pour ajouter de nouvelle paires de libellés, cliquez sur une cellule vide et entrez une valeur et son libellé. *Remarque* : L'ajout de paires valeur/valeur-libellé à ce tableau n'engendre l'ajout d'aucune nouvelle valeur au champ. Cela crée simplement des métadonnées pour la valeur du champ.

Les libellés indiquées dans le noeud Typer s'affichent dans de nombreux emplacements (sous forme d'info-bulles, de libellés de sortie, etc.), selon les éléments sélectionnés dans la boîte de dialogue Propriétés du flux.

Spécification des valeurs et des libellés pour des données nominales et ordinales

Les niveaux de mesure nominaux (ensemble) et ordinaux (ensemble ordonné) indiquent que les valeurs de données sont utilisées discrètement en tant que membres de l'ensemble. Les ensembles disposent des types de stockage suivants : chaîne, entier, nombre réel ou date/heure.

Les commandes suivantes sont propres aux champs nominaux et ordinaux et sont utilisées pour indiquer les valeurs et les libellés :

Valeurs. La colonne *Valeurs* du tableau vous permet de spécifier des valeurs, selon la connaissance que vous avez du champ actuel. Grâce à ce tableau, vous pouvez saisir des valeurs théoriques pour le champ et vérifier la conformité du jeu de données par rapport à ces valeurs à l'aide de la liste déroulante Vérifier les valeurs. A l'aide des flèches et du bouton Supprimer, vous pouvez modifier, réorganiser ou supprimer les valeurs existantes.

Libellés. La colonne *Libellés* permet de spécifier des libellés pour chaque valeur de l'ensemble. Ces libellés apparaissent dans divers emplacements, tels que des graphiques, des tableaux, des résultats et des navigateurs de modèle, selon vos sélections dans la boîte de dialogue Propriétés du flux.

Spécification des valeurs d'un champ booléen

Les champs booléens servent à afficher les données possédant deux valeurs distinctes. Les booléens disposent des types de stockage suivants : chaîne, entier, nombre réel ou date/heure.

Vrai. Spécifiez la valeur booléenne de la zone lorsque la condition est respectée.

Faux. Spécifiez la valeur booléenne du champ lorsque la condition n'est pas respectée.

Libellés. Spécifiez des libellés pour chaque valeur du champ booléen. Ces libellés apparaissent dans divers emplacements, tels que des graphiques, des tableaux, des résultats et des navigateurs de modèle, selon vos sélections dans la boîte de dialogue Propriétés du flux.

Spécification de valeurs pour les données de collection

Les champs de collection sont utilisés pour afficher des données non géospatiales qui figurent dans une liste.

Le seul élément que vous pouvez définir pour le niveau **Mesure** de collection est **Mesure de liste**. Par défaut, cette mesure est sans type mais vous pouvez sélectionner une autre valeur pour définir le niveau de mesure des éléments dans la liste. Vous pouvez choisir l'une des options suivantes :

- Sans type
- Continu
- Nominal
- Ordinal
- Indicateur

Spécification de valeurs pour les données geospatiales

Les champs géospatiaux sont utilisés pour afficher des données géospatiales qui se trouvent dans une liste.

Pour le niveau de **mesure** géospatial, vous pouvez définir les options suivantes afin de définir le niveau de mesure des éléments dans la liste :

Type Sélectionnez le sous-niveau de mesure du champ géospatial. Les sous-niveaux disponibles sont déterminés par la profondeur de la zone de liste. Les valeurs par défaut sont Point (pas de profondeur), Chaîne (profondeur de un) et Polygone (profondeur de un).

Pour plus d'informations sur les sous-niveaux, voir «Sous-niveaux de mesure géospatiaux», à la page 147.

Pour plus d'informations sur les profondeurs de liste, voir «Stockage de liste et niveaux de mesure associés», à la page 12.

Système de coordonnées Cette option est disponible uniquement si vous avez changé le niveau de mesure en remplaçant un niveau non-géospatial par un niveau géospatial. Pour appliquer un système de coordonnées à vos données géospatiales, sélectionnez cette case à cocher. Par défaut, le système de coordonnées défini dans le panneau **Outils** > **Propriétés du flux** > **Options** > **Géospatial** est affiché. Pour utiliser des systèmes de coordonnées différents, cliquez sur le bouton **Changer** afin d'ouvrir la boîte de dialogue Sélection d'un système de coordonnées et choisissez le système de votre choix.

Pour plus d'informations sur les systèmes de coordonnées, voir la rubrique sur la définition des options géospatiales pour les flux de la section Utilisation des flux dans le guide d'utilisation de SPSS Modeler.

Définition de valeurs manquantes

La colonne **Manquant** de l'onglet Types indique si le traitement des valeurs manquantes a été défini pour un champ. Les paramètres possibles sont :

Activé (*). Indique que le traitement des valeurs manquantes est défini pour ce champ. Ceci peut être effectué à l'aide d'un noeud Remplacer en aval, ou à travers une spécification explicite avec l'option Spécifier (voir ci-dessous).

Désactivé. Le champ n'a pas de traitement des valeurs manquantes défini.

Spécifier. Choisissez cette option pour afficher une boîte de dialogue dans laquelle vous pouvez déclarer des valeurs explicites à considérer comme des valeurs manquantes pour ce champ.

Vérification des valeurs de type

Activez l'option Vérifier de chaque champ pour examiner toutes les valeurs de ce champ, et déterminer si elles sont conformes aux paramètres de type actuels ou aux valeurs spécifiées dans la boîte de dialogue Indiquer les valeurs. Cette option est pratique pour nettoyer les jeux de données et réduire leur taille en une seule opération.

Le paramètre de la colonne *Vérifier* de la boîte de dialogue du noeud Typer détermine ce qui se produit si une valeur hors limites est découverte. Pour modifier les paramètres Vérifier d'un champ, utilisez la liste déroulante correspondante dans la colonne *Vérifier*. Pour définir les paramètres Vérifier de tous les champs, cliquez dans la colonne *Champ* et appuyez sur Ctrl+A. Utilisez ensuite la liste déroulante de n'importe quel champ de la colonne *Vérifier*.

Les paramètres Vérifier suivants sont disponibles :

Aucun. Les valeurs sont transmises sans être vérifiées. Il s'agit du paramètre par défaut.

Rendre nul. Convertit les valeurs hors limites en valeurs système nulles (`$null$`).

Forcer. Une recherche des valeurs situées hors de l'intervalle indiqué est effectuée sur les champs dont les niveaux de mesure sont complètement instanciés. Ces valeurs sont converties en valeurs adaptées au niveau de mesure selon les règles suivantes :

- Pour les booléens, les valeurs autres que "true" et "false" sont converties en valeurs "false".
- Pour les ensembles (nominaux et ordinaux), les valeurs inconnues sont converties en la valeur du premier membre des valeurs de l'ensemble.
- Les nombres supérieurs à la limite supérieure d'un intervalle sont remplacés par la valeur de cette limite.
- Les nombres inférieurs à la limite inférieure d'un intervalle sont remplacés par la valeur de cette limite.
- Les valeurs nulles d'un intervalle prennent la valeur médiane de cet intervalle.

Annuler. Lorsque des valeurs incorrectes sont trouvées, l'intégralité de l'enregistrement est supprimé.

Avertir. Le nombre d'éléments incorrects est calculé et reporté dans la boîte de dialogue des propriétés du flux une fois toutes les données lues.

Stopper. La première valeur incorrecte rencontrée met fin à l'exécution du flux. L'erreur est reportée dans la boîte de dialogue des propriétés du flux.

Définition du rôle de champ

Le rôle d'un champ indique comment il est utilisé lors de la création de modèles, par exemple, s'il s'agit d'un champ d'entrée ou d'un champ cible (chose prévue).

Remarque : Les rôles de partition, de fréquence et d'ID d'enregistrement peuvent être chacun appliqués à un seul champ.

Les rôles suivants sont disponibles :

Entrée. Le champ est utilisé comme entrée pour l'apprentissage automatique (champ prédicteur).

Cible. Le champ est utilisé comme sortie ou cible pour l'apprentissage automatique (l'un des champs que le modèle essaie de prédire).

Les deux. Le champ est utilisé comme entrée et sortie par le noeud Apriori. Tous les autres noeuds de modélisation ignorent ce champ.

Aucun. Le champ est ignoré par l'apprentissage automatique. Les champs dont le niveau de mesure est défini sur **Sans type** sont automatiquement définis sur **Aucun** dans la colonne **Rôle**.

Partition. Indique un champ utilisé pour partitionner les données en échantillons distincts pour l'apprentissage, le test et la validation (facultatif). Le champ doit être un type d'ensemble instancié avec deux ou trois valeurs possibles (telles qu'elles sont définies dans la boîte de dialogue Valeurs de champ). La première valeur représente l'échantillon d'apprentissage, le second l'échantillon de test et le troisième (s'il existe) l'échantillon de validation. Toutes les valeurs supplémentaires sont ignorées et les champs booléens ne peuvent pas être utilisés. Pour utiliser la partition dans une analyse, vous devez l'activer dans l'onglet Options de modèle du noeud de génération de modèle ou d'analyse approprié. Les enregistrements du champ de partition comportant des valeurs nulles sont exclus de l'analyse lorsque la fonction de partition est activée. Si plusieurs champs de partition ont été définis dans le flux, un champ de partition unique doit être indiqué dans l'onglet Champs de chaque noeud de modélisation applicable. Si aucun champ adapté n'existe encore dans vos données, vous pouvez en créer un via un noeud Partitionner ou Calculer. Pour plus d'informations, voir «Noeud Partitionner», à la page 185.

Split. (Champs nominaux, ordinaux et indicateurs) Spécifie qu'un modèle doit être construit pour chaque valeur possible du champ.

Fréquence. (Champs numériques uniquement) La définition de ce rôle permet d'utiliser la valeur du champ comme un facteur de pondération de fréquence pour l'enregistrement. Cette caractéristique est uniquement prise en charge par les modèles C&R Tree, CHAID, QUEST et linéaires ; tous les autres noeuds ignorent ce rôle. La pondération de fréquence est activée au moyen de l'option **Utiliser la pondération de fréquence** de l'onglet Champs de ces noeuds de modélisation qui prennent en charge cette caractéristique.

ID enregistrement. Le champ est utilisé comme identificateur d'enregistrement unique. Cette fonction est ignorée par la plupart des noeuds ; cependant, elle est prise en charge par les modèles linéaires et requise pour les noeuds d'exploration de la base de données Netezza IBM.

Copie d'attributs de type

Vous pouvez facilement copier les attributs d'un type, tels que les valeurs, les options de vérification et les valeurs manquantes, d'un champ à l'autre :

1. Cliquez avec le bouton droit de la souris sur le champ dont vous souhaitez copier les attributs.
2. Dans le menu contextuel, choisissez **Copier**.
3. Cliquez avec le bouton droit de la souris sur les champs dont vous souhaitez changer les attributs.
4. Dans le menu contextuel, sélectionnez **Collage spécial**. *Remarque* : Vous pouvez sélectionner plusieurs champs en appuyant sur Ctrl tout en cliquant sur les champs ou en choisissant l'option **Sélectionner les champs** dans le menu contextuel.

Une nouvelle boîte de dialogue apparaît ; elle vous permet de sélectionner les attributs spécifiques que vous souhaitez coller. Si vous collez des attributs dans plusieurs champs, les options sélectionnées ici s'appliquent à tous les champs cible.

Coller les attributs suivants. Sélectionnez parmi les options ci-dessous les attributs à coller d'un champ à l'autre.

- **Type.** Sélectionnez cette option pour coller le niveau de mesure.
- **Valeurs.** Sélectionnez cette option pour coller les valeurs de champ.
- **Manquant.** Sélectionnez cette option pour coller les paramètres des valeurs manquantes.
- **Vérifier.** Sélectionnez cette option pour coller les options de vérification des valeurs.
- **Rôle.** Sélectionnez cette option pour coller le rôle d'un champ.

Onglet Paramètres du champ

L'onglet Format des noeuds Table et Typer répertorie les champs actuels et non utilisés, ainsi que les options de formatage de chaque champ. Les colonnes du tableau de formatage des champs sont décrites ci-dessous :

Champ. Indique le nom du champ sélectionné.

Format. Double-cliquez sur une cellule de cette colonne pour spécifier le formatage de chacun des champs à l'aide de la boîte de dialogue appelée. Pour plus d'informations, voir «Définition des options de formatage des champs», à la page 158. Le formatage indiqué ici remplace celui indiqué dans les propriétés générales du flux.

Remarque : Les noeuds Export Statistiques et Sortie Statistiques exportent des fichiers *.sav* comportant dans leurs métadonnées un formatage par champ. Si l'un des formats par champ indiqués n'est pas pris en charge par le format de fichier IBM SPSS Statistics *.sav* , le noeud utilise le format IBM SPSS Statistics par défaut.

Justifier. Utilisez cette colonne pour indiquer le mode de justification des valeurs dans les colonnes du tableau. Le paramètre par défaut est **Auto** : il justifie les valeurs symboliques vers la gauche et les valeurs numériques vers la droite. Vous pouvez remplacer ce paramètre par défaut en sélectionnant **Gauche**, **Droite** ou **Au milieu**.

Largeur de colonne. Par défaut, les largeurs de colonne sont automatiquement calculées sur la base des valeurs du champ. Pour remplacer le calcul automatique de la largeur, cliquez sur une cellule du tableau et utilisez la liste déroulante pour sélectionner une nouvelle largeur. Pour entrer une largeur personnalisée non répertoriée ici, ouvrez la sous-boîte de dialogue Format de champ en double-cliquant sur une cellule de la colonne Champ ou Format dans le tableau. Vous pouvez également cliquer avec le bouton droit de la souris sur une cellule et sélectionner **Définir le format**.

Afficher les champs actuels. Par défaut, la boîte de dialogue contient la liste des champs actifs. Pour afficher la liste des champs inutilisés, sélectionnez **Afficher les paramètres de champ non utilisés**.

Menu Contexte. Le menu Contexte de cet onglet contient des options de sélection et de mise à jour des paramètres. Cliquez avec le bouton droit dans une colonne pour afficher ce menu.

- **Sélectionner tout.** Sélectionne l'ensemble des champs.
- **Ne rien sélectionner.** Supprime la sélection.
- **Sélectionner les champs.** Sélectionne des champs sur la base de leur type ou de leur caractéristique de stockage. Les options disponibles sont les suivantes : **Sélectionner catégoriel**, **Sélectionner continu** (données numériques), **Sélectionner Sans Type**, **Sélectionner Chaînes**, **Sélectionner Nombres** ou **Sélectionner Date/Heure**. Pour plus d'informations, voir «Niveaux de mesure», à la page 145.

- **Définir le format.** Ouvre une sous-boîte de dialogue permettant de spécifier les options de date, d'heure et décimales par champ.
- **Définir la justification.** Définit le mode de justification des champs sélectionnés. Les options sont les suivantes : **Auto**, **Au milieu**, **Gauche** ou **Droite**.
- **Définir la largeur de colonne.** Définit la largeur des champs sélectionnés. Indiquez **Automatique** pour lire la largeur dans les données. Vous pouvez également définir la largeur du champ sur 5, 10, 20, 30, 50, 100 ou 200.

Définition des options de formatage des champs

Le formatage des champs est spécifié dans une sous-boîte de dialogue disponible à partir de l'onglet Format des noeuds Type et Table. Si vous avez sélectionné plusieurs champs avant d'ouvrir cette boîte de dialogue, les paramètres du premier champ de la sélection sont utilisés pour tous les champs. Cliquez sur **OK** une fois les spécifications définies ici pour appliquer ces paramètres à tous les champs sélectionnés dans l'onglet Format.

Les options suivantes sont disponibles par champ. Vous pouvez également spécifier la plupart de ces paramètres dans la boîte de dialogue Propriétés du flux. Tous les paramètres définis au niveau du champ remplacent les paramètres par défaut indiqués pour le flux.

Format de date. Sélectionnez le format de date à utiliser pour les champs de stockage de date ou lorsque les chaînes sont interprétées comme des dates par les fonctions de date CLEM.

Format d'heure. Sélectionnez le format d'heure à utiliser pour les champs de stockage d'heure ou lorsque les chaînes sont interprétées comme des heures par les fonctions d'heure CLEM.

Format d'affichage des nombres. Vous pouvez sélectionner les formats d'affichage standard (**####.###**), scientifique (**#.###E+##**) ou monétaire (**### ## €**).

Symbole décimal. Sélectionnez la virgule (,) ou le point (.) comme séparateur décimal.

Symbole de regroupement. Pour les formats d'affichage des nombres, sélectionnez le symbole permettant de regrouper des valeurs (par exemple, l'espace dans *3 000,00*). Vous avez le choix entre les options suivantes : aucun, point, virgule, espace et paramètres régionaux définis (auquel cas la valeur par défaut des paramètres régionaux actuels est utilisée).

Nombre de décimales (au format standard, scientifique ou monétaire, ou à exporter). Pour les formats d'affichage des nombres, indique le nombre de décimales à utiliser pour l'affichage des nombres réels. Cette option apparaît séparément pour chaque format d'affichage. Notez que le paramètre **Exporter le nombre de décimales** ne s'applique qu'aux exportations de fichier à plat et qu'il remplace les propriétés de flux. La valeur par défaut du flux pour l'exportation de fichier à plat correspond à la valeur spécifiée pour le paramètre **Nombre de décimales (format standard)** des propriétés du flux. Le nombre de décimales exportées par le noeud Export XML est toujours de 6.

Justifier. Indique le mode de justification des valeurs dans la colonne. Le paramètre par défaut est **Auto** : il justifie les valeurs symboliques vers la gauche et les valeurs numériques vers la droite. Vous pouvez remplacer ce paramètre par défaut en sélectionnant Gauche, Droite ou Au milieu.

Largeur des colonnes. Par défaut, les largeurs de colonne sont automatiquement calculées sur la base des valeurs du champ. Vous pouvez spécifier des largeurs personnalisées par intervalles de cinq à l'aide des flèches situées à droite de la zone de liste.

Filtrage ou modification du nom des champs

Vous pouvez renommer ou exclure des champs à tout stade d'un flux. Par exemple, en tant que chercheur en médecine, vous n'êtes peut-être pas intéressé par le niveau de potassium (données de niveau champ) des patients (données de niveau enregistrement) ; vous pouvez donc filtrer le champ `K` correspondant. Vous pouvez réaliser ceci à l'aide d'un noeud Filtrer distinct ou d'un onglet Filtrer sur un noeud source ou de sortie. Cette fonctionnalité est identique quel que soit le noeud à partir duquel vous y accéder.

- A partir de noeuds source, tels que Délimité, Fixe, Fichier Statistiques, XML Ou importation d'extension, vous pouvez renommer ou filtrer les champs à mesure que les données sont lues dans IBM SPSS Modeler.
- Le noeud Filtrer permet de renommer ou de filtrer les champs en tout point du flux.
- A partir des noeuds Export Statistiques, Transformation Statistiques, Modèle Statistiques et Sortie Statistiques, vous pouvez filtrer ou renommer les champs pour respecter les conventions de dénomination IBM SPSS Statistics. Pour plus d'informations, voir «Changement du nom ou filtrage des champs pour IBM SPSS Statistics», à la page 383.
- Vous pouvez utiliser l'onglet Filtrer dans l'un des noeuds susmentionnés pour définir ou modifier des ensembles de réponses multiples. Pour plus d'informations, voir «Modification des ensembles de réponses multiples», à la page 161.
- Finalement, vous pouvez utiliser un noeud Filtrer pour mapper les champs entre un noeud source et un autre.

Définition des options de filtrage

Le tableau utilisé dans l'onglet Filtrer affiche le nom des champs dès qu'ils entrent ou sortent du noeud. Vous pouvez utiliser les options de ce tableau pour renommer ou filtrer les champs qui sont en double ou inutiles pour les opérations en aval.

- **Champ.** Affiche les champs d'entrée des sources de données actuellement connectées.
- **Filtrer.** Affiche l'état de filtrage de tous les champs d'entrée. Les champs filtrés comportent un X rouge dans cette colonne, ce qui indique qu'ils ne seront pas transférés en aval. Cliquez dans la colonne *Filtrer* d'un champ sélectionné pour activer ou désactiver le filtrage. Vous pouvez également sélectionner des options pour plusieurs champs en même temps, en utilisant la méthode de sélection Maj+clic.
- **Champ.** Affiche les champs lorsqu'ils quittent le noeud Filtrer. Les noms en double sont affichés en rouge. Pour éditer les noms des champs, cliquez dans la colonne et saisissez un nouveau nom. Vous pouvez également supprimer les champs en cliquant dans la colonne *Filtrer* pour désactiver les champs en double.

Vous pouvez trier toutes les colonnes du tableau en cliquant sur l'en-tête de la colonne.

Afficher les champs actuels. Sélectionnez cette option pour afficher les champs des jeux de données connectés au noeud Filtrer. Cette méthode standard d'utilisation des noeuds Filtrer est sélectionnée par défaut.

Afficher les paramètres de champ non utilisés. Sélectionnez cette option pour afficher les champs des jeux de données qui étaient auparavant connectés au noeud Filtrer. Cette option est utile lorsque vous copiez des noeuds Filtrer d'un flux à un autre, ou lorsque vous enregistrez ou rechargez des noeuds Filtrer.

Menu du bouton Filtrer

Cliquez sur le bouton Filtrer en haut à gauche de la boîte de dialogue pour accéder à un menu qui propose un certain nombre de raccourcis et d'autres options.

Vous pouvez :

- Supprimer tous les champs.
- Inclure tous les champs.
- Basculer tous les champs.
- Supprimer les doublons. Notez que la sélection de cette option entraîne la suppression de toutes les occurrences du nom en double, y compris la première.
- Renommer les champs et les ensembles de réponses multiples pour être en conformité avec d'autres applications . Pour plus d'informations, voir «Changement du nom ou filtrage des champs pour IBM SPSS Statistics», à la page 383.
- Tronquer les noms de champ.
- Anonymiser les noms de champs et d'ensembles de réponses multiples.
- Utiliser les noms de champ d'entrée.
- Modifier les ensembles de réponses multiples. Pour plus d'informations, voir «Modification des ensembles de réponses multiples», à la page 161.
- Définir l'état de filtrage par défaut.

Vous pouvez également utiliser les boutons bascule représentant une flèche, en haut de la boîte de dialogue, pour indiquer si vous souhaitez, par défaut, inclure ou ignorer les champs. Ces boutons sont particulièrement utiles lorsque vous travaillez avec des jeux de données volumineux dans lesquels seuls quelques champs doivent être inclus en aval. Par exemple, vous pouvez sélectionner uniquement les champs que vous souhaitez conserver et indiquer que tous les autres doivent être ignorés (au lieu de sélectionner chaque champ à ignorer).

Troncation des noms de champ

A partir du menu du bouton Filtrer (en haut à gauche de l'onglet Filtrer), vous pouvez choisir de tronquer des noms de champs.

Longueur maximale. Limitez la longueur des noms de champ en indiquant un nombre de caractères.

Nombre de chiffres. Si, une fois raccourci, le nom d'un champ n'est plus unique, il est de nouveau raccourci et assorti d'un chiffre permettant de le différencier des autres. Vous pouvez indiquer le nombre de chiffres à utiliser. Utilisez les flèches pour rectifier ce nombre.

Par exemple, le tableau ci-dessous indique comment les noms de champ d'un jeu de données médicales sont raccourcis en fonction des paramètres par défaut (Longueur maximale = 8 et Nombre de chiffres = 2).

Tableau 23. Troncature du nom de champ

Noms de champs	Noms de champ raccourcis
Entrée patient 1	Patien01
Entrée patient 2	Patien02
Rythme cardiaque	RythmeCa
TA	TA

Anonymisation des noms de champ

Vous pouvez anonymiser des noms de champs à partir d'un noeud quelconque qui comporte un onglet Filtrer en cliquant sur le menu du bouton Filtrer et en choisissant **Anonymiser des noms de champ**. Les noms de champ anonymisés sont formés d'un préfixe de chaîne suivi d'une valeur numérique unique.

Anonymiser des noms de. Choisissez **Champs sélectionnés uniquement** pour n'anonymiser que les noms des champs déjà sélectionnés dans l'onglet Filtrer. La valeur par défaut est **Tous les champs**, qui anonymise tous les noms de champ.

Préfixe des noms de champ. Le préfixe par défaut des noms de champ anonymisés est **anon_**. Si vous souhaitez le modifier, choisissez **Personnalisé** et saisissez votre propre préfixe.

Anonymiser des ensembles de réponses multiples. Anonymise le nom des ensembles de réponses multiples de la même manière que les champs. Pour plus d'informations, voir «Modification des ensembles de réponses multiples».

Pour restaurer les noms de champ d'origine, choisissez **Utiliser les noms de champ d'entrée** dans le menu Filtrer.

Modification des ensembles de réponses multiples

Vous pouvez ajouter ou modifier des ensembles de réponses multiples à partir d'un noeud quelconque qui comporte un onglet Filtrer en cliquant sur le menu du bouton Filtrer et en choisissant **Editer des ensembles de réponses multiples**.

Les ensembles de réponses multiples permettent de consigner des données pouvant comporter plusieurs valeurs pour chaque cas, par exemple, lorsque les personnes sondées sont interrogées sur les musées qu'ils ont visités ou les magazines qu'ils ont lus. Il est possible d'importer des ensembles de réponses multiples dans IBM SPSS Modeler à l'aide d'un noeud source Data Collection ou Statistics. En outre, vous pouvez les définir dans IBM SPSS Modeler à l'aide d'un noeud Filtrer.

Cliquez sur **Nouveau** pour créer un nouvel ensemble de réponses multiples ou sur **Modifier** pour les modifier.

Nom et libellé. Indique le nom et la description de l'ensemble.

Type. Les questions à réponses multiples peuvent être traitées de l'une des deux manières suivantes :

- **Ensemble de dichotomies multiples.** Un champ indicateur distinct est créé pour chaque réponse possible. Par conséquent, pour 10 magazines, il y a 10 champs indicateurs, dont chacun comprend des valeurs telles que 0 ou 1 pour *vrai* ou *faux*. La valeur calculée permet de préciser celle qui est considérée comme étant 'vrai'. Cette méthode est utile pour permettre aux personnes interrogées de choisir toutes les options applicables.
- **Ensemble de catégories multiples.** Un champ nominal est créé pour chaque réponse jusqu'au nombre maximum de réponses d'une personne interrogée donnée. Chaque champ nominal comprend des valeurs qui réprésentent les réponses possibles, telles que *1* pour *Temps*, *2* pour *Newsweek* et *3* pour *PC Week*. Cette méthode est très utile lorsque vous souhaitez limiter le nombre de réponses, par exemple, lorsque les personnes sondées sont interrogées sur les trois magazines les plus souvent lus.

Champs dans l'ensemble. Les icônes de droite permettent d'ajouter ou de supprimer des champs.

Commentaires

- Tous les champs inclus dans un ensemble de réponses multiples doivent disposer du même stockage.
- Les ensembles sont distincts des champs qu'ils comportent. Par exemple, la suppression d'un ensemble n'entraîne pas celle des champs inclus dans celui-ci, simplement des liens entre ces champs. L'ensemble reste visible en amont du point de suppression mais pas en aval.
- Si des champs sont renommés à l'aide d'un noeud Filtrer (directement sur l'onglet ou en choisissant les options Renommer pour IBM SPSS Statistics, **Tronquer**, ou **Anonymiser** sur le menu Filtrer), toute référence à ces champs utilisée dans plusieurs ensembles de réponses est également mise à jour. Toutefois, tout champ dans un ensemble de réponses multiples qui est supprimé par le noeud Filtrer ne l'est pas de l'ensemble de réponses multiples. De tels champs, bien que masqués dans le flux, sont encore référencés par l'ensemble de réponses multiples. Ceci peut être pris en compte lors de l'exportation, par exemple.

Noeud Calculer

L'une des fonctionnalités les plus performantes d'IBM SPSS Modeler est la capacité à modifier les valeurs de données et à calculer de nouveaux champs à partir de données existantes. Au cours des projets d'exploration de données très longs, il est courant d'effectuer plusieurs calculs tels que l'extraction d'un ID client d'une chaîne des données du log Web ou la création d'une valeur de durée de vie de client basée sur des données démographiques et de transaction. Toutes ces transformations peuvent être effectuées à l'aide des divers noeuds d'opérations sur les champs.

Plusieurs noeuds permettent de calculer de nouveaux champs :

Le noeud Calculer modifie les valeurs de données ou crée des nouveaux champs à partir d'un ou de plusieurs champs existants. Il crée des champs de type formule, indicateur, ensemble, nominal, statistiques, comptage et conditionnel.

Le noeud Recoder permet de transformer un ensemble de valeurs catégorielles en un autre. La recodification est utile pour réduire des catégories ou regrouper des données à analyser.

Le noeud Discrétiser crée automatiquement des champs nominaux (ensemble) sur la base des valeurs d'un ou de plusieurs champs continus (intervalle numérique) existants. Par exemple, vous pouvez transformer un champ continu de revenus en un nouveau champ catégoriel contenant des groupes de revenus comme écarts par rapport à la moyenne. Une fois les intervalles du nouveau champ créés, vous pouvez générer un noeud Calculer à partir des points de césure.

Le noeud Binariser calcule plusieurs champs indicateurs en fonction des valeurs catégorielles définies pour un ou plusieurs champs nominaux.

Le noeud Restructurer convertit un champ nominal ou un champ indicateur en un groupe de champs renseignés à partir des valeurs d'un autre champ. Par exemple, si l'on considère un champ nommé *type de paiement*, qui comporte les valeurs *crédit*, *liquide* et *débit*, trois champs sont alors créés (*crédit*, *liquide*, *débit*), chacun contenant la valeur du paiement réel effectué.

Le noeud Historiser crée des champs contenant des données provenant de champs d'enregistrements antérieurs. Les noeuds Historiser sont souvent utilisés pour les données séquentielles, telles que les séries temporelles. Avant d'utiliser un noeud Historiser, vous pouvez trier les données à l'aide d'un noeud Trier.

Utilisation du noeud Calculer

A l'aide du noeud Calculer, vous pouvez créer six types de nouveau champ à partir d'un ou de plusieurs champs existants :

- **Formule.** Le nouveau champ est le résultat d'une expression CLEM arbitraire.
- **Booléen.** Le nouveau champ est un indicateur, représentant une condition spécifique.

- **Nominal.** Le nouveau champ est un champ nominal. Autrement dit, ses membres constituent un groupe de valeurs spécifiées.
- **Etat.** Le nouveau champ est l'un de deux états. Le passage d'un état à l'autre est déclenché par une condition donnée.
- **Comptage.** Le nouveau champ est basé sur le nombre de fois qu'une condition est vraie (true).
- **Conditionnel.** Le nouveau champ est la valeur de l'une des deux expressions, selon la valeur d'une condition.

Chacun de ces noeuds contient un ensemble d'options particulières dans la boîte de dialogue du noeud Calculer. Ces options sont traitées dans des rubriques ultérieures.

Notez que l'utilisation de l'une des tâches suivantes peut changer l'ordre des lignes :
- Exécution dans une base de données via la répercussion SQL
- Exécution via le serveur IBM SPSS Analytic Server distant
- Utilisation de fonctions exécutées sur le serveur IBM SPSS Analytic Server imbriqué
- Calcul d'une liste (pour un exemple, voir «Calcul d'une liste ou d'un champ géospatial», à la page 166)
- Appel de l'une des fonctions décrites dans Fonctions spatiales

Définition des options de base du noeud dériver

Vous trouverez dans la partie supérieure de la boîte de dialogue des noeuds dériver, plusieurs options permettant de sélectionner le type de noeud dériver dont vous avez besoin.

Mode. Sélectionnez le mode **Simple** ou **Multiple**, selon que vous souhaitez ou non calculer plusieurs champs. Lorsque le mode **Multiple** est sélectionné, la boîte de dialogue change : de nouvelles options, adaptées à plusieurs champs de calcul, apparaissent.

Dériver champ. Pour les noeuds dériver simples, spécifiez le nom du champ que vous voulez calculer et ajouter à chaque enregistrement. Le nom par défaut est Dériver *N*, où *N* correspond au nombre de noeuds dériver créés jusqu'ici dans la session actuelle.

Dériver en tant que. Dans la liste déroulante, sélectionnez le type de noeud dériver, tel que Formule ou Nominal. Pour chaque type, un nouveau champ est créé en fonction des conditions que vous spécifiez dans la boîte de dialogue correspondante.

La sélection d'une option dans la liste déroulante a pour effet d'ajouter un nouvel ensemble de contrôles dans la boîte de dialogue principale selon les propriétés de chaque type de noeud dériver.

Type de champ. Sélectionnez le niveau de mesure du nouveau noeud calculé (par exemple, Continu, Catégoriel ou Indicateur). Cette option est commune à toutes les formes de noeuds dériver.

Remarque : le calcul des nouveaux champs demande souvent l'utilisation de fonctions ou expressions mathématiques particulières. Pour créer ces expressions, vous pouvez utiliser le Générateur de formules, disponible à partir de la boîte de dialogue de tous les types de noeud dériver. Il permet de vérifier les règles et fournit la liste complète des expressions CLEM.

Calcul à partir de plusieurs champs

Le fait de passer en mode **Multiple** dans le noeud Calculer vous permet d'effectuer un calcul à partir de plusieurs champs reposant sur la même condition et appartenant à un même noeud. Cette fonction vous permet de gagner du temps lorsque vous voulez appliquer des transformations identiques à plusieurs champs de votre jeu de données. Par exemple, si vous voulez créer un modèle de régression permettant de calculer le salaire actuel en fonction du salaire de départ et de l'expérience professionnelle, il peut s'avérer utile d'appliquer une transformation logarithmique à ces trois variables. Plutôt que d'ajouter un nouveau noeud Calculer pour chaque transformation, vous pouvez appliquer simultanément la même

fonction à tous les champs. Il vous suffit de sélectionner tous les champs à partir desquels vous voulez calculer un nouveau champ, puis de saisir la formule de calcul en utilisant la fonction `@FIELD` dans les parenthèses des champs.

Remarque : la fonction `@FIELD` est très pratique pour effectuer un calcul à partir de plusieurs champs en même temps. Elle vous permet de faire référence au contenu des champs actuels, sans spécifier leur nom exact. Par exemple, l'expression CLEM utilisée pour appliquer une transformation logarithmique à plusieurs champs est `log(@FIELD)`.

Les options suivantes apparaissent dans la boîte de dialogue lorsque vous sélectionnez le mode **Multiple** :

Calculer à partir de. Utilisez le sélecteur de champs pour sélectionner les champs à partir desquels calculer de nouveaux champs. Un champ de sortie est généré pour chaque champ sélectionné. *Remarque* : Les champs sélectionnés n'ont pas besoin d'avoir le même type de stockage. Néanmoins, l'opération de calcul échouera si la condition n'est pas valide pour *tous* les champs.

Extension nom de champ. Entrez l'extension à ajouter aux nouveaux noms de champ. Par exemple, vous pouvez ajouter au nom du nouveau champ contenant le logarithme du champ *Salaire actuel* l'extension *log_* et donc l'intituler *log_Salaire actuel*. A l'aide des cases d'option, spécifiez si l'extension doit être ajoutée au nom du champ en tant que préfixe (au début) ou en tant que suffixe (à la fin). Le nom par défaut est Derive *N*, où *N* correspond au nombre de noeuds Calculer créés jusqu'ici dans la session actuelle.

Comme pour les noeuds Calculer en mode Simple, vous devez créer l'expression qui sera utilisée pour calculer le nouveau champ. En fonction du type de l'opération de calcul sélectionnée, vous disposez de plusieurs options pour créer une condition. Ces options sont traitées dans des rubriques ultérieures. Pour créer une expression, vous pouvez simplement la saisir dans les champs de formule ou utiliser le Générateur de formules en cliquant sur le bouton représentant une calculatrice. N'oubliez pas d'utiliser la fonction `@FIELD` lorsque les manipulations portent sur plusieurs champs.

Sélection de plusieurs champs

Pour tous les noeuds qui effectuent des opérations sur plusieurs champs d'entrée, tels que les noeuds Calculer (mode Multiple), Agréger, Trier, Courbes et Tracé horaire, vous pouvez facilement sélectionner plusieurs champs à l'aide de la boîte de dialogue Sélection des champs.

Trier par. Vous pouvez trier les champs disponibles à afficher en sélectionnant l'une des options suivantes :

- **Naturel.** Permet d'afficher l'ordre dans lequel les champs ont été transmis via le flux de données dans le noeud actuel.
- **Nom.** Permet de trier les champs à afficher dans l'ordre alphabétique..
- **Type.** Permet d'afficher les champs triés par niveau de mesure. Cette option est utile lors de la sélection de champs avec un niveau de mesure particulier.

Sélectionnez les champs un par un dans la liste, ou utilisez la méthode de sélection Maj+clic ou Ctrl+clic pour sélectionner plusieurs champs. Vous pouvez également utiliser les boutons situés en dessous de la liste pour sélectionner des groupes de champs en fonction de leur niveau de mesure, ou pour sélectionner ou désélectionner tous les champs du tableau.

Paramétrage des options du noeud de formule Calculer

Les noeuds de formule Calculer créent un champ pour chaque enregistrement dans un jeu de données en fonction des résultats d'une expression CLEM. Cette expression ne peut pas être conditionnelle. Pour calculer des valeurs en fonction d'une expression conditionnelle, utilisez le type indicateur ou conditionnel du noeud Calculer.

Formule Spécifiez une formule dans le langage CLEM afin de calculer une valeur pour le nouveau champ.

Remarque : Etant donné que SPSS Modeler ne peut pas savoir quel sous-niveau de mesure doit être utilisé pour un champ de liste calculé, pour les niveaux de mesure Collection et Géospatial, vous pouvez cliquer sur **Spécifier...** afin d'ouvrir la boîte de dialogue Valeurs et de définir le sous-niveau de mesure requis. Pour plus d'informations, voir «Définition de valeurs de liste calculées».

Pour les champs géospatiaux, les seuls types de stockage pertinents sont **Réel** et **Entier** (le paramètre par défaut est **Réel**).

Définition de valeurs de liste calculées

La boîte de dialogue Valeurs s'affiche lorsque vous sélectionnez **Spécifier...** dans la liste déroulante **Type de champ** de la formule du noeud Calculer. Dans cette boîte de dialogue, vous définissez les valeurs des sous-niveaux de mesure à utiliser pour les niveaux de mesure **Type de champ** de formule Collection ou Géospatial.

Mesure Sélectionnez **Collection** ou **Géospatial**. Si vous sélectionnez un autre niveau de mesure, la boîte de dialogue affiche un message indiquant que les valeurs ne sont pas modifiables.

Résumé

Le seul élément que vous pouvez définir pour le niveau **Mesure** de collection est **Mesure de liste**. Par défaut, cette mesure a pour valeur Sans type, mais vous pouvez sélectionner une autre valeur pour définir le niveau de mesure des éléments dans la liste. Vous pouvez choisir l'une des options suivantes :

- Sans type
- Catégorielle
- Continu
- Nominal
- Ordinal
- Indicateur

Géospatial

Pour le niveau de **mesure** Géospatial, vous pouvez sélectionner les options suivantes afin de définir le niveau de mesure des éléments dans la liste :

Type Sélectionnez le sous-niveau de mesure du champ géospatial. Les sous-niveaux disponibles sont déterminés par la profondeur de la zone de liste. Les valeurs par défaut sont les suivantes :

- Point (pas de profondeur)
- Chaîne (profondeur de un)
- Polygone (profondeur de un)
- Multipoint (profondeur de un)
- Multichaîne (profondeur de deux)
- Multipolygone (profondeur de deux)

Pour plus d'informations sur les sous-niveaux, voir la rubrique Sous-niveaux de mesure géospatiaux dans la section Noeud Typer du guide Noeuds SPSS Modeler source, d'exécution et de sortie.

Pour plus d'informations sur les profondeurs de liste, voir la rubrique Stockage de liste et niveaux de mesure associés dans la section Noeud source du guide Noeuds SPSS Modeler source, d'exécution et de sortie.

Système de coordonnées Cette option est disponible uniquement si vous avez changé le niveau de mesure en remplaçant un niveau non-géospatial par un niveau géospatial. Pour appliquer un système de coordonnées à vos données géospatiales, sélectionnez cette case à cocher. Par défaut, le système de coordonnées défini dans le panneau **Outils** > **Propriétés du flux** > **Options** > **Géospatial** est affiché. Pour utiliser des systèmes de coordonnées différents, cliquez sur le bouton **Changer** afin d'ouvrir la boîte de dialogue Sélection d'un système de coordonnées et choisissez le système qui correspond à vos données.

Pour plus d'informations sur les systèmes de coordonnées, voir la rubrique sur la définition des options géospatiales pour les flux de la section Utilisation des flux dans le guide d'utilisation de SPSS Modeler.

Calcul d'une liste ou d'un champ géospatial

Parfois, des données devant être enregistrées sous forme d'éléments de liste sont importées dans SPSS Modeler avec des attributs incorrects, par exemple sous forme de champs géospatiaux distincts, comme une coordonnée x et une coordonnée y, ou en tant que longitude et latitude, sous forme de lignes individuelles dans un fichier `.csv`. Dans ce cas, vous devez combiner les champs individuels dans une zone de liste unique ; pour ce faire, vous pouvez utiliser le noeud Calculer.

Remarque : Vous devez savoir quel est le champ x (ou longitude) et quel est le champ y (ou latitude) lorsque vous combinez des données géospatiales. Vous devez combiner vos données pour que la zone de liste résultante présente les éléments dans l'ordre suivant : [x, y] ou [Longitude, Latitude], qui sont les formats standard pour les coordonnées géospatiales.

Les étapes ci-après illustrent un exemple simple de calcul d'une zone de liste.

1. Dans votre flux, connectez un noeud Calculer à votre noeud Source.
2. Dans l'onglet Paramètres du noeud Calculer, sélectionnez **Formule** dans la liste **Dériver en tant que**.
3. Dans **Type de champ**, sélectionnez **Collection** pour une liste non géospatiale ou **Géospatial**. Par défaut, SPSS Modeler utilise une approche "meilleure estimation" pour définir le détail de liste correct ; vous pouvez sélectionner **Spécifier...** pour ouvrir la boîte de dialogue Valeurs. Cette boîte de dialogue peut être utilisée pour une collection, pour entrer des informations supplémentaires sur les données de votre liste, et, pour les données géospatiales, elle peut être utilisée pour définir le type de données et spécifier le système de coordonnées des données.

   Remarque : Pour les données géospatiales, le système de coordonnées que vous spécifiez doit correspondre exactement au système de coordonnées des données. Si ce n'est pas le cas, la fonctionnalité des données géospatialles générera des résultats incorrects.
4. Dans le panneau **Formule**, entrez la formule permettant de combiner vos données au format de liste approprié. Vous pouvez aussi cliquer sur le bouton représentant une calculatrice pour ouvrir le générateur de formules.

   Un exemple simple de formule permettant de calculer une liste est [x, y], où x et y sont des champs distincts dans la source de données. Le nouveau champ calculé qui est créé est une liste dans laquelle la valeur pour chaque enregistrement correspond aux valeurs x et y concaténées pour cet enregistrement.

   Remarque : Les champs qui sont combinés dans une liste de cette façon doivent être associés au même type de stockage.

Pour plus d'informations sur les listes et les profondeurs de liste, voir «Stockage de liste et niveaux de mesure associés», à la page 12.

Définition des options du noeud de calcul Booléen

Les noeuds booléens Calculer permettent d'indiquer une condition spécifique, telle qu'une tension artérielle élevée ou l'inactivité d'un compte client. Un champ booléen est créé pour chaque enregistrement, et, lorsque la condition true (vrai) est satisfaite, la valeur booléenne correspondante est ajoutée dans le champ.

Valeur vraie (true). Indiquez la valeur à inclure dans le champ booléen pour les enregistrements qui respectent la condition spécifiée plus bas. La valeur par défaut est T.

Valeur fausse (false). Indiquez la valeur à inclure dans le champ indicateur pour les enregistrements qui ne respectent *pas* la condition spécifiée plus bas. La valeur par défaut est F.

Vrai quand. Indiquez une condition CLEM pour évaluer certaines valeurs de chaque enregistrement et attribuer à l'enregistrement la valeur true (vrai) ou false (faux) (définie plus haut). Remarque : la valeur true (vrai) est attribuée aux enregistrements dans le cas des valeurs numériques non fausses (false).

Remarque : pour renvoyer une chaîne vide, vous devez saisir des guillemets d'ouverture et de fermeture, sans rien entre les deux (""). Les chaînes vides sont souvent utilisées, par exemple, comme valeur false (faux) afin de permettre aux valeurs true (vrai) de ressortir plus clairement dans un tableau. De la même manière, ayez recours aux guillemets pour utiliser une valeur de chaîne qui serait traitée en tant que nombre autrement.

Exemple

Dans les versions d'IBM SPSS Modeler antérieures à 12.0, des réponses multiples ont été importées dans un champ unique, avec des valeurs séparées par des virgules. Par exemple :

```
musée_du_design,institut_des_textiles_et_de_la_mode
musée_du_design
musée_archéologique
$null$
musée_des_beaux_arts,musée_des_sciences,autre
```

Afin de préparer ces données pour l'analyse, vous pouvez utiliser la fonction `hassubstring` afin de générer un champ booléen distinct pour chaque réponse ; entrez pour cela une expression semblable à ce qui suit :

```
hassubstring(museums,"museum_of_design")
```

Définition des options du noeud de calcul Nominal

Les noeuds d'ensemble Nominal sont utilisés pour exécuter un ensemble de conditions CLEM afin de déterminer la condition remplie par chaque enregistrement. Chaque fois qu'une condition est remplie pour un enregistrement, une valeur (indiquant l'ensemble de conditions rempli) est ajoutée au nouveau champ calculé.

Valeur par défaut. Indiquez la valeur à utiliser dans le nouveau champ si aucune condition n'est satisfaite.

Définir le champ sur. Indiquez la valeur à entrer dans le nouveau champ lorsqu'une condition particulière est satisfaite. Chaque valeur de la liste est associée à une condition que vous spécifiez dans la colonne adjacente.

Si cette condition est vraie. Indiquez une condition pour chaque membre du champ d'ensemble à répertorier. Utilisez le Générateur de formules pour faire votre choix parmi les fonctions et les champs disponibles. Vous pouvez utiliser les flèches et le bouton Supprimer pour réorganiser ou supprimer des conditions.

Une condition fonctionne en testant les valeurs d'un champ particulier du jeu de données. Au fur et à mesure que les conditions sont testées, les valeurs spécifiées plus haut sont assignées au nouveau champ pour indiquer les éventuelles conditions satisfaites. Si aucune condition n'est satisfaite, la valeur par défaut est utilisée.

Définition des options du noeud de calcul Etat

Les noeuds d'état Calculer sont semblables aux noeuds booléens Calculer. Un noeud Booléen définit des valeurs si une condition *unique* est satisfaite ou non pour l'enregistrement actuel. Le noeud d'état Calculer, quant à lui, peut modifier les valeurs d'un champ en fonction de sa réponse à *deux conditions indépendantes*. Autrement dit, la valeur est modifiée (activée ou désactivée) en fonction de la réponse à chaque condition.

Etat initial. Indiquez si vous souhaitez attribuer à chaque enregistrement du nouveau champ la valeur initiale **Activé** ou **Désactivé**. Cette valeur peut changer au fur et à mesure que les conditions sont respectées.

Valeur "Activé". Indiquez la valeur du nouveau champ si la condition Activé est vérifiée.

Activer quand.. Choisissez la condition CLEM qui détermine le passage à l'état activé lorsque la condition a la valeur true (vrai). Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Valeur "Désactivé". Indiquez la valeur du nouveau champ si la condition Désactivé est vérifiée.

Désactiver quand. Choisissez la condition CLEM qui détermine le passage à l'état désactivé lorsque la condition a la valeur false (faux). Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Remarque : pour spécifier une chaîne vide, vous devez saisir des guillemets d'ouverture et de fermeture, sans rien entre les deux (""). De la même manière, ayez recours aux guillemets pour utiliser une valeur de chaîne qui serait traitée en tant que nombre autrement.

Définition des options du noeud de calcul Comptage

Les noeuds de calcul Calculer sont utilisés pour appliquer une série de conditions aux valeurs d'un champ numérique du jeu de données. Au fur et à mesure que les conditions sont respectées, la valeur du champ Comptage calculé augmente en fonction de l'incrément spécifié. Ce type de noeud Calculer est pratique pour les séries temporelles.

Valeur initiale. Définit une valeur utilisée pour le nouveau champ lors de l'exécution. La valeur initiale doit être une constante numérique. Utilisez les flèches pour augmenter ou diminuer la valeur.

Incrémenter quand. Spécifiez la condition CLEM qui, lorsqu'elle est satisfaite, modifie la valeur calculée en se basant sur le chiffre indiqué dans Incrémenter par. Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Incrémenter par. Définissez la valeur de l'incrément. Vous pouvez utiliser une constante numérique ou le résultat d'une expression CLEM.

Restaurer quand. Indiquez la condition qui, lorsqu'elle est satisfaite, restaure la valeur calculée sur sa valeur initiale. Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Définition des options du noeud de calcul Conditionnel

Les noeuds conditionnels Calculer utilisent la série d'instructions If-Then-Else pour le calcul de la valeur du nouveau champ.

Si. Indiquez une condition CLEM qui sera évaluée pour chaque enregistrement lors de l'exécution. Si cette condition a la valeur true (vrai) (ou qu'elle n'est pas false (faux), dans le cas de valeurs numériques), le nouveau champ reçoit la valeur indiquée à côté de l'expression Donc ci-dessous. Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Donc. Indiquez la valeur ou l'expression CLEM à utiliser pour le nouveau champ si l'instruction Si ci-dessus a la valeur true (vrai) (ou non-false (non fausse)). Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Sinon. Indiquez la valeur ou l'expression CLEM à utiliser pour le nouveau champ si l'instruction Si ci-dessus a la valeur false (faux). Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Recodage des valeurs à l'aide du noeud Calculer

Les noeuds Calculer peuvent également être utilisés pour recoder des valeurs, en convertissant, par exemple, un champ de type chaîne comportant des valeurs catégorielles en champ nominal (ensemble) numérique.

1. Pour l'option Calculer en tant que, sélectionnez le type de champ (Nominal, Indicateur, etc.) approprié.
2. Indiquez les conditions de recodage des valeurs. Vous pouvez, par exemple, affecter la valeur 1 si `Drug='drugA'`, 2 si `Drug='drugB'`, et ainsi de suite.

---

noeud Remplacer

Les noeuds Remplacer sont utilisés pour remplacer les valeurs de champ et pour modifier le stockage. Vous pouvez décider de remplacer les valeurs reposant sur une condition CLEM spécifiée, telle que `@BLANK(FIELD)`. Vous pouvez également choisir de remplacer tous les blancs ou toutes les valeurs nulles par une valeur précise. Les noeuds Remplacer sont souvent utilisés avec le noeud Typer pour remplacer des valeurs manquantes. Par exemple, vous pouvez remplacer des blancs avec la valeur moyenne d'un champ en spécifiant une expression telle que `@GLOBAL_MEAN`. Cette expression remplace tous les blancs par la valeur moyenne calculée par le noeud Valeurs globales.

Renseigner les champs. A l'aide du sélecteur de champs (bouton situé à droite du champ de texte), sélectionnez les champs du jeu de données dont vous souhaitez analyser et remplacer les valeurs. Le comportement par défaut consiste à remplacer les valeurs en fonction des expressions Condition et Remplacer par spécifiées plus bas. Vous pouvez également choisir une autre méthode de remplacement à l'aide des options Remplacer ci-dessous.

Remarque : si vous sélectionnez plusieurs champs à remplacer par une valeur définie par l'utilisateur, il est important que les types de champ soient similaires (tous numériques ou tous symboliques).

Remplacer - Sélectionnez cette option pour remplacer les valeurs des champs sélectionnés à l'aide de l'une des méthodes suivantes :

- **Basé sur une condition.** Cette option active le champ Condition et le Générateur de formules pour vous permettre de créer une expression utilisée comme condition pour le remplacement par la valeur spécifiée.
- **Toujours.** Remplace toutes les valeurs du champ sélectionné. Par exemple, vous pouvez utiliser cette option pour convertir le stockage du revenu en une chaîne, grâce à l'expression CLEM suivante `(to_string(income))`.
- **Valeurs non renseignées.** Remplace toutes les valeurs vides spécifiées par l'utilisateur dans le champ sélectionné. La condition standard `@BLANK(@FIELD)` est utilisée pour sélectionner les blancs. *Remarque* : vous pouvez définir les blancs via l'onglet Types du noeud source ou via un noeud type.
- **Valeurs nulles.** Remplace toutes les valeurs système nulles dans le champ sélectionné. La condition standard `@NULL(@FIELD)` est utilisée pour sélectionner les valeurs nulles.
- **Valeurs nulles et non renseignées.** Remplace les valeurs non renseignées et les valeurs système nulles dans le champ sélectionné. Cette option est utile lorsque vous ne savez pas avec certitude si les valeurs nulles ont été définies comme valeurs manquantes.

Condition. Cette option est disponible lorsque vous avez sélectionné l'option **Basé sur une condition**. Cette zone de texte vous permet d'indiquer une expression CLEM pour l'évaluation des champs sélectionnés. Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Remplacer par. Indiquez une expression CLEM pour attribuer une nouvelle valeur aux champs sélectionnés. Vous pouvez également remplacer la valeur par une valeur nulle en entrant `undef` dans la zone de texte. Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules.

Remarque : Si les champs sélectionnés sont des chaînes, vous devez les remplacer par des valeurs de chaîne. Si vous utilisez la valeur par défaut 0 ou une autre valeur numérique en tant que valeur de remplacement pour les champs de type chaîne, une erreur est générée.

Notez que l'utilisation de l'une des tâches suivantes peut changer l'ordre des lignes :

- Exécution dans une base de données via la répercussion SQL
- Exécution via le serveur IBM SPSS Analytic Server distant
- Utilisation de fonctions exécutées sur le serveur IBM SPSS Analytic Server imbriqué
- Calcul d'une liste (pour un exemple, voir «Calcul d'une liste ou d'un champ géospatial», à la page 166)
- Appel de l'une des fonctions décrites dans Fonctions spatiales

Conversion du stockage à l'aide du noeud Remplacer

En utilisant la condition Remplacer d'un noeud Remplacer, vous pouvez facilement convertir le type de stockage d'un ou de plusieurs champs. Par exemple, la fonction de conversion `to_integer` permet de convertir le type chaîne d'un *revenu* en un type entier, grâce à l'expression CLEM suivante : `to_integer(income)`.

Vous pouvez afficher les fonctions de conversion disponibles et créer automatiquement une expression CLEM en utilisant le Générateur de formules. Dans la liste déroulante Fonctions, sélectionnez **Conversion** pour afficher la liste des fonctions de conversion de stockage. Les fonctions de conversion suivantes sont disponibles :

- `to_integer(ITEM)`
- `to_real(ITEM)`
- `to_number(ITEM)`
- `to_string(ITEM)`
- `to_time(ITEM)`
- `to_timestamp(ITEM)`
- `to_date(ITEM)`
- `to_datetime(ITEM)`

Conversion des valeurs date et heure. Notez que les fonctions de conversion (et toutes les autres fonctions qui nécessitent un type spécifique d'entrée, par exemple une valeur de date ou d'heure) dépendent des formats actuels indiqués dans la boîte de dialogue des options de flux. Par exemple, si vous souhaitez convertir un champ de type chaîne avec des valeurs *Jan 2003*, *Fév 2003*, etc., en stockage de date, sélectionnez **MOIS AAAA** comme format de date par défaut pour le flux.

Les fonctions de conversion sont également disponibles depuis le noeud Calculer pour la conversion temporaire lors d'un calcul. Vous pouvez également utiliser le noeud Calculer pour effectuer d'autres manipulations, telles que la modification du codage des champs de type chaîne contenant des valeurs catégorielles. Pour plus d'informations, voir «Recodage des valeurs à l'aide du noeud Calculer», à la page 169.

Noeud Recoder

Le noeud Recoder permet de transformer un ensemble de valeurs catégorielles en un autre. La recodification est utile pour réduire des catégories ou regrouper des données à analyser. Par exemple, vous pouvez recoder les valeurs du nom *Produit* en trois groupes, comme *Ustensiles de cuisine*, *Salle de bains et linge* et *Appareils ménagers*. Cette opération est généralement exécutée directement à partir d'un noeud Proportion par regroupement des valeurs et génération d'un noeud Recoder. Pour plus d'informations, voir «Utilisation d'un noeud Proportion», à la page 251.

La recodification peut s'effectuer pour un ou plusieurs champs symboliques. Vous pouvez également décider de remplacer les valeurs d'un champ existant par de nouvelles valeurs ou de générer un nouveau champ.

Quand faut-il utiliser un noeud Recoder ?

Avant d'utiliser un noeud Recoder, assurez-vous qu'aucun autre noeud d'opérations sur les champs n'est plus adéquat pour cette tâche :

- Pour transformer des intervalles numériques en ensembles à l'aide d'une méthode automatique, telle que celle des rangs ou des centiles, vous devez utiliser un noeud Discrétiser. Pour plus d'informations, voir «Noeud Discrétiser», à la page 175.
- Pour classifier manuellement des intervalles numériques en ensembles, vous devez utiliser un noeud Calculer. Par exemple, si vous souhaitez réduire des valeurs de salaire en catégories d'intervalle salarial, vous devez utiliser un noeud Calculer pour définir chaque catégorie manuellement.
- Pour créer un ou plusieurs champs booléens sur la base des valeurs d'un champ catégoriel, tel que *Type_hypothèque*, vous devez utiliser un noeud Binariser.
- Pour convertir un champ catégoriel en stockage numérique, vous pouvez utiliser un noeud Calculer. Vous pouvez, par exemple, convertir les valeurs *Non* et *Oui* respectivement en valeurs 0 et 1. Pour plus d'informations, voir «Recodage des valeurs à l'aide du noeud Calculer», à la page 169.

Paramétrage des options du noeud Recoder

L'utilisation du noeud Recoder se divise en trois étapes :

1. Tout d'abord, choisissez si vous souhaitez recodifier plusieurs champs ou un seul champ.
2. Ensuite, choisissez soit de recoder un champ existant, soit de créer un nouveau champ.
3. Enfin, utilisez les options dynamiques de la boîte de dialogue du noeud Recoder pour mapper les ensembles comme vous le souhaitez.

   Mode. Sélectionnez **Simple** pour recodifier les catégories d'un champ. Sélectionnez **Multiple** pour activer les options permettant de transformer plusieurs champs simultanément.

   Recoder dans. Sélectionnez **Nouveau champ** pour conserver le champ nominal d'origine et calculer un autre champ contenant les valeurs recodifiées. Sélectionnez **Champ existant** pour écraser les valeurs du champ d'origine et les remplacer par les nouvelles classifications. Il s'agit avant tout d'une opération de type remplacement.

   Lorsque vous avez indiqué le mode et les options de remplacement, vous devez sélectionner le champ de transformation et indiquer les nouvelles valeurs de classification à l'aide des options dynamiques situées dans la moitié inférieure de la boîte de dialogue. Ces options varient en fonction du mode sélectionné plus haut.

   Recoder les champs. Utilisez le sélecteur de champs à droite pour sélectionner un (mode Simple) ou plusieurs (mode Multiple) champs catégoriels.

   Nouveau nom de champ. Indiquez le nom du nouveau champ nominal contenant les valeurs recodées. Cette option n'est disponible qu'en mode Simple lorsque l'option **Nouveau champ** plus haut est sélectionnée. Lorsque l'option **Champ existant** est sélectionnée, le nom du champ d'origine est conservé. Lorsque vous travaillez en mode Multiple, cette option est remplacée par des contrôles

permettant de spécifier une extension ajoutée à chaque nouveau champ. Pour plus d'informations, voir «Recodification de plusieurs champs».

Recoder les valeurs. Ce tableau établit un mappage clair entre les anciennes valeurs d'ensemble et celles que vous indiquez ici.

- **Valeur d'origine.** Cette colonne répertorie les valeurs existantes des champs sélectionnés.
- **Nouvelle valeur.** Utilisez cette colonne pour saisir les nouvelles valeurs de catégorie ou sélectionnez-en une dans la liste déroulante. Lorsque vous générez automatiquement un noeud Recoder avec les valeurs provenant d'un graphique Proportion, ces valeurs sont incluses dans la liste déroulante. Cela vous permet de mapper les valeurs existantes rapidement avec un ensemble de valeurs connu. Par exemple, les organisations de santé regroupent parfois les diagnostics différemment selon le réseau et les paramètres régionaux. Après une fusion ou un rachat, toutes les parties doivent recodifier les données nouvelles ou même existantes de manière homogène. Au lieu d'attribuer un type manuellement à chaque cible à partir d'une longue liste, vous pouvez lire la principale liste des valeurs dans IBM SPSS Modeler, exécuter un graphique Proportion pour le champ *Diagnostic*, et générer un noeud Recoder (valeurs) pour ce champ directement à partir du graphique. Ce processus rend toutes les valeurs Diagnostic cible disponibles à partir de la liste déroulante Nouvelles valeurs.

4. Cliquez sur **Obtenir** pour lire les valeurs d'origine d'un ou de plusieurs des champs sélectionnés plus haut.
5. Cliquez sur **Copier** pour coller les valeurs d'origine dans la colonne *Nouvelle valeur* pour les champs qui n'ont pas encore été mappés. Les valeurs d'origine non mappées sont ajoutées à la liste déroulante.
6. Cliquez sur **Tout effacer** pour effacer toutes les spécifications de la colonne *Nouvelle valeur*. *Remarque* : cette option n'efface pas les valeurs de la liste déroulante..
7. Cliquez sur **Auto** pour générer automatiquement des entiers consécutifs pour chacune des valeurs d'origine. Seules les valeurs entières peuvent être générées (les valeurs réelles telles que 1,5 ou 2,5 ne peuvent pas l'être).

Par exemple, vous pouvez générer automatiquement des numéros d'ID consécutifs pour des noms de produit ou des numéros pour des cours proposés par une université. Cette fonctionnalité correspond à la recodification automatique des ensembles de IBM SPSS Statistics.

A utiliser pour les valeurs non spécifiées. Cette option est utilisée pour remplacer les valeurs non spécifiées dans le nouveau champ. Vous pouvez choisir de conserver la valeur d'origine en sélectionnant **Valeur d'origine** ou d'indiquer une valeur par défaut.

Recodification de plusieurs champs

Pour mapper simultanément les valeurs de catégorie de plusieurs champs, paramétrez le mode sur **Multiple**. Les nouveaux paramètres décrits ci-dessous sont alors activés dans la boîte de dialogue Recoder.

Recoder les champs. Utilisez le sélecteur de champs situé à droite pour sélectionner les champs à transformer. Le sélecteur de champs permet de sélectionner tous les champs à la fois ou des champs de même type, par exemple des champs nominaux ou indicateurs.

Extension nom de champ. Lorsque vous recodez plusieurs champs simultanément, il est plus efficace d'indiquer une extension commune à ajouter à tous les nouveaux champs plutôt qu'un nom différent pour chaque champ. Indiquez une extension telle que `_recode`, et précisez si cette extension doit figurer au début ou à la fin des noms de champ d'origine.

Stockage et niveau de mesure des champs recodifiés

Le noeud Recoder crée toujours un champ de type nominal à partir de l'opération de recodification. Ceci peut entraîner, dans certains cas, la modification du niveau de mesure de champ si vous utilisez le mode de recodification **Champ existant**.

Le stockage du nouveau champ (mode de *stockage* des données et non *utilisation*) est calculé sur la base des options suivantes de l'onglet Paramètres :

- Si les valeurs non spécifiées sont paramétrées pour utiliser une valeur par défaut, le type de stockage approprié est déterminé par l'examen des nouvelles valeurs et par celui de la valeur par défaut. Par exemple, si toutes les valeurs sont analysées comme des entiers, le champ aura le type de stockage Entier.
- Si les valeurs non spécifiées sont paramétrées pour utiliser les valeurs d'origine, le type de stockage est déterminé par celui du champ d'origine. Si toutes les valeurs sont analysées comme disposant du stockage du champ d'origine, ce stockage est conservé ; sinon, il est déterminé par la recherche du type de stockage le plus adéquat, pour les anciennes comme pour les nouvelles valeurs. Par exemple, la recodification 4 => 0, 5 => 0 de l'ensemble d'entiers { 1, 2, 3, 4, 5 } génère un nouvel ensemble d'entiers { 1, 2, 3, 0 }, tandis que la recodification 4 => valeurs supérieures à 3, 5 => valeurs supérieures à 3 génère la chaîne { 1, 2, 3, "valeurs "supérieures à 3" génère la chaîne { “1”, “2”, “3”, “valeurs "supérieures à 3” }.

Remarque : si le type d'origine n'était pas instancié, le nouveau type ne l'est pas non plus.

Noeud Anonymiser

Le noeud Anonymiser permet de masquer les noms de champ, les valeurs de champ ou les deux types de données lorsque vous travaillez avec des données à inclure dans un modèle situé en aval du noeud. De cette façon, vous pouvez distribuer librement le modèle généré (par exemple à l'assistance technique) sans craindre que des utilisateurs non autorisés aient la possibilité de visualiser des données confidentielles telles que les fichiers du personnel ou les dossiers médicaux de patients.

Selon l'endroit où vous placez le noeud Anonymiser dans le flux, il se peut que vous deviez apporter des changements à d'autres noeuds. Par exemple, si vous insérez un noeud Anonymiser en amont d'un noeud Sélectionner, les critères de sélection de ce dernier doivent être modifiés s'ils agissent sur des valeurs qui sont désormais anonymisées.

La méthode à utiliser pour l'anonymisation dépend de plusieurs facteurs. Pour les noms de champ, ainsi que pour toutes les valeurs de champ excepté les niveaux de mesure continus, les données sont remplacées par une chaîne du type :

*préfixe*_S*n*

où *prefix_* est une chaîne définie par l'utilisateur ou la chaîne par défaut `anon_` et *fieldcounter* est une valeur entière qui commence à 0 et qui est incrémentée pour chaque valeur unique (par exemple, `anon_S0`, `anon_S1`, etc.).

Les valeurs de champ du type Continu doivent être transformées car les intervalles numériques se rapportent à des valeurs entières ou réelles plutôt qu'à des chaînes. En tant que telles, elles peuvent être anonymisées uniquement par la transformation de l'intervalle en un intervalle différent. Les données d'origine sont ainsi masquées. La transformation d'une valeur *x* de l'intervalle est exécutée de la façon suivante :

`A*(`*x* `+ B)`

où :

A est un facteur d'échelle, obligatoirement supérieur à 0.

B est un décalage de translation à ajouter aux valeurs.

Exemple

Soit un champ *AGE* avec le facteur d'échelle *A* défini sur 7 et le décalage de translation *B* défini sur 3, les valeurs relatives à *AGE* sont transformées de la façon suivante :

```
7*(AGE + 3)
```

Définition des options du noeud Anonymiser

Ici, vous pouvez choisir les champs qui auront leurs valeurs masquées plus en aval.

Les champs de données doivent être instanciés en amont du noeud Anonymiser pour que les opérations d'anonymisation puissent être exécutées. Pour instancier les données, cliquez sur le bouton **Lire les valeurs** d'un noeud Typer ou sur l'onglet Types d'un noeud source.

Champ. Répertorie les champs du jeu de données actuel. Si des noms de champ ont déjà été anonymisés, ils apparaissent ici.

Mesure. Niveau de mesure du champ.

Anonymiser des valeurs. Sélectionnez un ou plusieurs champs, cliquez sur cette colonne et choisissez **Oui** pour anonymiser la valeur de champ à l'aide du préfixe par défaut **anon_** ; choisissez **Spécifier** pour afficher une boîte de dialogue qui permet de saisir votre propre préfixe ou, dans le cas de valeurs de champ de type *Continu*, indiquez si la transformation des valeurs de champ doit utiliser des valeurs aléatoires ou définies par l'utilisateur. Il n'est pas possible de spécifier au cours de la même opération des types de champ *Continu* et non-*Continu* ; vous devez spécifier chaque type séparément.

Afficher les champs actuels. Sélectionnez cette option pour afficher les champs des jeux de données connectés au noeud Anonymiser. Par défaut, cette option est sélectionnée.

Afficher les paramètres de champ non utilisés. Sélectionnez cette option pour afficher les champs des jeux de données qui étaient auparavant connectés au noeud Filtrer. Cette option est utile lorsque vous copiez des noeuds d'un flux à un autre, ou lorsque vous enregistrez ou rechargez des noeuds.

Spécification des modalités d'anonymisation des valeurs de champ

La boîte de dialogue Remplacer les valeurs permet de choisir entre l'utilisation du préfixe par défaut pour les valeurs de champ anonymisées et l'utilisation d'un préfixe personnalisé. Lorsque vous cliquez sur **OK** dans cette boîte de dialogue, le paramètre de l'option Anonymiser des valeurs de l'onglet Paramètres devient **Oui** pour le ou les champs sélectionnés.

Préfixe des valeurs de champ. Le préfixe par défaut pour les valeurs de champ anonymisées est **anon_** ; sélectionnez **Personnalisé** et entrez, si vous le souhaitez, votre propre préfixe.

La boîte de dialogue Transformer les valeurs apparaît uniquement pour les champs du type Continu ; elle permet de spécifier si la transformation des valeurs de champ doit utiliser des valeurs aléatoires ou définies par l'utilisateur.

Aléatoire. Sélectionnez cette option afin d'utiliser des valeurs aléatoires pour la transformation. L'option **Définir une valeur de départ aléatoire** est sélectionnée par défaut. Spécifiez une valeur dans le champ **Valeur de départ** ou utilisez la valeur par défaut.

Colonne fixe. Sélectionnez cette option afin de définir vos propres valeurs pour la transformation.

- **Mise à l'échelle.** Nombre par lequel les valeurs de champ sont multipliées dans la transformation. La valeur minimale est 1. La valeur maximale est normalement de 10, mais elle peut être diminuée pour éviter tout dépassement.

- **Traduire par.** nombre qui sera ajouté aux valeurs de champ dans la transformation. La valeur minimale est 0. La valeur maximale est normalement de 1 000, mais elle peut être diminuée pour éviter tout dépassement.

Anonymisation des valeurs de champ

Les valeurs des champs sélectionnés pour l'anonymisation dans l'onglet Paramètres sont anonymisées :

- lorsque vous exécutez le flux contenant le noeud Anonymiser ;
- lorsque vous prévisualisez les valeurs.

Pour prévisualiser les valeurs, cliquez sur le bouton **Anonymiser des valeurs** dans l'onglet Valeurs anonymisées. Sélectionnez ensuite un nom de champ dans la liste déroulante.

Si le niveau de mesure est de type Continu, les éléments suivants s'affichent :

- valeurs minimale et maximale de l'intervalle d'origine
- équation utilisée pour transformer les valeurs

Si le niveau de mesure est différent de la valeur Continue, l'écran affiche la valeur d'origine et la valeur anonymisée pour ce champ.

Un affichage sur fond jaune indique que le paramètre du champ sélectionné a changé depuis la dernière anonymisation des valeurs ou que des changements ont été apportés aux données situées en amont du noeud Anonymiser, de sorte que les valeurs anonymisées ne sont peut-être plus correctes. L'ensemble actuel de valeurs apparaît ; cliquez de nouveau sur le bouton **Anonymiser des valeurs** pour générer un nouvel ensemble de valeurs conforme au paramètre actuel.

Anonymiser des valeurs. Crée des valeurs anonymisées pour le champ sélectionné et les affiche dans le tableau. Si vous utilisez une valeur de départ aléatoire pour un champ de type Continu, le fait de cliquer sur ce bouton à plusieurs reprises crée un ensemble de valeurs différent à chaque fois.

Effacer les valeurs. Efface les valeurs d'origine et les valeurs anonymisées du tableau.

Noeud Discrétiser

Le noeud Discrétiser permet de créer automatiquement de nouveaux champs nominaux sur la base des valeurs d'un ou de plusieurs champs continus numériques existants (intervalle numérique). Par exemple, vous pouvez transformer un champ continu de revenus en un nouveau champ catégoriel contenant des groupes de revenus de largeur égale ou comme écarts par rapport à la moyenne. Vous pouvez également sélectionner un champ de superviseur catégoriel afin de conserver la force de l'association d'origine entre deux champs.

La création d'intervalles peut s'avérer utile pour un certain nombre de raisons, notamment :

- **Matrice de diagramme de dispersion.** Certains algorithmes, Naive Bayes ou la régression logistique par exemple, ont besoin d'entrées catégorielles.
- **Performances.** Les algorithmes comme la logistique multinomiale peuvent obtenir de meilleures performances si le nombre de valeurs distinctes des champs d'entrée est réduit. Utilisez par exemple la valeur médiane ou moyenne pour chaque noeud plutôt que la valeur d'origine.
- **Confidentialité des données.** Pour les informations personnelles et confidentielles, par exemple les salaires, vous pouvez indiquer des intervalles plutôt que les chiffres exacts afin de protéger la confidentialité.

Un certain nombre de méthode de regroupements par casiers sont disponibles. Une fois les casiers du nouveau champ créés, vous pouvez générer un noeud Calculer à partir des points de césure.

Quand faut-il utiliser un noeud Discrétiser ?

Avant d'utiliser un noeud Discrétiser, assurez-vous qu'aucune autre technique n'est plus adéquate pour cette tâche :

- Pour indiquer manuellement les points de césure des catégories, telles que des intervalles salariaux prédéfinis, utilisez un noeud Calculer. Pour plus d'informations, voir «Noeud Calculer», à la page 162.
- Pour créer de nouvelles catégories pour des ensembles existants, utilisez un noeud Recoder. Pour plus d'informations, voir «Noeud Recoder», à la page 171.

Traitement des valeurs manquantes

Le noeud Discrétiser traite les valeurs manquantes de l'une des manières suivantes :

- **Blancs définis par l'utilisateur.** Les valeurs manquantes définies comme des blancs sont incluses dans la transformation. Par exemple, si vous avez indiqué -99 pour indiquer une valeur non renseignée à l'aide du noeud type, cette valeur sera incluse dans la création des casiers. Pour ignorer les blancs au cours de la création des casiers, utilisez un noeud Remplacer pour remplacer les valeurs non renseignées par la valeur système nulle.
- **Valeurs manquantes système ($null$).** Les valeurs nulles sont ignorées lors de la transformation des casiers. Elles restent nulles après la transformation.

L'onglet Paramètres propose les options des différentes techniques. L'onglet Affichage affiche les points de césure établis pour les données précédemment passées dans ce noeud.

Définition des options du noeud Discrétiser

Le noeud Discrétiser permet de générer automatiquement des intervalles (catégories) à l'aide des techniques suivantes :

- Création d'intervalles à largeur fixe
- Quantiles (effectifs égaux ou somme)
- Moyenne et écart-type
- Rangs
- Optimisé par rapport à un champ de superviseur catégoriel

La partie inférieure de la boîte de dialogue change de manière dynamique en fonction de la méthode de regroupement par casiers sélectionnée.

Champs de casier. Les champs continus (intervalle numérique) en attente de transformation sont affichés ici. Le noeud Discrétiser permet de créer des casiers pour plusieurs champs simultanément. Ajoutez ou supprimez des champs à l'aide des boutons sur la droite.

Méthode de regroupement par casiers. Sélectionnez la méthode utilisée pour déterminer les points de césure des nouveaux casiers de champ (catégories). Les rubriques suivantes décrivent les options disponibles dans chaque cas.

Seuils des casiers. Spécifiez comment sont calculés les seuils des casiers.

- **Toujours recalculer.** Les points de césure et les affectations de casier sont toujours recalculés lors de l'exécution du noeud.
- **Lire dans l'onglet Valeurs de casier si disponible.** Les points de césure et les attributions de casiers sont calculés uniquement en fonction des besoins (par exemple, quand de nouvelles données sont ajoutées).

Les rubriques suivantes présentent les options des méthodes de création de casiers disponibles.

Intervalles à largeur fixe

Lorsque vous choisissez la méthode de regroupement par casiers **Largeur fixe**, un nouvel ensemble d'options apparaît dans la boîte de dialogue.

Extension du nom. Spécifiez l'extension à utiliser pour les champs générés. L'extension par défaut est *_BIN*. Vous pouvez également indiquer si l'extension doit être ajoutée au début (**Préfixe**) ou à la fin (**Suffixe**) du nom de champ. Par exemple, vous pouvez générer un nouveau champ intitulé *revenu_BIN*.

Largeur de casier. Spécifiez la valeur (entier ou réel) utilisée pour le calcul de la "largeur" du casier. Par exemple, vous pouvez utiliser la valeur par défaut, 10, pour créer les casiers du champ *Age*. Le champ *Age* couvrant l'intervalle 18–65, les casiers générés sont les suivants :

Tableau 24. Casiers du champ Age qui couvre l'intervalle 18–65

Casier 1	Casier 2	Casier 3	Casier 4	Casier 5	Casier 6
>=13 à <23	>=23 à <33	>=33 à <43	>=43 à <53	>=53 à <63	>=63 à <73

Le début des intervalles de casier est calculé de la manière suivante : plus valeur faible analysée moins la moitié de la largeur de l'intervalle de casier indiqué. Par exemple, dans les casiers ci-dessus, la valeur 13, qui correspond au début des intervalles, a été obtenue grâce au calcul suivant : 18 *[valeur de données la plus faible]* – 5 *[0.5 × (largeur du casier 10)]* = 13.

Nbre de casiers. Utilisez cette option pour indiquer un entier déterminant le nombre de casiers de largeur fixe (catégories) des nouveaux champs.

Lorsque vous avez exécuté le noeud Discrétiser dans le cadre d'un flux, vous pouvez afficher les seuils de casier générés en cliquant sur l'onglet **Aperçu** de la boîte de dialogue du noeud Discrétiser. Pour plus d'informations, voir «Prévisualisation des intervalles générés», à la page 181.

Quantiles (effectifs égaux ou somme)

La méthode de regroupement par casiers de type quantile génère des champs nominaux qui peuvent être utilisés pour scinder des enregistrements analysés en groupes de type centiles (ou quartiles, déciles, etc.), de sorte que chaque groupe contienne le même nombre d'enregistrements, ou que la somme des valeurs de chaque groupe soit égale. Les enregistrements sont classés dans l'ordre croissant de la valeur du champ d'intervalle indiqué ; les enregistrements présentant les valeurs les moins élevées pour la variable d'intervalle sélectionnée se voient ainsi attribuer le rang 1, l'ensemble d'enregistrements suivant le rang 2, et ainsi de suite. Les valeurs de seuil de chaque intervalle sont générées automatiquement en fonction des données et de la méthode des quantiles utilisée.

Extension du nom du quantile. Spécifiez l'extension utilisée pour les champs générés à l'aide de centiles standard. L'extension par défaut est *_TILE* plus *N*, *N* étant le numéro du quantile. Vous pouvez également indiquer si l'extension doit être ajoutée au début (**Préfixe**) ou à la fin (**Suffixe**) du nom de champ. Par exemple, vous pouvez générer un nouveau champ intitulé *revenu_TILE4*.

Extension personnalisée du nombre de quantiles. Spécifiez l'extension utilisée pour un intervalle de type quantile personnalisé. La valeur par défaut est *_TILEN*. Dans ce cas, *N* n'est *pas* remplacé par le nombre personnalisé.

Les centiles disponibles sont les suivants :

- **Quartile.** Génère 4 casiers, chacun contenant 25% des observations.
- **Quintile.** Génère 5 casiers, chacun contenant 20 % des observations.
- **Décile.** Génère 10 casiers, chacun contenant 10 % des observations.
- **Vingtile.** Génère 20 casiers, chacun contenant 5% des observations.
- **Percentile.** Génère 100 intervalles, chacun contenant 1 % des observations.

- **N personnalisé.** Sélectionnez cette option pour indiquer le nombre d'intervalles. Par exemple, une valeur de 3 produirait 3 catégories (deux points de césure), chacune contenant 33,3 % des observations.

Si les données contiennent moins de valeurs discrètes que le nombre de quantiles indiqué, tous les quantiles ne sont pas utilisés. La nouvelle proportion peut alors refléter la proportion d'origine des données.

Méthode des quantiles. Indique la méthode utilisée pour affecter des enregistrements à des intervalles.

- **Nombre d'enregistrements.** Cherche à attribuer un nombre égal d'enregistrements à chaque intervalle.
- **Somme des valeurs.** Cherche à attribuer des enregistrements à des intervalles de sorte que la somme des valeurs de chaque intervalle soit égale. Lorsque vous vous intéressez aux efforts de ventes par exemple, cette méthode peut être utilisée pour attribuer des prospects à des groupes de type décile en fonction de la valeur par enregistrement (les prospects qui présentent les valeurs les plus élevées étant placés dans l'intervalle supérieur). Par exemple, une entreprise pharmaceutique peut classer les médecins en groupes de type décile en fonction du nombre d'ordonnances qu'ils rédigent. Alors que chaque décile contient environ le même nombre d'ordonnances, le nombre de personnes à l'origine de ces ordonnances est différent (les personnes qui écrivent le plus d'ordonnances étant regroupées dans le décile 10). Cette approche suppose que toutes les valeurs soient supérieures à zéro ; si tel n'est pas le cas, elle risque de renvoyer des résultats inattendus.

Ex aequo. On parle de condition ex aequo lorsque des valeurs de part et d'autre d'un point de césure sont identiques. Par exemple, si vous utilisez des déciles, et que plus de 10 % des enregistrements présentent la même valeur pour le champ d'intervalle, ces enregistrements ne peuvent pas tous tenir dans le même intervalle sans forcer le seuil d'une façon ou d'une autre. Les valeurs ex aequo peuvent être déplacées vers le haut dans l'intervalle suivant ou conservées dans l'intervalle actuel, à condition qu'elles soient résolues de sorte que tous les enregistrements comportant des valeurs identiques se trouvent dans le même intervalle, et ce, même si cela génère un nombre d'enregistrements par intervalle plus important que prévu. Il est, pour cela, également possible d'ajuster les seuils des intervalles suivants ; les valeurs d'un même ensemble de nombres sont ainsi affectées différemment en fonction de la méthode utilisée pour résoudre les valeurs ex aequo.

- **Ajouter au suivant.** Sélectionnez cette option pour déplacer les valeurs ex aequo vers l'intervalle supérieur suivant.
- **Conserver dans l'élément actuel.** Conserve les valeurs ex aequo dans l'intervalle (inférieur) actuel. Cette méthode peut générer un nombre inférieur d'intervalles.
- **Attribuer de façon aléatoire.** Sélectionnez cette option pour attribuer les valeurs ex æquo de façon aléatoire à un intervalle. Ceci permet de conserver le nombre d'enregistrements dans chaque intervalle de façon égale.

Exemple : Création de quantiles en fonction du nombre d'enregistrements

Le tableau ci-dessous illustre la façon dont les valeurs de champ simplifiées sont classées en quartiles lors de la création de quantiles en fonction du nombre d'enregistrements. Les résultats varient en fonction de l'option de valeurs ex aequo sélectionnée.

Tableau 25. Exemple de création de quantiles en fonction du nombre d'enregistrements.

Valeurs	Ajouter au suivant	Conserver dans l'élément actuel
10	1	1
13	2	1
15	3	2
15	3	2
20	4	3

Le nombre d'éléments par intervalle est calculé de la façon suivante :

```
total number of value / number of tiles
```

Dans l'exemple simplifié ci-dessus, le nombre souhaité d'éléments par intervalle est de 1,25 (5 valeurs / 4 quartiles). La valeur 13 (valeur numéro 2) chevauche le seuil de comptage souhaité de 1,25 ; elle est par conséquent traitée différemment selon l'option d'ex aequo sélectionnée. En mode **Ajouter au suivant**, elle est ajoutée à l'intervalle 2. En mode **Conserver dans l'élément actuel**, elle reste dans l'intervalle 1, ce qui place l'intervalle des valeurs de l'intervalle 4 en dehors de l'intervalle des valeurs de données existantes. Par conséquent, seuls trois casier sont créés et les seuils de chaque casier sont ajustés en conséquence, comme illustré dans le tableau ci-après.

Tableau 26. Résultat de l'exemple de regroupement par casiers.

Intervalle	Inférieur	Supérieur
1	>=10	<15
2	>=15	<20
3	>=20	<=20

Remarque : L'activation du traitement parallèle peut augmenter la vitesse de création d'intervalles par quantiles.

Observations des rangs

Lorsque vous choisissez la méthode de création d'intervalles **Rangs**, un nouvel ensemble d'options apparaît dans la boîte de dialogue.

Le classement crée de nouveaux champs contenant des rangs, des rangs fractionnaires et des valeurs de centile pour les champs numériques, conformément aux options décrites ci-dessous.

Ordre des rangs. Sélectionnez **Croissant** (la valeur la plus faible est marquée 1) ou **Décroissant** (la valeur la plus élevée est marquée 1).

Rang. Sélectionnez cette option pour classer les observations dans l'ordre croissant ou décroissant, comme indiqué ci-avant. L'intervalle des valeurs du nouveau champ sera 1–*N*, *N* étant le nombre de valeurs discrètes présentes dans le champ d'origine. Les valeurs ex aequo reçoivent la moyenne de leur rang.

Rang fractionnaire. Sélectionnez cette option pour classer les observations dans lesquelles la valeur du nouveau champ équivaut au rang divisé par la somme des pondérations des observations non manquantes. Les rangs fractionnaires sont compris dans l'intervalle 0–1

Rang fractionnaire de pourcentage. Chaque rang est divisé par le nombre d'enregistrements avec valeurs valides et multiplié par 100. Les rangs fractionnaires de pourcentage sont compris dans l'intervalle 1-100..

Extension. Pour toutes les options de rang, vous pouvez créer des extensions personnalisées, et indiquer si l'extension doit être ajoutée au début (**Préfixe**) ou à la fin (**Suffixe**) du nom de champ. Par exemple, vous pouvez générer un nouveau champ intitulé *revenu_P_RANK*.

Moyenne/écart-type

Lorsque vous choisissez la méthode de création d'intervalles **Moyenne/écart-type**, un nouvel ensemble d'options apparaît dans la boîte de dialogue.

Cette méthode génère un ou plusieurs nouveaux champs avec catégories en fonction des valeurs de moyenne et d'écart-type de la proportion des champs spécifiés. Sélectionnez le nombre d'écarts à utiliser plus bas.

Extension du nom. Spécifiez l'extension à utiliser pour les champs générés. L'extension par défaut est *_SDBIN*. Vous pouvez également indiquer si l'extension doit être ajoutée au début (**Préfixe**) ou à la fin (**Suffixe**) du nom de champ. Par exemple, vous pouvez générer un nouveau champ intitulé *revenu_SDBIN*.

- **Ecart-type +/- 1.** Sélectionnez cette option pour générer trois intervalles.
- **Ecarts-types +/– 2.** Sélectionnez cette option pour générer cinq intervalles.
- **Ecarts-types +/– 3.** Sélectionnez cette option pour générer sept intervalles.

Par exemple, la sélection de l'option Ecart-type +/–1 génère améliorer les trois intervalles calculés dans le tableau ci-dessous.

Tableau 27. Exemple de casier d'écart-type.

Casier 1	Casier 2	Casier 3
x < (Moyenne – Ecart- type)	(Moyenne – Ecart- type) <= x <= (Moyenne + Ecart- type)	x > (Moyenne + Ecart-type. type)

Dans une proportion normale, 68 % des observations sont comprises dans un écart-type par rapport à la moyenne, 95 % dans deux écarts-types et 99 % dans 3 écarts-types. La création de catégories basées sur les écarts-types peut résulter en des intervalles définis en dehors de l'intervalle de données réel et même en dehors de l'intervalle des valeurs de données possibles (par exemple, un intervalle salarial négatif).

Création d'intervalles optimale

Si le champ dans lequel vous souhaitez créer des intervalles est fortement associé à un autre champ catégoriel, vous pouvez sélectionner ce dernier comme champ de superviseur afin de créer les intervalles de façon à préserver la force de l'association d'origine entre les deux champs.

Supposez par exemple que vous ayez utilisé l'analyse des clusters pour regrouper les Etats en fonction du taux de prêts immobiliers en souffrance, avec les taux les plus élevés dans le premier cluster. Dans ce cas, vous pouvez choisir *Pourcentage d'arriéré* et *Pourcentage de forclusion* comme champs d'intervalle, et le champ devant contenir les clusters d'appartenance généré par le modèle comme champ de superviseur.

Extension du nom Indiquez l'extension à utiliser pour les champs générés et déterminez si elle doit être ajoutée au début (**Préfixe**) ou à la fin (**Suffixe**) du nom du champ. Vous pouvez par exemple générer deux nouveaux champs nommés `arriéré_OPTIMAL` et `forclusion_OPTIMAL`.

Champ de superviseur Champ catégoriel utilisé pour construire les intervalles.

Préregrouper les champs pour améliorer les performances avec les jeux de données volumineux Indique s'il convient de procéder à un prétraitement pour simplifier la création d'intervalles optimale. Ceci permet de regrouper les valeurs d'échelle en un grand nombre d'intervalles en utilisant une méthode de création d'intervalles simple et non supervisée. Elle représente en outre les valeurs au sein de chaque intervalle par la moyenne et ajuste la pondération d'observation en conséquence avant de passer à la création d'intervalles supervisée. En pratique, cette méthode perd un certain degré de précision mais gagne en vitesse d'exécution. Elle est donc recommandée pour les grands jeux de données. Vous pouvez également indiquer le nombre maximal de casiers dans lesquels doit se trouver toute variable après le prétraitement une fois cette option utilisée.

Fusionner les casiers comportant relativement peu d'observations avec un voisinage plus large. Si cette option est activée, indique qu'un casier est fusionné si le rapport de sa taille (nombre d'observations) avec celle d'un casier voisin est inférieur au seuil spécifié. Notez que des seuils plus élevés sont susceptibles d'entraîner une fusion plus importante.

Paramètres de point de césure

La boîte de dialogue Paramètres de point de césure vous permet de choisir des options avancées pour l'algorithme de création d'intervalles optimale. Ces options indiquent à l'algorithme comment calculer les intervalles à l'aide du champ cible.

Points de fin de casier. Vous pouvez indiquer si les points de fin inférieur ou supérieur doivent être inclusifs (inférieur <= *x*) ou exclusifs (inférieur < *x*).

Premiers et derniers casiers. Pour le premier et le dernier casier, vous pouvez indiquer s'ils doivent être non délimités (tendant vers l'infini positif ou négatif) ou délimités par les points de données inférieur ou supérieur.

Prévisualisation des intervalles générés

L'onglet Valeurs d'intervalle du noeud Discrétiser permet de visualiser les seuils des intervalles générés. Avec le menu Générer, vous pouvez également générer un noeud Calculer qui peut être utilisé pour appliquer ces seuils d'un jeu de données à l'autre.

Champ regroupé en casiers. Dans la liste déroulante, sélectionnez le champ à afficher. A des fins de clarté, les noms de champ affichés reprennent le nom du champ d'origine.

Quantiles. Dans la liste déroulante, sélectionnez le quantile, tel que 10 ou 100, à afficher. Cette option est disponible uniquement lorsque les intervalles ont été générés à l'aide de la méthode des quantiles (effectifs égaux ou somme égale).

Seuils des casiers. Les valeurs de seuil sont affichées ici pour chaque intervalle généré, avec le nombre d'enregistrements qui correspondent à chaque intervalle. Pour la méthode de création d'intervalles optimale uniquement, le nombre d'enregistrements dans chaque intervalle est présenté comme un pourcentage du total. Il est impossible d'appliquer des seuils lorsque la méthode de création d'intervalles par rang est utilisée.

Lire les valeurs. Lit les valeurs mises en intervalles du jeu de données. Notez que les seuils sont également remplacés dès que de nouvelles données passent dans le flux.

Génération d'un noeud Calculer

Vous pouvez utiliser le menu Générer pour créer un noeud Calculer fondé sur les seuils actuels. Cela est utile lors de l'application de seuils d'intervalle établis d'un jeu de données à un autre. Par ailleurs, si les points de séparation sont connus, l'opération Calculer est plus efficace (c'est-à-dire plus rapide) que l'opération Discrétiser dans le cas des jeux de données volumineux.

Noeud Analyse RFM

Le noeud Analyse RFM (Récence, Effectif, Monétaire) permet de déterminer de façon quantitative les clients susceptibles d'être les meilleurs par l'étude de leur dernier achat (récence), l'effectif de leurs achats (effectif), et la somme dépensée lors de toutes les transactions (monétaire).

Le raisonnement derrière l'analyse RFM est que les clients qui achètent un produit ou un service une fois sont susceptibles de l'acheter à nouveau. Les données clients catégorisées se divisent en un certain nombre d'intervalles, avec les critères de création d'intervalles ajustés selon les besoins. Dans chacun des intervalles, un score est attribué aux clients. Ces scores sont ensuite combinés pour offrir un score RFM global. Ce score est une représentation de l'appartenance du client aux intervalles créés pour chacun des paramètres RFM. Ces données mises en intervalles peuvent s'avérer suffisantes pour vos besoins, par exemple, en identifiant les clients importants les plus fidèles. Elles peuvent être également transmises dans un flux pour une modélisation et une analyse plus approfondies.

Remarque : bien que la capacité à analyser et à classer les scores RFM est un outil pratique, vous devez cependant garder à l'esprit certains facteurs lors de son utilisation. Il peut être tentant de cibler les clients avec les meilleurs classements. Toutefois, une sur-sollicitation de ces clients peut conduire à un certain ressentiment et une baisse effective de l'activité commerciale continue. Cela vaut également la peine de garder à l'esprit que les clients avec des scores bas ne doivent pas être négligés mais plutôt encouragés pour qu'ils deviennent de meilleurs clients. Inversement, des scores élevés seuls ne reflètent pas forcément une bonne perspective de ventes, selon le marché. Par exemple, un client dans l'intervalle 5 pour la récence, indiquant qu'il a effectué des achats très récemment, peut ne pas être le meilleur client cible pour une personne vendant des produits coûteux plus durables tels que des voitures ou des télévisions.

Remarque : Selon le mode de stockage de vos données, vous devrez peut être faire précéder le noeud Analyse RFM par un noeud Agréger RFM pour transformer les données en un format utilisable. Par exemple, les données d'entrée doivent être au format client avec une seule ligne par client. Si les données des clients sont au format transactionnel, utilisez un noeud Agréger RFM en amont pour calculer les champs Récence, Effectif et Montant. Pour plus d'informations, voir «Noeud Agréger RFM», à la page 87.

Les noeuds Agréger RFM et Analyse RFM d'IBM SPSS Modeler sont configurés pour utiliser la création d'intervalles indépendants ; en d'autres termes, ils classent et espacent les données sur chaque mesure de valeur de proximité dans le temps, d'effectif et de valeur monétaire, sans tenir compte de leur valeur ni des deux autres mesures.

Paramètres du noeud Analyse RFM

Récence. A l'aide du sélecteur de champs (bouton à droite de la zone de texte), sélectionnez le champ Récence. Il peut s'agir d'une date, d'un horodatage ou d'un simple nombre. Remarque : lorsqu'une date ou un horodatage représente la date de la transaction la plus récente, la valeur la plus élevée est considérée comme étant la plus récente. Là où un nombre est indiqué, il représente le temps écoulé depuis la transaction la plus récente et la valeur la plus basse est considérée comme la plus récente.

Remarque : Si le noeud Analyse RFM est précédé dans le flux par un noeud Agréger RFM, les champs Récence, Effectif et Monétaire générés par le noeud Agréger RFM doivent être sélectionnés comme entrées dans le noeud Analyse RFM.

Fréquence. A l'aide du sélecteur de champs, sélectionnez le champ Effectif à utiliser.

Monétaire. A l'aide du sélecteur de champs, sélectionnez le champ Monétaire à utiliser.

Nombre de casiers. Pour chacun des trois types de sorties, sélectionnez le nombre de casiers à créer. La valeur par défaut est 5.

Remarque : Le nombre minimum de casiers est 2, et le maximum est 9.

Pondération. Par défaut, la plus haute importance lors du calcul des scores est accordée aux données de récence, suivies de l'effectif, puis du montant. Si besoin est, vous pouvez modifier la pondération affectant un ou plusieurs de ces éléments pour changer celui qui se voit accorder la plus haute importance.

Le score RFM est calculé comme suit : (Score de récence x pondération de récence) + (score de fréquence x pondération de fréquence) + (score du montant x pondération du montant).

Ex aequo. Indiquez la manière dont les scores (ex æquo) identiques doivent être mis en intervalles. Les options sont les suivantes :

- **Ajouter au suivant.** Sélectionnez cette option pour déplacer les valeurs ex aequo vers l'intervalle supérieur suivant.

- **Conserver dans l'élément actuel.** Conserve les valeurs ex aequo dans l'intervalle (inférieur) actuel. Cette méthode peut générer un nombre inférieur d'intervalles. (Il s'agit de la valeur par défaut).

Seuils des casiers. Indiquez si les scores de RFM et les affectations de casiers sont toujours recalculés lors de l'exécution du noeud, ou s'ils sont uniquement calculés selon les besoins (par exemple, lors de l'ajout de données). Si vous sélectionnez l'option **Lire à partir de l'onglet Valeurs de casier si disponible**, vous pouvez éditer les points de césures supérieurs et inférieurs pour les différents casiers dans l'onglet Valeurs de casier.

Une fois exécuté, le noeud Analyse RFM met en casiers les champs Récence, Effectif et Monétaire bruts et ajoute les champs suivants au jeu de données :

- Score de récence. Un classement (valeur d'intervalle) pour la récence
- Score de fréquence. Un classement (valeur d'intervalle) pour l'effectif
- Score monétaire. Un classement (valeur d'intervalle) pour Monétaire
- Score RFM. Le total pondéré des scores de récence, effectif et monétaire.

Ajouter les valeurs éloignées aux casiers de fin. Si vous cochez cette case, les enregistrements qui figurent au-dessous du casier inférieur sont ajoutés au casier inférieur, et ceux au-dessus du casier supérieur sont ajoutés au casier le plus grand sinon, une valeur nulle leur est attribuée. Cette case n'est disponible que si vous sélectionnez **Lire dans l'onglet Valeurs de casier si disponible**.

Mise en intervalle du noeud Analyse RFM

L'onglet Valeurs d'intervalle permet d'afficher, et dans certains cas, modifier les seuils des intervalles générés.

Remarque : vous ne pouvez modifier les valeurs dans cet onglet que si vous sélectionnez l'option **Lire dans l'onglet Valeurs d'intervalle si disponible** dans l'onglet Paramètres..

Champ regroupé en casiers. Dans la liste déroulante, sélectionnez un champ pour la séparation en intervalles. Les valeurs disponibles sont celles sélectionnées dans l'onglet Paramètres.

Tableau des valeurs de casier. Les valeurs de seuil de chaque casier généré sont affichées ici. Si vous sélectionnez l'option **Lire dans l'onglet Valeurs de casier si disponible** dans l'onglet Paramètres, vous pouvez modifier les points de césure pour chaque casier en double-cliquant sur la cellule pertinente.

Lire les valeurs. Lit les valeurs mises en casiers à partir du jeu de données et renseigne le tableau de valeurs de casier. Remarque : si vous sélectionnez **Toujours recalculer** dans l'onglet Paramètres, les seuils de casier seront écrasés lors de l'exécution des nouvelles données via le flux.

Noeud Ensemble

Le noeud Ensemble combine deux ou plusieurs nuggets de modèles pour obtenir des prévisions plus précises que celles acquises à partir des modèles individuels. En combinant les prévisions à partir de plusieurs modèles, il est possible d'éviter les limitations dans les modèles individuels. Ce qui entraîne une plus grande précision globale. Les modèles combinés de cette manière fonctionne généralement aussi bien, sinon mieux, que les modèles individuels.

Cette combinaison de noeuds se produit automatiquement dans les noeuds de modélisation automatisée : Discriminant automatique, Numérisation automatique et Cluster automatique.

Après avoir utilisé un noeud Ensemble, vous pouvez utiliser un noeud Analyse ou Evaluation pour comparer la précision des résultats combinés avec chacun des modèles d'entrée. Pour ce faire, assurez-vous que l'option **Filtrer les champs générés par des modèles combinés** n'est pas sélectionnée dans l'onglet Paramètres du noeud Ensemble.

Champs de sortie

Chaque noeud Ensemble génère un champ contenant les scores combinés. Le nom est basé sur le champ cible spécifié et comporte le préfixe *$XF_*, *$XS_* ou *$XR_*, selon le niveau de mesure de champ (Indicateur, nominal (ensemble) ou continu (intervalle), respectivement). Par exemple, si la cible est un champ booléen nommé *réponse*, le champ de sortie est *$XF_response*.

Champs de confiance ou de propension. Pour les champs indicateurs et nominaux, d'autres champs de confiance ou de propension sont créés selon la méthode d'ensemble, comme illustré dans le tableau suivant.

Tableau 28. Création d'un champ Méthode d'ensemble.

Méthode d'ensemble	Nom de champ
Vote Vote pondéré par la confiance Vote pondéré par la propension brute Vote pondéré - propension ajustée La plus grande confiance gagne	*$XFC_<field>*
Propension brute moyenne	*$XFRP_<field>*
Propension brute moyenne ajustée	*$XFAP_<field>*

Paramètres du noeud Ensemble

Champ cible pour ensemble. Sélectionnez un champ unique qui est utilisé comme cible pour deux ou plusieurs modèles en amont. Les modèles en amont peuvent utiliser des champs cible Indicateur, Nominal ou Continu. Toutefois, au moins deux des modèles doivent partager la même cible afin de combiner les scores.

Filtrer des champs générés par des modèles combinés. Supprime de la sortie tous les champs supplémentaires générés par les modèles individuels qui sont intégrés au noeud Ensemble. Cochez cette case si seul le score combiné de tous les modèles d'entrée vous intéresse. Assurez-vous que cette option est désélectionnée si, par exemple, vous voulez utiliser un noeud Analyse ou Evaluation pour comparer la précision du score combiné avec chacun des modèles d'entrée individuels.

Les paramètres disponibles dépendent du niveau de mesure de champ sélectionné comme cible.

Cibles continues

Pour une cible continue, la moyenne des scores est effectuée. Il s'agit de la seule méthode disponible pour la combinaison des scores.

Lorsque vous effectuez la moyenne des scores ou des évaluations, le noeud Ensemble utilise un calcul d'erreur standard pour déterminer la différence entre les valeurs mesurées ou estimées et les valeurs réelles et pour montrer la correspondance proche de ces évaluations. Le calcul d'erreur standard est généré par défaut pour les nouveaux modèles ; vous pouvez néanmoins décocher la case des modèles existants, par exemple s'ils doivent être régénérés.

Cibles catégorielles

Pour ce type de cible, un certain nombre de méthodes sont prises en charge, dont le **vote**, qui fonctionne en comptant le nombre de fois où chaque valeur prédite possible est choisie et en sélectionnant la valeur avec le total le plus élevé. Par exemple, si trois modèles sur cinq prédisent *oui* et que les deux autres prédisent *non*, le*oui* remporte par un vote de 3 contre 2. Les votes peuvent être aussi **pondérés** selon la valeur de confiance ou de propension pour chaque prédiction. La somme des pondérations est ensuite

effectuée, et la valeur avec le total le plus élevé est à nouveau sélectionné. La fiabilité de la prédiction finale est la somme des pondérations pour la valeur gagnante divisée par le nombre de modèles inclus dans l'ensemble.

Tous les champs catégoriels. Pour les champs Indicateur et Nominal, les méthodes suivantes sont prises en charge :

- Vote
- Vote pondéré par la confiance
- La plus grande confiance gagne

Champs indicateur uniquement. Pour les champs booléens uniquement, un certain nombre de méthodes basées sur la propension sont également disponibles :

- Vote pondéré par la propension brute
- Vote pondéré par la propension ajustée
- Propension brute moyenne
- Propension moyenne ajustée

Ex æquo du vote. Pour les méthodes de vote, vous pouvez indiquer le mode de résolution des ex æquo.

- **Sélection aléatoire.** Une des valeurs ex æquo est choisie au hasard.
- **Confiance la plus grande.** La valeur ex aequo prédite avec la plus grande fiabilité gagne. Remarque : ce n'est pas forcément la même que la plus grande fiabilité de toutes les valeurs prédites.
- **Propension brute ou ajustée (champs indicateurs uniquement).** La valeur ex æquo prédite avec la plus grande propension absolue, où la propension absolue est calculée avec :

```
abs(0.5 - propensity) *
2
```

Ou, dans le cas d'une propension ajustée :

```
abs(0.5 - adjusted propensity) * 2
```

Noeud Partitionner

Les noeuds Partitionner sont utilisés pour générer un champ de partition qui sépare les données en sous-ensembles ou en échantillons distincts pour les phases d'apprentissage, de test et de validation de la génération de modèle. L'utilisation d'un échantillon pour la génération du modèle et d'un échantillon distinct pour le tester vous permet d'avoir une bonne indication de la manière dont le modèle peut se généraliser à des jeux de données plus importants, semblables aux données actuelles.

Le noeud Partitionner génère un champ nominal dont le rôle est configuré sur **Partitionner**. Si vos données comportent déjà un champ adapté, vous pouvez également le désigner en tant que partition à l'aide d'un noeud Typer. Dans ce cas, vous n'avez pas besoin d'un noeud Partitionner distinct. Tout champ nominal instancié comportant deux ou trois valeurs peut être utilisé en tant que partition à l'exception des champs indicateurs. Pour plus d'informations, voir «Définition du rôle de champ», à la page 155.

Vous pouvez définir plusieurs champs de partition dans un flux, mais vous devrez alors sélectionner un champ de partition unique dans l'onglet Champs de chaque noeud de modélisation utilisant la partition. (Dans le cas d'une seule partition, cette partition est automatiquement utilisée lorsque la fonction de partition est activée.)

Activation des partitions. Pour utiliser la partition dans une analyse, vous devez l'activer dans l'onglet Options de modèle du noeud de génération de modèle ou d'analyse approprié. Désélectionnez cette option pour pouvoir désactiver la partition sans supprimer le champ.

Pour créer un champ de partition en fonction de certains critères, tels qu'un intervalle de date ou un emplacement, vous pouvez également utiliser un noeud Calculer. Pour plus d'informations, voir «Noeud Calculer», à la page 162.

Exemple. Lors de la création d'un flux RFM pour identifier les clients récents qui ont réagi favorablement aux précédentes campagnes de marketing, le service marketing d'une société de ventes utilise un noeud Partitionner pour diviser les données en partitions de formation et de test.

Options du noeud Partitionner

Champ de partition. Indique le nom du champ créé par le noeud.

Partitions. Vous pouvez séparer les données en deux (apprentissage et test) ou trois (apprentissage, test et validation) échantillons.

- **Apprentissage et test.** Partitionne les données en deux échantillons et vous permet de former le modèle avec un échantillon et de le tester avec l'autre.
- **Apprentissage, test et validation.** Partitionne les données en trois échantillons, et vous permet de former le modèle avec un échantillon, de le tester et de l'affiner avec le deuxième et de valider les résultats à l'aide du troisième. La taille de chaque partition s'en trouve réduite, ce qui peut s'avérer très pratique lors de l'utilisation d'un jeu de données très volumineux.

Taille de la partition. Indique la taille relative de chaque partition. Si la somme des tailles de partition est inférieure à 100 %, les enregistrements non inclus dans une partition sont ignorés. Par exemple, un utilisateur dispose de 10 millions d'enregistrements et de tailles de partition d'apprentissage de 5 % et de test de 10 %. Une fois le noeud exécuté, environ 500 000 enregistrements d'apprentissage et un million d'enregistrements de test doivent exister, le reste ayant été ignoré.

Valeurs. Indique les valeurs utilisées pour représenter chaque échantillon de partition dans les données.

- **Utiliser les valeurs définies par le système ("1", "2" et "3").** Utilise un entier pour représenter chaque partition ; par exemple, tous les enregistrements appartenant à l'échantillon d'apprentissage ont la valeur 1 pour le champ de partition. Cela garantit la portabilité des données entre les différents paramètres régionaux, ainsi que la conservation de l'ordre de tri (de sorte que 1 représente toujours la partition d'apprentissage) si le champ de partition est à nouveau instancié ailleurs (par exemple, lors de la nouvelle lecture des données provenant d'une base de données). Cependant, les valeurs nécessitent une interprétation.
- **Ajouter des libellés aux valeurs définies par le système.** Combine l'entier avec un libellé ; par exemple, les enregistrements de partition d'apprentissage ont la valeur `1_Apprentissage`. Lors de la consultation des données, vous pouvez ainsi déterminer à quoi correspondent les valeurs ; en outre, cela permet de préserver l'ordre de tri. Cependant, les valeurs sont propres à un paramètre régional donné.
- **Utiliser les libellés en tant que valeurs.** Utilise le libellé sans entier ; par exemple, **Apprentissage**. Cela permet d'indiquer les valeurs en modifiant les libellés. Cependant, les données sont alors régionales et la réinstanciation d'une colonne de partition trie les valeurs dans leur ordre de tri naturel, qui ne correspond pas forcément à leur ordre «sémantique».

Valeur de départ. Disponible seulement lorsque l'option **Affectation de partition répétable** est sélectionnée. Lors de l'échantillonnage ou du partitionnement d'enregistrements en fonction d'un pourcentage aléatoire, cette option vous permet de dupliquer les mêmes résultats dans une autre session. Indiquez la valeur de départ utilisée par le générateur de nombres aléatoires pour vous assurer que les mêmes enregistrements sont affectés à chaque exécution du noeud. Entrez la valeur de départ souhaitée ou cliquez sur le bouton **Générer** pour générer automatiquement une valeur aléatoire. Si cette option n'est pas sélectionnée, un échantillon différent est généré à chaque exécution du noeud.

Remarque : Si vous utilisez l'option **Valeur de départ** avec des enregistrements lus depuis une base de données, il peut être nécessaire d'exécuter un noeud Trier avant l'échantillonnage afin de garantir le

même résultat à chaque exécution du noeud. En effet, la valeur de départ aléatoire dépend de l'ordre des enregistrements, et il n'est pas garanti que cet ordre reste inchangé dans une base de données relationnelle. Pour plus d'informations, voir «Noeud Trier», à la page 88.

Utiliser un champ unique pour affecter des partitions. Disponible seulement lorsque l'option **Affectation de partition répétable** est sélectionnée. (Uniquement pour des bases de données de niveau 1) Cochez cette case pour utiliser les conversions SQL afin d'affecter des enregistrements à des partitions. Dans la liste déroulante, sélectionnez un champ avec des valeurs uniques (par exemple un champ d'ID) pour vous assurer que les enregistrements sont affectés de façon aléatoire mais pouvant être répétée.

Les niveaux de base de données sont expliqués dans la description du noeud source de la base de données. Pour plus d'informations, voir «noeud source de base de données», à la page 18.

Génération de noeuds Sélectionner

Le menu Générer du noeud Partitionner permet de générer automatiquement un noeud Sélectionner pour chaque partition. Par exemple, vous pouvez sélectionner tous les enregistrements de la partition d'apprentissage afin d'obtenir une évaluation ou une analyse plus poussées avec cette seule partition.

Noeud Binariser

Le noeud Binariser est utilisé pour calculer des champs indicateurs en fonction des valeurs catégorielles définies pour un ou plusieurs champs nominaux. Par exemple, votre jeu de données peut contenir un champ nominal, *TA* (tension artérielle), ainsi que les valeurs *Elevée*, *Normale* et *Faible*. Pour faciliter la manipulation des données, vous pouvez créer un champ indicateur pour la tension artérielle élevée ; celui-ci indique alors si le patient a une tension artérielle élevée.

Paramétrage des options du noeud Binariser

Champs d'ensemble. Répertorie tous les champs de données ayant un niveau de mesure *Nominal* (ensemble). Sélectionnez un champ dans la liste pour afficher les valeurs de l'ensemble. Vous pouvez choisir l'une de ces valeurs pour créer un champ booléen. Pour que vous puissiez voir les champs nominaux disponibles (et leurs valeurs), les données doivent d'abord être entièrement instanciées à l'aide du noeud Type ou Source en amont. Pour plus d'informations, voir «Noeud Typer», à la page 144.

Extension nom de champ. Sélectionnez cette option pour permettre aux contrôles de spécifier une extension qui sera ajoutée au nouveau champ booléen en tant que suffixe ou préfixe. Par défaut, les nouveaux noms de champ sont automatiquement créés en combinant le nom de champ d'origine et la valeur du champ afin d'obtenir un libellé de type *Nomchamp_valeurchamp*.

Valeurs d'ensemble disponibles. Les valeurs de l'ensemble sélectionné plus haut apparaissent ici. Sélectionnez les valeurs pour lesquelles générer des booléens. Par exemple, si les valeurs d'un champ appelé *tension_artérielle* sont *Elevée*, *Moyenne* et *Faible*, vous pouvez sélectionner *Elevée* et l'ajouter à la liste sur la droite. Cette opération entraîne la création d'un champ comportant un booléen pour les enregistrements dotés d'une valeur indiquant une tension artérielle élevée.

Créer des champs indicateurs. Les nouveaux champs booléens sont répertoriés ici. Vous pouvez spécifier des options déterminant l'attribution du nom aux nouveaux champs à l'aide des contrôles d'extension de nom de champ.

Valeur vraie (true). Indiquez la valeur true (vrai) utilisée par le noeud lors de la définition d'un booléen. Par défaut, cette valeur est **T**.

Valeur fausse (false). Indiquez la valeur false (faux) utilisée par le noeud lors de la définition d'un booléen. Par défaut, cette valeur est **F**.

Clés d'agrégation. Sélectionnez cette option pour grouper les enregistrements sur la base des champs-clés spécifiés plus bas. Lorsque l'option **Clés d'agrégation** est sélectionnée, tous les champs booléens d'un groupe sont activés si l'*un* des enregistrements a été défini comme true (vrai). Utilisez le sélecteur de champs pour choisir les champs-clés à utiliser pour agréger les enregistrements.

Noeud Restructurer

Le noeud Restructurer peut être utilisé pour générer plusieurs champs en fonction des valeurs d'un champ nominal ou d'un champ indicateur. Les champs nouvellement générés peuvent contenir des valeurs issues d'un autre champ ou de champs booléens numériques (0 et 1). La fonctionnalité de ce noeud est semblable à celle du noeud Binariser. Toutefois, il offre une plus grande souplesse d'utilisation. Il permet de créer des champs de tout type (y compris les booléens numériques), à l'aide des valeurs issues d'un autre champ. Vous pouvez ainsi effectuer une agrégation ou d'autres manipulations avec d'autres noeuds situés en aval. (Grâce au noeud Binariser, vous pouvez agréger des champs en une seule étape ; cela peut s'avérer utile lorsque vous créez des champs indicateurs.)

Par exemple, le jeu de données suivant contient un champ nominal *Compte*, et les valeurs *Epargne* et *Courant*. Le solde d'ouverture et le solde actuel sont enregistrés pour chaque compte. Certains clients possèdent plusieurs comptes de chaque type. Supposons que vous souhaitiez savoir si chaque client possède un type de compte particulier et, si tel est le cas, la somme figurant sur chaque type de compte. Utilisez le noeud Restructurer pour générer un champ pour chacune des valeurs du champ *Compte* et sélectionnez *Solde_actuel* comme valeur. Chaque nouveau champ est renseigné par le solde actuel de l'enregistrement concerné.

Tableau 29. Exemple de données avant restructuration.

IDclient	Compte	Solde_ouverture	Solde_actuel
12701	Mode brouillon	1000	1005.32
12702	Epargne	100	144.51
12703	Epargne	300	321.20
12703	Epargne	150	204.51
12703	Mode brouillon	1200	586.32

Tableau 30. Exemple de données après restructuration.

IDclient	Compte	Solde_ouverture	Solde_actuel	Compte_Courant_ Solde_actuel	Compte_Epargne _Solde_actuel
12701	Mode brouillon	1000	1005.32	1005.32	$null$
12702	Epargne	100	144.51	$null$	144.51
12703	Epargne	300	321.20	$null$	321.20
12703	Epargne	150	204.51	$null$	204.51
12703	Mode brouillon	1200	586.32	586.32	$null$

Utilisation du noeud Restructurer avec le noeud Agréger

Dans de nombreux cas, vous pouvez combiner le noeud Restructurer avec le noeud Agréger. Dans l'exemple précédent, un client (doté de l'ID 12703) possède trois comptes. Vous pouvez utiliser un noeud Agréger pour calculer le solde total de chaque type de compte. Le champ-clé est *IDclient* et les champs d'agrégation sont les nouveaux champs restructurés, *Compte_Courant_Solde_actuel* et *Compte_Epargne_Solde_actuel*. Le tableau ci-dessous présente les résultats obtenus.

Tableau 31. Exemple de données après restructuration et agrégation.

IDclient	Effectif	Compte_Courant_Solde_ actuel_Somme	Compte_Epargne_Solde_ actuel_Somme
12701	1	1005.32	$null$
12702	1	$null$	144.51
12703	3	586.32	525.71

Paramétrage des options du noeud Restructurer

Champs disponibles. Répertorie tous les champs de données ayant un niveau de mesure *Nominal* (ensemble) ou *Indicateur*. Sélectionnez un champ dans la liste pour afficher les valeurs de l'ensemble ou de l'indicateur, puis choisissez les valeurs souhaitées pour créer les champs restructurés. Pour que vous puissiez voir les champs disponibles (et leurs valeurs), les données doivent d'abord être entièrement instanciées à l'aide du noeud Typer ou source en amont. Pour plus d'informations, voir «Noeud Typer», à la page 144.

Valeurs disponibles. Les valeurs de l'ensemble sélectionné plus haut apparaissent ici. Sélectionnez les valeurs pour lesquelles générer des champs restructurés. Par exemple, si les valeurs d'un champ appelé *Tension artérielle* sont *Elevée*, *Moyenne* et *Faible*, vous pouvez sélectionner *Elevée* et l'ajouter à la liste figurant sur la droite. Un champ est ainsi créé et renseigné à partir d'une valeur définie (voir ci-dessous) pour les enregistrements présentant la valeur *Elevée*.

Créer des champs restructurés. Les nouveaux champs restructurés sont répertoriés ici. Par défaut, les nouveaux noms de champ sont automatiquement créés en combinant le nom de champ d'origine et la valeur du champ afin d'obtenir un libellé de type *Nomchamp_valeurchamp*.

Inclure le nom des champs. Désélectionnez cette option pour ne pas inclure le nom du champ d'origine comme préfixe dans les nouveaux noms de champ.

Utiliser les valeurs d'autres champs. Indiquez un ou plusieurs champs dont la valeur sera utilisée pour renseigner les champs restructurés. Utilisez pour cela le sélecteur de champs. Un champ est créé pour chaque champ sélectionné. Le nom du champ de valeur est ajouté au nom du champ restructuré ; par exemple *TA_Elevée_Age* ou *TA_Faible_Age*. Chaque nouveau champ hérite du type du champ de valeur d'origine.

Créer des indicateur de valeur numérique. Sélectionnez cette option pour renseigner les nouveaux champs à l'aide de booléens de valeur numérique (0 pour false (faux) et 1 pour true (vrai)), et non à partir d'une valeur d'un autre champ.

Noeud Transposer

Par défaut, les colonnes correspondent aux champs, et les lignes aux enregistrements ou aux observations. Vous pouvez utiliser, si nécessaire, un noeud Transposer pour faire permuter les données des lignes et des colonnes afin que les champs deviennent des enregistrements et que les enregistrements deviennent des champs. Par exemple, si vous disposez de séries temporelles, où chaque série est une ligne et non une colonne, vous pouvez transposer les données avant de procéder à l'analyse.

Définition des options du noeud Transposer

Dans la liste déroulante **Méthode de transposition**, sélectionnez la méthode à effectuer par le noeud Transposer : **Champs et enregistrements**, **Enregistrements vers champs** ou **Champs vers enregistrements**. Les paramètres de chacune des trois méthodes sont décrits dans les sections ci-après.

Restriction : Les méthodes **Enregistrements vers champs** et **Champs vers enregistrements** ne sont prises en charge que sous Windows 64 bits, Linux 64 bits et Mac.

Champs et enregistrements

Les nouveaux noms de champ peuvent être générés automatiquement en fonction d'un préfixe spécifié ou lus à partir d'un champ existant dans les données.

Utiliser un préfixe. Cette option génère automatiquement de nouveaux noms de champ à partir du préfixe indiqué (`Champ1`, `Champ2`, etc.). Vous pouvez personnaliser le préfixe comme souhaité. Lorsque vous utilisez cette option, vous devez indiquer le nombre de champs à créer, quel que soit le nombre de lignes présentes dans les données d'origine. Par exemple, si l'option **Nombre de nouveaux champs** est paramétrée sur 100, toutes les données figurant au-delà des 100 premières lignes sont ignorées. Si les données d'origine contiennent moins de 100 lignes, certains champs prennent la valeur nulle. (Vous pouvez augmenter le nombre de champs selon vos besoins ; il convient toutefois d'éviter d'utiliser ce paramètre pour transposer un million d'enregistrements en un million de champs, ce qui produirait un résultat ingérable.)

Par exemple, supposons que vous disposez de données avec des séries en lignes et un champ distinct (colonne) pour chaque mois. Vous pouvez transposer ces données de telle manière que chaque série apparaisse dans un champ distinct, avec une ligne pour chaque mois.

Lire à partir du champ. Lit les noms de champ à partir d'un champ existant. Avec cette option, le nombre de nouveaux champs est déterminé par les données, jusqu'à atteindre la limite maximale indiquée. Chaque valeur du champ sélectionné devient un nouveau champ dans les données de sortie. Le champ sélectionné peut présenter n'importe quel type de stockage (entier, chaîne, date, heure, etc.), mais afin d'éviter des noms de fichier en double, chaque valeur du champ sélectionné doit être unique (en d'autres termes, le nombre de valeurs doit correspondre au nombre de lignes). Lorsque des noms de champ en double sont détectés, un avertissement apparaît.

- **Lire les valeurs.** Si le champ sélectionné n'a pas été instancié, sélectionnez cette option pour renseigner la liste des nouveaux noms de champ. Si le champ a déjà été instancié, cette étape n'est pas nécessaire.
- **Nombre maximal de valeurs à lire.** Lors de la lecture des noms de champ à partir des données, une limite supérieure est définie afin d'éviter de créer un nombre de champs trop important. (Comme indiqué ci-dessus, la transposition d'un million d'enregistrements en un million de champs produirait un résultat ingérable.)

Par exemple, si la première colonne de données indique le nom de chaque série, vous pouvez utiliser ces valeurs en tant que noms de champ dans les données transposées.

Transposer. Par défaut, seuls les champs continus (intervalle numérique) sont transposés (stockage de type entier ou nombre réel). Si nécessaire, vous pouvez sélectionner un sous-ensemble de champs numériques ou transposer des champs de type chaîne. Toutefois, tous les champs transposés doivent comporter le même type—either de stockage (numérique ou chaîne, mais pas les deux) ; en effet, l'utilisation de champs d'entrée mixtes générerait des valeurs mixtes au sein de chaque colonne de sortie, ce qui irait à l'encontre de la règle selon laquelle toutes les valeurs d'un champ doivent être dotées du même type de stockage. Les autres types de stockage (date, heure, horodatage) ne peuvent pas être transposés.

- **Tous les champs numériques.** Transpose tous les champs numériques (stockage de type entier ou nombre réel). Le nombre de lignes dans la sortie correspond au nombre de champs numériques dans les données d'origine.
- **Toutes les chaînes.** Transpose tous les champs de type chaîne.
- **Personnalisé.** Permet de sélectionner un sous-ensemble de champs numériques. Le nombre de ligne dans la sortie correspond au nombre de champs sélectionnés. Cette option n'est disponible que pour les champs numériques.

Nom d'ID de ligne. Indique le nom du champ d'ID de ligne créé par le noeud. Les valeurs de ce champ sont déterminées en fonction du nom des champs figurant dans les données d'origine.

Conseil : si vous transposez des séries temporelles de lignes en colonnes et que les données d'origine incluent une ligne, telle que Date, Mois ou Année, qui sert d'étiquetage de période à chaque mesure, veillez à lire ces libellés comme des noms de champ dans IBM SPSS Modeler (comme le montrent les exemples précédents, qui affichent le mois ou le jour comme noms de champ dans les données d'origine, respectivement) au lieu d'inclure le libellé dans la première ligne de données. Ainsi, vous éviterez de mélanger les libellés et les valeurs dans chaque colonne (ce qui obligerait les nombres à être lus comme des chaînes car les types de stockage ne peuvent pas être mélangés dans une colonne).

Enregistrements vers champs

Champs. La liste Champs contient tous les champs qui accèdent au noeud Transposer.

Index. La section Index permet de sélectionner les champs à utiliser comme champs d'index.

Champs. La section Champs permet de sélectionner les champs à utiliser comme champs.

Valeur. La section Valeur permet de sélectionner les champs à utiliser comme champs de valeur.

Fonction d'agrégation. S'il existe plusieurs enregistrements pour un index, vous devez les agréger en un seul. Utilisez la liste déroulante **Fonction d'agrégation** pour indiquer comment agréger les enregistrements à l'aide de l'une des fonctions ci-après. Notez que l'agrégation a un impact sur tous les champs.

- **Moyenne.** Renvoie les valeurs moyennes de chaque combinaison de champs-clés. La moyenne est une mesure de tendance centrale et est la moyenne arithmétique (la somme divisée par le nombre de cas).
- **Somme.** Renvoie les valeurs additionnées de chaque combinaison de champs-clés. La somme ou le total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.
- **Min.** Renvoie les valeurs minimales de chaque combinaison de champs-clés.
- **Max.** Renvoie les valeurs maximales de chaque combinaison de champs-clés.
- **Médiane.** Renvoie les valeurs médianes de chaque combinaison de champs-clés. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées. Elle est également nommée 50ème centile ou 2ème quartile.
- **Comptage.** Renvoie le nombre de valeurs non null de chaque combinaison de champs-clés.

Champs vers enregistrements

Champs. La liste Champs contient tous les champs qui accèdent au noeud Transposer.

Index. La section Index permet de sélectionner les champs à utiliser comme champs d'index.

Valeur. La section Valeur permet de sélectionner les champs à utiliser comme champs de valeur. Si vous ne sélectionnez pas de champs de valeur, tous les champs numériques non affectés sont utilisés comme valeurs. Toutefois, si des champs non numériques sont disponibles, tous les champs de chaîne non affectés sont utilisés.

Noeud Historiser

Les noeuds Historiser sont souvent utilisés pour les données séquentielles, telles que les séries temporelles. Ils servent à créer des champs contenant des données provenant de champs d'enregistrements antérieurs. Lorsque vous utilisez un noeud Historiser, si vous souhaitez obtenir des données prétriées selon un champ particulier, vous pouvez utiliser le noeud Trier.

Paramétrage des options du noeud Historiser

Champs sélectionnés. A l'aide du sélecteur de champs (bouton à droite de la zone de texte), sélectionnez les champs pour lesquels vous souhaitez obtenir un historique. Chaque champ sélectionné est utilisé pour créer des champs pour tous les enregistrements du jeu de données.

Décalage. Indiquez le dernier enregistrement avant l'enregistrement actuel à partir duquel extraire les valeurs de champ historiques. Par exemple, si le décalage est défini sur 3, au fur et à mesure que chaque enregistrement passe dans le noeud, les valeurs de champ du troisième enregistrement précédent sont incluses dans l'enregistrement actuel. Utilisez les paramètres d'amplitude pour indiquer jusqu'à quel enregistrement portera l'extraction. Utilisez les flèches pour rectifier la valeur de décalage.

Amplitude. Indiquez le nombre d'enregistrements précédents desquels extraire des valeurs. Par exemple, si le décalage est défini sur 3 et l'amplitude sur 5, chaque enregistrement qui passe dans le noeud se verra ajouter cinq champs pour chacun des champs spécifiés dans la liste Champs sélectionnés. Autrement dit, lorsque le noeud traite l'enregistrement 10, des champs provenant des enregistrements 7 à 3 sont ajoutés. Utilisez les flèches pour rectifier la valeur d'amplitude.

Lorsque l'historique n'est pas disponible. Sélectionnez l'une des options suivantes pour traiter les enregistrements qui n'ont pas de valeurs historiques. Il s'agit généralement des premiers enregistrements du jeu de données, pour lesquels aucun enregistrement précédent ne peut être utilisé en tant qu'historique.

- **Supprimer les enregistrements.** Sélectionnez cette option pour supprimer les enregistrements dans lesquels aucune valeur d'historique n'est disponible pour le champ sélectionné.
- **Conserver l'historique non défini.** Sélectionnez cette option pour conserver les enregistrements dans lesquels aucune valeur d'historique n'est disponible. Une valeur non définie apparaît dans le champ d'historique (`$null$`).
- **Remplacer les valeurs avec.** Indiquez la valeur ou la chaîne à utiliser pour les enregistrements dans lesquels aucune valeur d'historique n'est disponible. La valeur de remplacement par défaut est *undef*, la valeur système nulle. Les valeurs nulles sont indiquées par la chaîne `$null$`.

Lorsque vous sélectionnez une valeur de remplacement, gardez à l'esprit les règles suivantes pour que l'exécution se déroule correctement :

- Les champs sélectionnés doivent être du même type de stockage.
- Si tous les champs sélectionnés présentent un stockage numérique, la valeur de remplacement doit être analysée en tant qu'entier.
- Si tous les champs sélectionnés présentent un stockage réel, la valeur de remplacement doit être analysée en tant que nombre réel.
- Si tous les champs sélectionnés présentent un stockage symbolique, la valeur de remplacement doit être analysée en tant que chaîne.
- Si tous les champs sélectionnés présentent un stockage date/heure, la valeur de remplacement doit être analysée en tant que champ date/heure.

Si l'une des conditions ci-dessus n'est pas remplie, une erreur se produit lors de l'exécution du noeud Historiser.

Noeud Re-trier

Le noeud Re-trier permet de définir l'ordre naturel utilisé pour afficher les champs situés en aval. Cet ordre a une incidence sur l'affichage des champs en différents endroits : tableaux, listes et sélecteur de champs. Cette opération est utile, par exemple, lorsque vous utilisez des jeux de données volumineux pour rendre plus visibles les champs intéressants.

Paramétrage des options du noeud Re-trier

Il existe deux méthodes de réorganisation des champs : ordre personnalisé et tri automatique.

Ordre personnalisé

Sélectionnez **Ordre personnalisé** pour activer une table de noms et de types de champ dans laquelle vous pouvez afficher tous les champs et utiliser les flèches pour créer un ordre personnalisé.

Pour réorganiser les champs :

1. Sélectionnez un champ dans le tableau. Utilisez la méthode Ctrl+clic pour sélectionner plusieurs champs.
2. Utilisez les boutons représentant une simple flèche pour déplacer les champs d'un rang vers le haut ou vers le bas.
3. Utilisez les boutons représentant une flèche et une ligne pour placer les champs tout en bas ou tout en haut de la liste.
4. Spécifiez l'ordre des champs qui ne sont pas inclus ici en déplaçant la ligne séparatrice, indiquée par `[ autres champs ]`, vers le haut ou vers le bas.

Plus d'informations sur [autres champs]

Autres champs. L'objectif de la ligne séparatrice **[autres champs]** est de diviser la table en deux parties.

- Les champs qui apparaissent au-dessus de la ligne séparatrice sont ordonnés (tels qu'ils apparaissent dans la table) avant tous les ordres naturels utilisés pour afficher les champs en aval de ce noeud.
- Les champs qui apparaissent au-dessous de la ligne séparatrice sont ordonnés (tels qu'ils apparaissent dans la table) après tous les ordres naturels utilisés pour afficher les champs en aval de ce noeud.

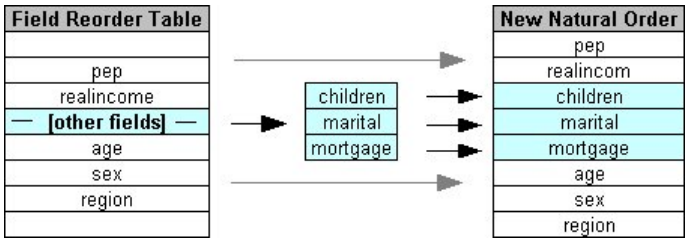


Figure 6. Diagramme illustrant la manière dont les "autres champs" sont incorporés dans le nouvel ordre des champs

- Tous les champs qui n'apparaissent pas dans la table de réorganisation des champs figurent entre les champs "supérieurs" et "inférieurs" à l'emplacement de la ligne séparatrice.

Voici d'autres options de tri personnalisé :

- Triez les champs dans l'ordre croissant ou décroissant en cliquant sur les flèches situées au-dessus de chaque en-tête de colonne (**Type**, **Nom** et **Stockage**). Lorsque vous effectuez un tri par colonne, les champs qui n'y sont pas mentionnés (ceux indiqués par la ligne `[ autres champs ]`) sont triés en dernier dans leur ordre naturel.
- Cliquez sur **Effacer les éléments non utilisés** pour supprimer du noeud Re-trier tous les champs inutilisés. Les champs inutilisés sont affichés en rouge dans le tableau. Cette couleur indique que le champ a été supprimé dans des opérations en amont.
- Indiquez l'ordre de tous les nouveaux champs (les nouveaux champs ou les champs non spécifiés sont identifiés par une icône représentant un éclair). Lorsque vous cliquez sur **OK** ou sur **Appliquer**, l'icône disparaît.

Remarque : si des champs sont ajoutés en amont après l'application d'un tri personnalisé, ces nouveaux champs sont ajoutés à la fin de la liste personnalisée.

Tri automatique

Sélectionnez **Tri automatique** pour indiquer le paramètre de tri. La boîte de dialogue change de manière dynamique pour proposer les options de tri automatique.

Trier par. Sélectionnez l'un des trois modes de tri des champs du noeud Réorganiser. Les flèches indiquent si l'ordre est croissant ou décroissant. Sélectionnez une option pour apporter un changement.

- Nom
- Type
- Stockage

Les champs ajoutés en amont du noeud Re-trier après l'application d'un tri automatique sont automatiquement placés à l'endroit qui convient, en fonction du type de tri sélectionné.

Noeud Intervalles de temps

Le noeud Intervalles de temps d'origine, qui était disponible dans SPSS Modeler version 17.1 et les versions antérieures, n'était pas compatible avec Analytic Server (AS) et est devenu obsolète dans SPSS Modeler version 18.0.

Le noeud Intervalles de temps qui le remplace contient un certain nombre de modifications par rapport au noeud Intervalles de temps d'origine. Ce nouveau noeud peut être utilisé avec Analytic Server ou dans SPSS Modeler lui-même.

Vous utilisez le noeud Intervalles de temps pour spécifier des intervalles et calculer un nouveau champ Heure pour l'estimation ou la prévision. Un ensemble complet d'intervalles de temps, allant des secondes aux années, est pris en charge.

Utilisez le noeud pour dériver une nouvelle zone Heure ; le nouveau champ possède le même type de stockage que la champ Heure d'entrée de votre choix. Le noeud génère les éléments suivants :

- Le champ spécifié dans l'onglet Champs comme **Zone Heure**, avec le préfixe/suffixe choisi. Par défaut, le préfixe est `$TI_`.
- Les champs spécifiés dans l'onglet Champs comme **Zones Dimension**.
- Les champs spécifiés dans l'onglet Champs comme **Zones à agréger**.

Plusieurs champs supplémentaires peuvent également être générés en fonction de la période ou de l'intervalle sélectionné (par exemple, la minute ou la seconde à laquelle une mesure échoue).

Intervalles de temps - Options de champ

Utilisez l'onglet Champs dans le noeud Intervalles de temps pour sélectionner les données à partir duquel le nouvel intervalle de temps est calculé.

Champs Affiche tous les champs d'entrée dans le noeud, avec leurs icônes de type de mesure. Tous les champs de temps possèdent le type de mesure 'Continu'. Sélectionnez le champ à utiliser comme entrée.

Zone Heure Affiche le champ d'entrée depuis lequel le nouvel intervalle de temps est calculé ; un champ continu unique est admis seulement. Le champ est utilisé pour le noeud Intervalles de temps comme clé d'agrégation pour la conversion de l'intervalle. Le nouveau champ possède le même type de stockage que le champ d'heure d'entrée choisi. Si vous sélectionnez un champ de type entier, il est considéré comme un index de temps.

Zones Dimension En option, vous pouvez ajouter des champs ici pour créer une série temporelle individuelle en fonction des valeurs de champ. Comme exemple tout simple, avec des données

géospatiales, vous pouvez utiliser un champ de point comme dimension. Dans cet exemple, la sortie de données du noeud Intervalles de temps est triée en séries temporelles pour chaque valeur de point dans le champ de point.

Les dimensions sont idéales lorsque vous utilisez des données multidimensionnelles à plat, comme celles générées par le noeud, ou pour prendre en charge des données d'un type plus complexe, telles que les données géospatiales. Pour résumer, vous pouvez envisager d'utiliser les **Champs de dimension** comme l'équivalent d'une clause **`Group By`** dans une requête SQL ou comme des **Champs-clés** dans le noeud Agréger, mais les **Champs de dimension** sont plus sophistiqués de par leur nature en raison de leur capacité à traiter des structures de données plus compliquées que de seules données de ligne et de colonne traditionnelles.

Zones à agréger Sélectionnez les champs à agréger dans le cadre du changement de période du champ d'heure. Seuls les champs que vous sélectionnez ici sont disponibles dans l'onglet Génération pour la table **Paramètres personnalisés pour les zones spécifiées**. Les champs qui ne sont pas inclus sont filtrés et exclus des données qui quittent le noeud. Cela signifie que les champs restants dans la liste **Champs** sont exclus des données.

Intervalles de temps - Options de génération

Utilisez l'onglet Génération pour spécifier des options permettant de changer l'intervalle de temps et la façon dont les champs dans les données sont agrégés, en fonction de leur type de mesure.

Lorsque vous agrégez des données, les champs de date, d'heure et d'horodatage existants sont remplacés par les champs générés et sont supprimés de la sortie. Les autres champs sont agrégés en fonction des options que vous spécifiez dans cet onglet.

Intervalle de temps Sélectionnez l'intervalle et la périodicité pour la génération des séries.

Paramètres par défaut Sélectionnez l'agrégation par défaut à appliquer aux données de différents types. Elle est appliquée en fonction du niveau de mesure ; par exemple, des champs continus sont agrégés en fonction de la somme, alors que les champs nominaux utilisent le mode. Vous pouvez définir l'agrégation par défaut pour trois niveaux de mesure différents :

- **Continu** Les fonctions disponibles pour les champs continus sont **Somme**, **Moyenne**, **Min.**, **Max.**, **Médiane**, **1er quartile** et **3ème quartile**.
- Les options **Nominal** incluent **Mode**, **Min.** et **Max.**.
- Les options **Indicateur** sont **Valeur vraie (le cas échéant)** ou **Faux si l'un des éléments est faux**.

Paramètres personnalisés pour les zones spécifiées Vous pouvez spécifier des exceptions aux paramètres d'agrégation par défaut pour des champs individuels. Utilisez les icônes de droite pour ajouter ou retirer des champs de la table, ou cliquez sur la cellule dans la colonne appropriée pour changer la fonction d'agrégation de ce champ. Les champs sans type sont exclus de la liste et ne peuvent pas être ajoutés au tableau.

Extension du nom du nouveau champ Spécifiez le **préfixe** ou le **suffixe** appliqué à tous les champs générés par le noeud.

Noeud de reprojection

Avec les données géospatiales ou de carte, les deux systèmes les plus souvent utilisés pour identifier les coordonnées sont le système de coordonnées géographiques et le système de coordonnées projetées. Dans IBM SPSS Modeler, les éléments tels que les fonctions spatiales du générateur de formules, le noeud de prévision spatio-temporelle (STP) et le noeud de visualisation de carte utilisent le système de coordonnées projetées ; par conséquent, les données que vous importez et qui sont enregistrées avec un système de coordonnées géographiques doivent être reprojetées. Si possible, les champs géospatiaux (c'est-à-dire les champs avec un niveau de mesure géospatial) sont reprojetés automatiquement lorsqu'ils

sont utilisés (et non lorsqu'ils sont importés). Lorsque des champs ne peuvent pas être reprojetés automatiquement, vous utilisez le noeud de reprojection pour changer leur système de coordonnées. Avec ce mode de reprojection, vous pouvez corriger les situations dans lesquelles une erreur survient en raison de l'utilisation d'un système de coordonnées incorrect.

La liste suivante répertorie des exemples de situations dans lesquelles il peut être nécessaire d'effectuer une reprojection afin de changer le système de coordonnées :

- *Ajout* Si vous tentez d'ajouter deux jeux de données avec des systèmes de coordonnées différents pour un champ géospatial, SPSS Modeler affiche le message d'erreur suivant : `Coordinate systems of <Field1> and <Field2> are not compatible. Reproject one or both fields to the same coordinate system.`

  <Field1> et *<Field2>* sont les noms des champs géospatiaux à l'origine de l'erreur.
- *Expression if/else* Si vous utilisez une expression qui contient une instruction if/else avec des champs géospatiaux ou des types de retour dans les deux parties de l'expression, mais qui utilisent des systèmes de coordonnées différents, SPSS Modeler affiche le message d'erreur suivant : `The conditional expression contains incompatible coordinate systems: <arg1> and <arg2>.`

  <arg1> and *<arg2>* sont les arguments then ou else qui renvoient un type géospatial avec des systèmes de coordonnées différents.
- *Construction d'une liste de champs géospatiaux* Pour créer une zone de liste qui contient de nombreux champs géospatiaux, tous les arguments de champ géospatial qui sont fournis dans l'expression de liste doivent être exprimés dans le même système de coordonnées. Sinon, le message d'erreur suivant s'affiche : `Coordinate systems of <Field1> and <Field2> are not compatible. Reproject one or both fields to the same coordinate system.`

Pour plus d'informations sur les systèmes de coordonnées, voir la rubrique sur la définition des options géospatiales pour les flux de la section Utilisation des flux dans le guide d'utilisation de SPSS Modeler.

Définition des options pour le noeud Reprojeter

Champs

Champs géographiques
: Par défaut, la liste est vide. Vous pouvez déplacer des champs géospatiaux dans cette liste depuis la liste **Zones à reprojeter** pour qu'elles ne soient pas reprojetées.

Zones à reprojeter
: Par défaut, cette liste contient toutes les champs géospatiaux qui sont ajoutés à ce noeud. Tous les champs de cette liste sont reprojetés dans le système de coordonnées que vous définissez dans le champ **Système de coordonnées**.

Système de coordonnées

Flux par défaut
: Sélectionnez cette option pour utiliser le système de coordonnées par défaut.

Spécifier
: Si vous sélectionnez cette option, vous pouvez utiliser le bouton **Changer** pour afficher la boîte de dialogue Sélection d'un système de coordonnées et choisir le système de coordonnées à utiliser pour la reprojection.

  Pour plus d'informations sur les systèmes de coordonnées, voir la rubrique sur la définition des options géospatiales pour les flux de la section Utilisation des flux dans le guide d'utilisation de SPSS Modeler.

Chapitre 5. Noeuds Graphiques

Fonctions communes des noeuds Graphiques

Au cours de plusieurs étapes du processus d'exploration de données, des graphiques et des diagrammes sont utilisés pour explorer les données introduites dans IBM SPSS Modeler. Par exemple, vous pouvez connecter un noeud Tracé ou Proportion à une source de données pour obtenir un aperçu des types de données et des proportions. Vous pouvez ensuite effectuer des manipulations de champ et d'enregistrement afin de préparer les données pour des opérations de modélisation en aval. Les graphiques permettent également de vérifier les distributions et les relations entre des champs nouvellement calculés.

La palette Graphiques contient les noeuds suivants :

Le noeud Représentation Graphique offre de nombreux types de graphiques différents dans un seul noeud. Ce noeud permet de choisir les champs de données que vous souhaitez explorer puis de sélectionner un graphique parmi ceux disponibles pour les données sélectionnées. Le noeud filtre automatiquement tous les types de graphiques ne fonctionnant pas avec les sélections de champs.

Le noeud Nuage montre les relations existant entre les champs numériques. Vous pouvez créer un graphique Nuage à l'aide de points (diagramme de dispersion) ou de lignes.

Le noeud Proportion fournit l'occurrence des valeurs symboliques (catégorielles), comme un type de prêt hypothécaire ou le sexe d'un individu. Ce noeud est souvent utilisé pour montrer les déséquilibres des données, déséquilibres que vous pouvez rectifier à l'aide d'un noeud Equilibrer avant la création d'un modèle.

Le noeud Histogramme montre l'occurrence des valeurs des champs numériques. Il est souvent utilisé pour explorer les données avant toute génération de modèle ou manipulation. Semblable au noeud Proportion, le noeud Histogramme sert souvent à montrer les déséquilibres des données.

Le noeud Résumé fournit la proportion de valeurs d'un champ numérique par rapport aux valeurs d'un autre champ. (Il génère des graphiques semblables aux histogrammes.) Il est utile pour illustrer une variable ou un champ dont les valeurs changent avec le temps. Grâce à la représentation graphique en 3D, vous pouvez en outre inclure un axe symbolique affichant les proportions par catégorie.

Le noeud Courbes génère un graphique qui affiche plusieurs champs *Y* pour un seul champ *X*. Les champs *Y* sont représentés par des lignes colorées. Chacun équivaut à un noeud Nuage dont le style est défini sur **Ligne** et le mode X sur **Trier**. Les graphiques Courbes sont utiles lorsque vous souhaitez étudier la fluctuation de plusieurs variables au fil du temps.

Le noeud Relations illustre la force de la relation existant entre les valeurs de plusieurs champs symboliques (catégoriels). Le graphique utilise des lignes d'épaisseur différente pour représenter les forces de connexion. Par exemple, vous pouvez utiliser un noeud Relations pour explorer la relation avec l'achat d'un ensemble d'articles sur un site de commerce électronique.

Le noeud Tracé horaire affiche un ou plusieurs jeux de données temporelles. En règle générale, vous utilisez un noeud Intervalles de temps, en premier lieu, pour créer un champ *TimeLabel* qui servira de libellé à l'axe *x*.

Le noeud Evaluation permet d'évaluer et de comparer des modèles prédictifs. Le graphique d'évaluation montre l'aptitude des modèles à prédire des résultats spécifiques. Il trie les enregistrements en fonction de la valeur prédite et de la fiabilité dans cette prévision. Il scinde les enregistrements en groupes de taille égale (**quantiles**), puis reporte la valeur du critère traité pour chaque quantile, du plus élevé au plus faible. Les divers modèles apparaissent sous forme de lignes dans le graphique.

Le noeud Visualisation de carte peut accepter plusieurs connexions d'entrée et afficher des données géospatiales sur une carte sous forme de série de couches. Chaque couche est un champ géospatial unique. Par exemple, la couche de base peut être la carte d'un pays ; sur cette couche, il peut y avoir une couche pour les routes, une couche pour les rivières et une couche pour les villes.

Le noeud E-Tracé (Bêta) montre les relations existant entre les champs numériques. Il ressemble au noeud Tracé, mais ses options sont différentes et sa sortie utilise une nouvelle interface de graphique spécifique à ce noeud. Utilisez le noeud de niveau bêta pour vous familiariser avec les nouvelles fonctions de graphique.

t-SNE (t-Distributed Stochastic Neighbor Embedding) est un outil permettant de visualiser des données en grande dimension. Il convertit les analogies de points de données en probabilités. Ce noeud t-SNE de SPSS Modeler est mis en oeuvre dans Python et nécessite la bibliothèque Python `scikit-learn`©.

Une fois que vous avez ajouté un noeud Graphique à un flux, vous pouvez double-cliquer sur le noeud pour ouvrir une boîte de dialogue qui permet de définir des options. La plupart des graphiques contiennent un certain nombre d'options spécifiques figurant sur un ou plusieurs onglets. Les onglets comportent également des options communes à tous les graphiques. Les sections suivantes contiennent des informations supplémentaires sur ces options communes.

Une fois que vous avez configuré les options d'un noeud Graphique, vous pouvez exécuter ce dernier dans la boîte de dialogue ou au sein d'un flux. Dans la fenêtre du graphique créé, vous pouvez générer des noeuds Calculer (Binariser) et Sélectionner en fonction d'une sélection ou d'une zone de données, ce qui entraîne la définition de sous-jeux de données. Par exemple, vous pouvez utiliser la puissance de cette fonction pour identifier et exclure les valeurs éloignées.

Apparences, superpositions, panneaux et animation

Superpositions et apparences

Les apparences (et les superpositions) ajoutent la dimensionnalité⌂ à une visualisation. L'effet d'une apparence (regroupement, juxtapostion ou empilement) dépend du type de visualisation, du type de champ (variable), ainsi que du type de l'élément graphique et des statistiques. Il est par exemple possible d'utiliser un champ catégoriel de couleur pour grouper des points dans un nuage de points ou pour créer les piles d'un graphique à barres superposées. Un intervalle numérique continu de couleur peut également permettre d'indiquer les valeurs d'intervalle pour chaque point d'un diagramme de dispersion.

Vous devez faire des essais avec les différentes apparences et superpositions pour trouver la solution qui répond le mieux à vos besoins. Les descriptions suivantes peuvent vous aider à faire le bon choix.

Remarque : Certaines apparences ou superpositions ne conviennent pas à certains types de visualisation.

- **Couleur.** Lorsque la couleur est définie par un champ catégoriel⌂, cela scinde la visualisation basée sur les catégories individuelles, avec une couleur pour chaque catégorie. Lorsque la couleur est une plage numérique continue, la couleur varie selon la valeur du champ de plage. Si l'élément graphique (par exemple, une barre ou une zone) représente plus d'un enregistrement/observation et qu'un champ de plage est utilisé pour la couleur, la couleur varie selon la *moyenne* du champ de plage.
- **Forme.** La forme est définie par un champ catégoriel⌂ qui scinde la visualisation en éléments de formes différentes, un élément pour chaque catégorie.
- **Transparence.** Lorsque la transparence est définie par un champ catégoriel⌂, cela scinde la visualisation basée sur les catégories individuelles, avec un niveau de transparence pour chaque catégorie. Lorsqu'une transparence est une plage numérique continue, cela modifie la transparence basée sur la valeur du champ de plage. Si l'élément graphique (par exemple, une barre ou une zone) représente plus d'un enregistrement/observation et qu'un champ de plage est utilisé pour la transparence, la couleur varie selon la *moyenne* du champ de plage. A la valeur la plus élevée, les éléments graphiques sont complètement transparents. Pour la valeur la plus faible, ils sont totalement opaques.
- **Libellé de données :** Les libellés de données sont définis par tout type de champ dont les valeurs sont utilisées pour créer des libellés qui sont joints aux éléments graphiques.
- **Taille.** Lorsque la taille est définie par un champ catégoriel⌂, cela scinde la visualisation basée sur les catégories individuelles, avec une taille pour chaque catégorie. Lorsque la taille est une plage numérique continu, cela modifie la taille basée sur la valeur du champ de plage. Si l'élément graphique (par exemple, une barre ou une zone) représente plus d'un enregistrement/observation et qu'un champ de plage est utilisé pour la taille, la taille varie selon la *moyenne* du champ de plage.

Panneaux et animation

Panélisation. La division en panneaux (ou création de facettes) crée un tableau de graphiques. Un graphique est généré pour chaque catégorie dans les champs de division en panneaux, mais tous les panels apparaissent en même temps. La division en panneaux est utile lorsqu'il s'agit de vérifier si la visualisation est soumise ou non aux conditions de champs de division en panels. Supposons par exemple que vous souhaitiez diviser en panels un histogramme par sexe de manière à déterminer si les distributions des effectifs sont les mêmes entre les hommes et les femmes. C'est-à-dire que vous pouvez vérifier si le salaire varie en fonction du sexe. Sélectionnez un champ catégoriel pour la division en panneaux.

Animation. L'animation ressemble à la division en panneaux car plusieurs graphiques sont créés à partir des valeurs du champ d'animation, mais ces graphiques ne sont pas affichés ensemble. Utilisez plutôt les contrôle⌂s dans le mode Explorer pour animer les sorties et parcourir une séquence de graphiques individuels. De plus, à la différence de la division en panneaux, l'animation ne requiert pas de champ catégoriel⌂. Vous pouvez spécifier un champ continu dont les valeurs sont automatiquement scindées en plages. Vous pouvez modifier la taille de la plage à l'aide des contrôle⌂s d'animation dans le mode Explorer. Seules certaines visualisations permettent l'animation.

Utilisation de l'onglet Sortie

Vous pouvez définir les options suivantes pour tous les types de graphique : elles concernent les noms de fichier et l'affichage des graphiques générés.

Remarque : Les graphiques de noeud distribution ont des fonctions supplémentaires.

Nom de la sortie - Spécifie le nom du graphique généré lorsque le noeud est exécuté. L'option **Automatique** sélectionne un nom en fonction du noeud qui génère la sortie. Si vous le souhaitez, vous pouvez choisir **Personnalisé** pour indiquer un autre nom.

Sortie à l'écran. Sélectionnez cette option pour générer et afficher le graphique dans une nouvelle fenêtre.

Sortie dans un fichier. Sélectionnez cette option pour enregistrer la sortie sous forme de fichier.

- **Graphique de sortie.** Sélectionnez cette option pour produire la sortie sous forme de graphique. Disponible uniquement dans les noeuds Proportion.
- **Table de sortie.** Sélectionnez cette option pour produire la sortie sous forme de tableau. Disponible uniquement dans les noeuds Proportion.
- **Nom de fichier.** Indiquez le nom de fichier du graphique ou du tableau généré. Utilisez le bouton **...** pour indiquer l'emplacement d'un fichier spécifique.
- **Type de fichier.** Spécifiez le type de fichier dans la liste déroulante. Pour tous les noeuds Graphiques, à l'exception du noeud distribution avec une option **Tableau de sortie**, les types de fichiers graphiques disponibles sont les suivants.

  - Bitmap (`.bmp`)
  - PNG (`.png`)
  - Objet de sortie (`.cou`)
  - JPEG (`.jpg`)
  - HTML (`.html`)
  - Document ViZml (`.xml`) à utiliser dans d'autres applications IBM SPSS Statistics.

  Pour l'option **Tableau de sortie** dans le noeud distribution, les types de fichiers disponibles sont les suivants.

  - Données délimitées par des tabulations (`.tab`)
  - Données délimitées par une virgule (`.csv`)
  - HTML (`.html`)
  - Objet de sortie (`.cou`)

Paginer la sortie. Lorsque vous enregistrez la sortie au format HTML, cette option est activée pour vous permettre de contrôler la taille de chaque page HTML. (S'applique uniquement au noeud Proportion.)

Lignes par page. Lorsque vous choisissez l'option **Paginer la sortie**, cette option est activée pour vous permettre de déterminer la longueur de chaque page HTML. Le paramètre par défaut est de 400. (S'applique uniquement au noeud Proportion.)

Utilisation de l'onglet Annotations

Utilisé pour tous les noeuds, cet onglet propose des options permettant de renommer les noeuds, de créer des info-bulles personnalisées et de stocker de longues annotations.

Graphiques en 3D

Les graphiques Nuage et Résumé d'IBM SPSS Modeler permettent d'afficher des informations sur un troisième axe. Vous disposez ainsi d'une flexibilité accrue lorsque vous visualisez vos données pour sélectionner des sous-ensembles ou calculez de nouveaux champs en vue d'une modélisation.

Une fois que vous avez créé un graphique en 3D, vous pouvez cliquer dessus et faire glisser votre souris pour le faire tourner et le voir sous n'importe quel angle.

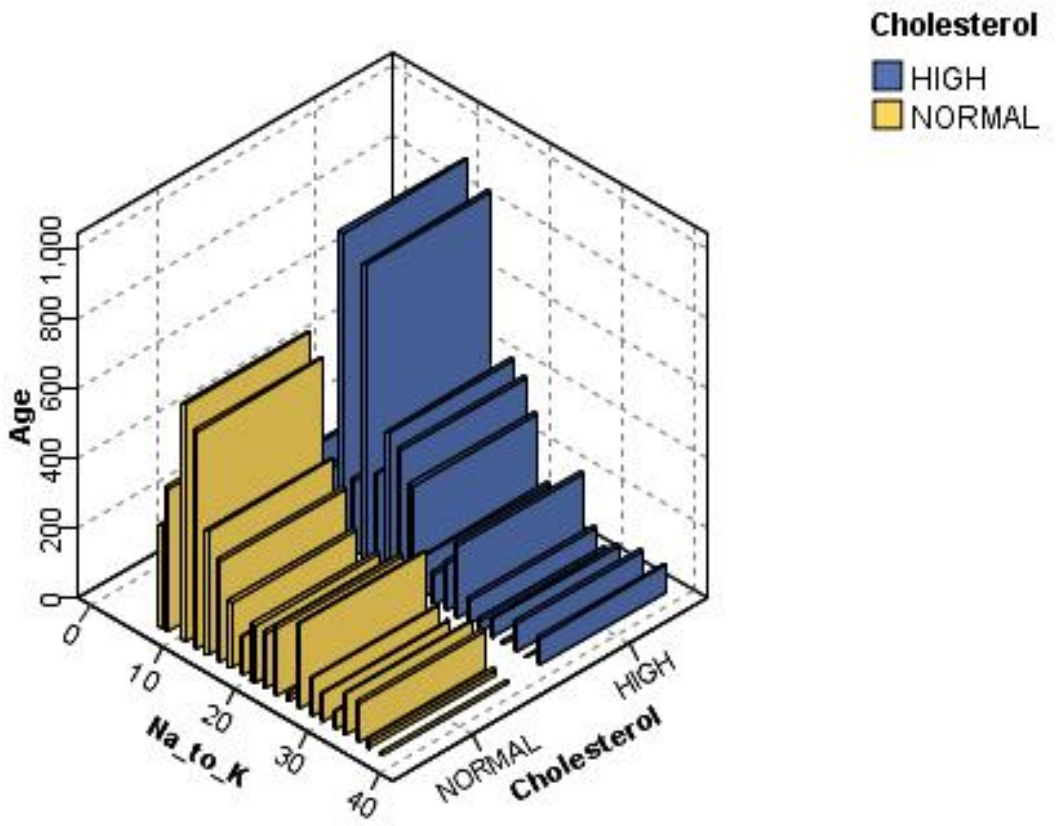


Figure 7. Graphique Résumé avec axes x, y et z

Il existe deux méthodes pour créer des graphiques en 3D dans IBM SPSS Modeler : représenter des informations sur un troisième axe (véritables graphiques en 3D) ou afficher le graphique avec un effet 3D. Ces deux méthodes sont disponibles pour les graphiques Nuage et Résumé.

Pour représenter des informations sur un troisième axe :

1. Dans la boîte de dialogue du noeud Graphiques, cliquez sur l'onglet **Nuage**.
2. Cliquez sur le bouton 3D afin d'activer les options de l'axe *z*.
3. Utilisez le sélecteur de champs pour sélectionner le champ de l'axe *z*. Dans certains cas, seuls les champs symboliques sont autorisés. Le sélecteur de champs affiche les champs appropriés.

Pour ajouter un effet 3D à un graphique :

1. Une fois le graphique créé, cliquez sur l'onglet **Graphiques** dans la fenêtre de sortie.
2. Cliquez sur le bouton 3D pour convertir la vue en un graphique en trois dimensions.

Noeud Représentation Graphique

Le noeud Représentation Graphique vous permet de choisir parmi de nombreuses sorties graphiques différentes (graphiques à barres, graphiques circulaires, histogrammes, nuages de points, cartes thermiques, etc.) dans un seul noeud. Dans le premier onglet, vous commencez par choisir les champs de données que vous souhaitez explorer, puis le noeud vous présente une sélection de types de graphiques fonctionnant pour vos données. Le noeud filtre automatiquement tous les types de graphiques ne fonctionnant pas avec les sélections de champs. Vous pouvez définir des options de graphiques détaillées ou plus avancées dans l'onglet Détaillées.

Remarque : Vous devez connecter le noeud Représentation Graphique à un flux avec des données afin d'éditer le noeud ou de sélectionner des types de graphiques.

Deux boutons vous permettent de contrôler les modèles de visualisation (et les feuilles de style et les cartes) disponibles :

Gérer. Gérer les modèles de visualisation, les feuilles de style et les cartes sur votre ordinateur. Vous pouvez importer, exporter, renommer et supprimer les modèles de visualisation, les feuilles de style et les cartes depuis votre ordinateur local. Pour plus d'informations, voir «Gérer les modèles, les feuilles de style et les fichiers cartes», à la page 229.

Emplacement. Modifier l'emplacement dans lequel les modèles de visualisation, les feuilles de style et les cartes sont stockés. L'emplacement actuel est indiqué à droite du bouton. Pour plus d'informations, voir la rubrique «Définition de l'emplacement des modèles, des feuilles de style et des cartes.», à la page 228.

Onglet de base de la représentation graphique

Si vous n'êtes pas certain du type de visualisation qui représenterait au mieux vos données, utilisez l'onglet de base. Quand vous sélectionnez vos données, un sous-ensemble de types de visualisations appropriées aux données vous est présenté. Pour des exemples, voir «Représentation graphiqueExemples», à la page 216.

1. Sélectionnez un ou plusieurs champs (variables) dans la liste. Pour sélectionner plusieurs champs, cliquez dessus tout en maintenant la touche Ctrl enfoncée.

   Remarquez que le niveau de mesure du champ détermine le type des visualisations disponibles. Vous pouvez modifier le niveau de mesure en cliquant avec le bouton droit de la souris sur le champ dans la liste et en choisissant une option. Pour plus d'informations sur les types de niveau de mesure disponibles, voir «Types de champ (variable)», à la page 204.
2. Sélectionnez un type de visualisation. Pour obtenir des descriptions des types disponibles, voir «Types de visualisation des Représentations graphiques intégrées disponibles», à la page 208.
3. Pour certaines visualisations, vous pouvez sélectionner une statistique récapitulative. Différents sous-ensembles de statistiques sont disponibles selon que les statistiques sont basées sur l'effectif ou calculées à partir d'un champ continu. Les statistiques disponibles dépendent aussi du modèle même. Une liste complète des statistiques qui peuvent être disponibles se trouve à la suite de l'étape suivante.
4. Si vous souhaitez définir davantage d'options, telles qu'un type esthétique facultatif et des champs de panel, cliquez sur **Détaillé**. Pour plus d'informations, voir «Représentation graphique Onglet Détaillé», à la page 206.

Statistiques récapitulatives calculées à partir d'un champ continu

- *Moyenne*. Mesure de la tendance centrale. Moyenne arithmétique : somme divisée par le nombre d'observations.
- *Médiane*. Valeur au-dessus et au-dessous de laquelle se trouvent la moitié des observations, le 50e percentile. Si le nombre de cellules est pair, la médiane correspond à la moyenne des deux cellules du milieu lorsqu'elles sont triées dans l'ordre croissant ou décroissant. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées.

- *Mode*. Valeur qui revient le plus souvent. Si plusieurs valeurs partagent la plus grande fréquence d'occurrence, chacune d'elles constitue un mode.
- *minimum*. La plus petite valeur d'une variable numérique.
- *maximum*. La plus grande valeur d'une variable numérique.
- *Intervalle*. Indique la différence entre les valeurs minimale et maximale.
- *Milieu de la plage*. Le milieu de la plage, c'est-à-dire la valeur dont la différence par rapport au minimum est égale à sa différence par rapport au maximum.
- *somme*. Somme ou total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.
- *Somme cumulative*. La somme cumulative des valeurs. Chaque élément graphique montre la somme d'un sous-groupe plus la somme totale de tous les groupes précédents.
- *Somme des pourcentages*. Le pourcentage dans chaque sous-groupe basé sur une zone totalisé par rapport à la somme dans tous les groupes.
- *Somme cumulative des pourcentages*. Pourcentage cumulé dans chaque sous-groupe basé sur un champ calculé comparé à la somme de tous les groupes. Chaque élément graphique montre le pourcentage d'un sous-groupe plus le pourcentage total de tous les groupes précédents.
- *Variance*. Mesure de dispersion autour de la moyenne, égal à la somme des écarts au carré par rapport à la moyenne divisé par le nombre d'observations moins un. La variance se mesure en unités, qui sont égales au carré des unités de la variable.
- *écart type*. Mesure de dispersion autour de la moyenne. Dans le cas d'une distribution normale, 68 % des observations se situent à l'intérieur d'un écart-type de la moyenne et 95 % se situent à l'intérieur de deux écarts-types. Par exemple, si l'âge moyen est 45 ans avec un écart type égal à 10, une distribution normale verra 95 % des observations se situer entre 25 et 65.
- *Erreur standard*. Mesure du degré de variation des valeurs de statistiques de test d'un échantillon à l'autre. Il s'agit de l'écart-type de la distribution de l'échantillon pour des statistiques. Par exemple, l'erreur standard de la moyenne est l'écart type des moyennes d'échantillon.
- *Kurtosis*. Mesure de l'importance des valeurs extrêmes. Dans le cas d'une distribution normale, la valeur de la statistique d'aplatissement est égale à zéro. Un kurtosis positif indique qu'on observe dans les données plus de valeurs extrêmes que dans une distribution normale. Une valeur négative indique que les données comportent moins de valeurs extrêmes qu'une distribution normale.
- *Asymétrie*. Mesure de l'asymétrie d'une distribution. La distribution normale est symétrique et présente une valeur de décalage de zéro. Une distribution avec un important décalage positif présente une longue queue vers la droite. Une distribution caractérisée par une importante asymétrie négative présente une extrémité gauche plus allongée. Pour simplifier, une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.

Les statistiques de région suivantes peuvent produire plus d'un élément graphique par sous-groupe. Lors de l'utilisation d'éléments graphiques d'intervalle, de région ou de périmètre, les statistiques régionales produisent un élément graphique affichant la plage. Tous les autres éléments graphiques résultent en deux éléments séparés, l'un montrant le début de l'intervalle et l'autre en montrant la fin.

- **Région : Plage**. Plage de valeurs entre les valeurs minimum et maximum.
- **Région : Intervalle de confiance de moyenne de 95 %**. Plage de valeurs qui, dans 95 % des cas, comprend la moyenne de la population.
- **Région : Intervalle de confiance de 95 % d'un individu**. Plage de valeurs qui, dans 95 % des cas, comprend la valeur prédite d'après l'observation individuelle.
- **Région : Ecart-type de 1 inférieur/supérieur à la moyenne**. Un intervalle de valeurs entre 1 *écart-type* au-dessus et en dessous de la *moyenne*.
- **Région : Erreur standard de 1 inférieure/supérieure à la moyenne**. Un intervalle de valeurs entre 1 *erreur standard* au-dessus et en dessous de la *moyenne*.

Statistiques récapitulatives d'après l'effectif

- **Comptage**. Nombre de lignes/d'observations.

- **Effectifs cumulés**. Nombre cumulé de lignes/cellules. Chaque élément graphique montre les effectifs d'un sous-groupe plus les effectifs totaux de tous les groupes précédents.
- **Pourcentage des effectifs**. Le pourcentage de lignes/cellules dans chaque sous-groupe par rapport au nombre total de lignes/cellules.
- **Pourcentage d'effectif cumulé**. Pourcentage cumulé de lignes/d'observations dans chaque sous-groupe comparé au nombre total de lignes/d'observations. Chaque élément graphique montre le pourcentage d'un sous-groupe plus le pourcentage total de tous les groupes précédents.

Types de champ (variable)

Des icônes s'affichent en regard des champs dans les listes de champs et indiquent le type de champ et le type de données. Les icônes indiquent aussi des jeux de réponses multiples.

Tableau 32. Icones de niveau de mesure.

Niveau de mesure	Numérique	Chaîne	Date	Heure
Continu		n/a		
Vecteur ordonné				
Définir				

Tableau 33. Icônes Ensemble à réponses multiples.

Type d'ensemble à réponses multiples	Icône
Jeu de réponses multiples - catégories multiples	
Jeu de réponses multiples - dichotomies multiples	

Niveau de mesure

Le niveau de mesure d'un champ est important lors de la création d'une visualisation. Vous trouverez ci-dessous une description des niveaux de mesure. Vous pouvez modifier le niveau de mesure de façon temporaire. Pour cela, cliquez sur un champ avec le bouton droit de la souris dans la liste de champs et choisissez une option. Dans la plupart des cas, vous ne devez prendre en compte que les deux classifications de champ les plus importantes, les qualitatives et les continues :

Qualitatives. Données possédant un nombre limité de valeurs ou de catégories distinctes (par exemple, le sexe ou la religion). Les champs catégoriel⌂s peuvent être des données chaîne (alphanumérique) ou des champs numériques qui utilisent des codes chiffrés pour représenter les catégories (par exemple, 0 = *Masculin* et 1 = *Féminin*). Elles sont parfois également qualifiées de données qualitatives. Les vecteurs, vecteurs ordonnés et indicateur⌂s sont tous des champs catégoriel⌂s.

- *Ensemble*. Champ/variable dont les valeurs représentent des modalités sans classement intrinsèque (par exemple, le service de la société dans lequel travaille un employé). La région, le code postal ou l'appartenance religieuse sont des exemples de variables nominales. Aussi nommé une variable nominale.
- *Ensemble ordonné*. Champ/variable dont les valeurs représentent des modalités associées à un classement intrinsèque (par exemple, des niveaux de satisfaction allant de Très mécontent à Très satisfait). Exemples de vecteurs ordonnés : des scores d'attitude représentant le degré de satisfaction ou de confiance, et des scores de classement des préférences. Aussi nommé une variable ordinale.
- *Indicateur*. Champ/variable avec deux valeurs différentes, telles que Yes et No, ou 1 et 2. Egalement appelé variable dichotomique ou binaire.

Continuer. Données mesurées sur une échelle d'intervalle ou une échelle de rapport, où les valeurs des données indiquent à la fois l'ordre des valeurs et la distance entre ces valeurs. Par exemple, un salaire de 72 195 dollars est supérieur à un salaire de 52 398 dollars, et la distance entre les deux valeurs est 19 797 dollars. Egalement appelées données quantitatives ou d'échelle ou données d'intervalle numérique.

Les champs catégoriel⌂s définissent des catégories de la visualisation, généralement pour dessiner des éléments graphiques distincts ou pour regrouper ces mêmes éléments. Les champs continus sont souvent récapitulés au sein des catégories de champs catégoriel⌂s. Par exemple, une visualisation par défaut présentant la variable Revenu pour les catégories de la variable Sexe afficherait le revenu moyen des hommes et le revenu moyen des femmes. Les valeurs brutes pour les champs continus peuvent également être tracées, comme dans un nuage de points. Par exemple, un nuage de points peut afficher le salaire actuel et le salaire d'embauche pour chaque observation. Un champ catégoriel⌂ peut être utilisé pour grouper les observations en fonction du sexe.

Types de données

Le niveau de mesure n'est pas la seule propriété d'un champ qui détermine son type. Un champ est aussi stocké en tant que type de données spécifique. Parmi les types de données possibles, on trouve les chaînes (données non numériques telles que des lettres), les valeurs numériques (nombres réels) et les dates. A la différence du niveau de mesure, un type de données de champ ne peut pas être modifié temporairement. Vous devez modifier la façon dont les données sont stockées dans le jeu de données d'origine.

Jeux de réponses multiples

Certains fichiers de données prennent également en charge un type spécifique de « champ » nommé **jeu de réponses multiples**. Les jeux de réponses multiples ne sont pas réellement des « champs » au sens habituel du terme. Les jeux de réponses multiples utilisent plusieurs champs pour enregistrer les réponses à des questions auxquelles le répondant peut donner plusieurs réponses. Les jeux de réponses multiples sont traités comme des champs catégoriel⌂s, et la plupart des actions appliquées à ces derniers peuvent également l'être aux jeux de réponses multiples.

Les jeux de réponses multiples peuvent être des jeux de dichotomies multiples ou des jeux de catégories multiples.

Jeux de dichotomies multiples. Un ensemble de dichotomies multiples consiste généralement en plusieurs champs dichotomiques : des champs pouvant prendre deux valeurs uniquement comme oui/non, présent/absent, coché/décoché. Bien que les champs peuvent ne pas être strictement dichotomiques, tous les champs du vecteur sont codées de la même manière.

Par exemple, une enquête pose la question " Parmi les sources suivantes, quelles sont les plus fiables dans le domaine de l'information ? " et propose cinq réponses possibles. Le répondant peut indiquer

plusieurs choix en cochant la case située en regard de chaque proposition. Les cinq réponses deviennent cinq champs dans le fichier de données, codées par 0 pour *Non* (non sélectionné) et 1 pour *Oui* (sélectionné).

Jeux de catégories multiples. Un jeu de catégories multiples comporte plusieurs champs, tous codés de la même façon, souvent avec un grand nombre de catégories de réponses possibles. Par exemple, une enquête comporte la question " Nommez jusqu'à trois nationalités décrivant le mieux votre héritage ethnique ". Des centaines de réponses sont possibles, mais pour des questions de codification, la liste est limitée aux 40 nationalités les plus courantes, le reste étant relégué dans une catégorie " Autre ". Dans le fichier de données, les trois choix deviennent trois champs, chacun comportant 41 catégories (40 nationalités codées et une catégorie « autre »).

Représentation graphique Onglet Détaillé

Utilisez l'onglet détaillé lorsque vous savez quel type de visualisation vous souhaitez créer ou si vous souhaitez ajouter des types esthétiques facultatifs, des panels et/ou une animation à la visualisation. Pour des exemples, voir «Représentation graphiqueExemples», à la page 216.

1. Si vous avez sélectionné un type de visualisation dans l'onglet de base, celui-ci est affiché. Sinon, sélectionnez-en un dans la liste déroulante. Pour plus d'informations sur les types de visualisation, voir «Types de visualisation des Représentations graphiques intégrées disponibles», à la page 208.
2. Juste à droite de la vignette de la visualisation se trouvent les contrôle⌂s permettant de spécifier les champs (variables) nécessaires au type de visualisation. Vous devez spécifier l'intégralité de ces champs.
3. Pour certaines visualisations, vous pouvez sélectionner une statistique récapitulative. Dans certains cas (tels que les graphiques à barres⌂), vous pouvez utiliser l'une de ces options récapitulatives pour le type transparent d'esthétique. Pour obtenir des descriptions des statistiques récapitulatives, voir «Onglet de base de la représentation graphique», à la page 202.
4. Vous pouvez sélectionner un ou plusieurs types esthétiques facultatifs. Ils peuvent ajouter la dimensionnalité⌂ en vous permettant d'inclure d'autres champs dans la visualisation. Par exemple, vous pouvez utiliser un champ pour faire varier la taille des points dans un nuage de points. Pour plus d'informations sur les apparences facultatives, voir «Apparences, superpositions, panneaux et animation», à la page 198. Notez que le type transparent d'esthétique n'est pas pris en charge par les scripts.
5. Si vous créez une visualisation de carte, le groupe **Fichiers cartes** affiche le ou les fichiers cartes qui seront utilisés. S'il existe un fichier carte par défaut, ce fichier apparaît. Pour modifier le fichier carte, cliquez sur **Sélectionnez un fichier carte** pour afficher la boîte de dialogue Sélectionner les cartes. Vous pouvez également spécifier le fichier carte par défaut dans cette boîte de dialogue. Pour plus d'informations, voir «Sélection des fichiers cartes pour les visualisations de carte».
6. Vous pouvez sélectionner une ou plusieurs options de division en panneaux ou d'animation. Pour plus d'informations sur les options de création de panneaux et d'animation, reportez-vous à la section «Apparences, superpositions, panneaux et animation», à la page 198.

Sélection des fichiers cartes pour les visualisations de carte

Si vous sélectionnez le modèle de visualisation de carte, il vous faut un fichier carte qui définit les informations géographiques permettant de dessiner la carte. Si un fichier carte par défaut existe, il sera utilisé pour la visualisation de carte. Pour choisir un autre fichier carte, cliquez sur **Sélectionnez un fichier carte** pour afficher l'onglet Détaillé de la boîte de dialogue Sélectionner les cartes.

La boîte de dialogue Sélectionner les cartes vous permet de choisir un fichier carte principal et un fichier carte de référence. Les fichiers cartes définissent les informations géographiques permettant de dessiner la carte. Votre application est installée avec un ensemble de fichiers cartes standard. S'il existe d'autres fichiers de formes ESRI que vous souhaitez utiliser, vous devez d'abord convertir ces fichiers en fichiers SMZ. Pour plus d'informations, voir «Conversion et distribution des fichiers de formes Carte», à la page 230

230. Après avoir converti la carte, cliquez sur **Gérer...** dans la boîte de dialogue Sélecteur de modèles pour importer la carte dans le système de gestion afin qu'il soit disponible dans la boîte de dialogue Sélectionner les cartes.

Tenez compte des points suivants lors de la spécification des fichiers cartes :

- Tous les modèles de carte ont besoin d'au moins un fichier carte.
- Le fichier carte relie généralement un attribut de clé de carte à la clé de données.
- Si le modèle ne nécessite aucune clé de carte reliée à une clé de données, il nécessite un fichier de carte de référence et des champs qui spécifient les coordonnées (comme la longitude et la latitude) pour dessiner les éléments sur la carte de référence.
- Les modèles de carte en superposition nécessitent deux cartes : un fichier carte principal et un fichier carte de référence. La carte de référence est d'abord dessinée afin qu'elle se trouve derrière le fichier carte principal.

Pour des informations sur la terminologie des cartes comme les attributs et les fonctions, voir «Concepts principaux des cartes», à la page 231.

Fichier de mappe. Vous pouvez sélectionner n'importe quel fichier carte se trouvant dans le système de gestion. Il contient des fichiers cartes préinstallés et les fichiers cartes que vous avez importés. Pour plus d'informations sur la gestion des fichiers cartes, voir «Gérer les modèles, les feuilles de style et les fichiers cartes», à la page 229.

Légende de la carte. Spécifiez l'attribut à utiliser comme clé qui relie le fichier carte à la clé de données.

Enregistrez le fichier carte et les paramètres par défaut. Sélectionnez cette case si vous souhaitez utiliser le fichier carte sélectionné par défaut. Si vous avez spécifié un fichier carte par défaut, il est inutile de spécifier un fichier carte à chaque fois que vous créez une visualisation de carte.

Légende des données. Ce contrôle répertorie la même valeur que celle qui apparaît dans l'onglet Détaillé du sélecteur de modèles. Elle apparaît ici au cas où vous auriez besoin de modifier la clé en raison du fichier carte spécifique que vous avez choisi.

Afficher toutes les fonctions des cartes dans la visualisation. Lorsque cette option est sélectionnée, toutes les fonctions de la carte apparaissent dans la visualisation même si aucune valeur de clé de données correspondante n'existe. Si vous souhaitez uniquement voir les fonctions pour lesquelles vous avez des données, désélectionnez cette option. Les fonctions identifiées par les clés de carte affichées dans la liste **Clés de carte non correspondantes** n'apparaîtront pas dans la visualisation.

Comparer les valeurs de la carte et de données. La clé de carte et la clé de données sont reliées l'une à l'autre pour créer la visualisation de carte. La clé de carte et la clé de données doivent être tirées du même domaine (par exemple, pays et régions). Cliquez sur **Comparer** pour vérifier si les valeurs de clé de données et de clé de carte correspondent. L'icône qui apparaît vous informe de l'état de la comparaison. Ces icônes sont décrites ci-dessous. Si une comparaison a été effectuée et qu'il existe des valeurs de clé de données sans valeurs de clé de carte correspondantes, les valeurs de clé de données apparaissent dans la liste **Clés de données non correspondantes**. Dans la liste **Clés de carte non correspondantes**, vous pouvez également voir les valeurs de clés de carte n'ayant pas de valeurs de clés de données correspondantes. Si **Afficher toutes les fonctions sur des cartes dans la visualisation** n'est pas coché, les fonctions identifiées par ces valeurs de clés de carte ne seront pas rendues.

Tableau 34. Icônes de comparaison.

Icône	Description
	Aucune comparaison n'a été effectuée. Il s'agit de l'état par défaut avant de cliquer sur **Comparer**. Vous devez être prudent car vous ne savez pas si les valeurs de la clé de données et de la clé de carte correspondent.

Tableau 34. Icônes de comparaison (suite).

Icône	Description
	Une comparaison a été effectuée et les valeurs de la clé de données et de la clé de carte sont en correspondance parfaite. Pour chaque valeur de la clé de données, il existe une fonction correspondante identifiée par la clé de carte.
	Une comparaison a été effectuée et certaines valeurs de la clé de données et de la clé de carte ne correspondent pas. Pour certaines des valeurs de la clé de données, il n'existe pas de fonction correspondante identifiée par la clé de carte. Vous devez procéder avec prudence. Si vous continuez, la visualisation n'inclura pas toutes les valeurs de données.
	Une comparaison a été effectuée et aucune des valeurs de la clé de données et de la clé de carte ne correspond. Vous devez choisir une autre clé de données ou une autre clé de carte car aucune carte ne sera représentée si vous continuez.

Types de visualisation des Représentations graphiques intégrées disponibles

Vous pouvez créer plusieurs types différents de visualisations. Tous les types intégrés suivants sont disponibles dans les onglets de base et détaillé. Certaines des descriptions de ces modèles (particulièrement les modèles de cartes) identifient les champs (variables) spécifiés sur l'onglet Détaillé à l'aide de **texte spécial**.

Tableau 35. Types de graphiques disponibles.

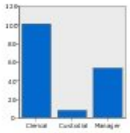
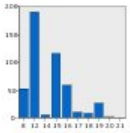
Icône Graphique	Description	Icône Graphique	Description
	**Barre** Calcule une statistique récapitulative d'un champ numérique continu et affiche les résultats de chaque catégorie d'un champ catégoriel⌂ sous la forme de barres. *Nécessite* : un champ catégoriel et un champ continu.		**Barres d'effectifs** Affiche la proportion de lignes/d'observations dans chaque catégorie d'un champ catégoriel⌂ sous la forme de barres. Vous pouvez aussi utiliser le noeud de graphique de distribution pour générer ce graphique. Ce noeud propose quelques options supplémentaires. Pour plus d'informations, voir «Noeud Proportion», à la page 250. *Nécessite* : un seul champ catégoriel.

Tableau 35. Types de graphiques disponibles (suite).

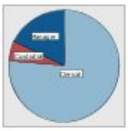
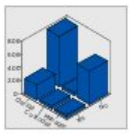
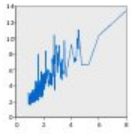
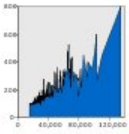
Icône Graphique	Description	Icône Graphique	Description
	**Graphique circulaire** Calcule la somme d'un champ numérique continu et affiche la proportion de cette somme distribuée dans chaque catégorie d'un champ catégoriel⌂ sous la forme de tranches d'un graphique circulaire. *Nécessite* : un champ catégoriel et un champ continu.		**Secteur d'effectifs** Affiche la proportion de lignes/d'observations dans chaque catégorie d'un champ catégoriel⌂ sous la forme de tranches d'un graphique circulaire. *Nécessite* : un seul champ catégoriel.
	**Barres 3D** Calcule une statistique récapitulative d'un champ numérique continu et affiche les résultats de l'intersection de catégories de deux champs catégoriel⌂s. *Nécessite* : une paire de champs catégoriels et un champ continu.		**Graphique circulaire 3-D** Il est identique au graphique circulaire à l'exception de l'effet 3-D supplémentaire. *Nécessite* : un champ catégoriel et un champ continu.
	**Line** Calcule une statistique récapitulative pour un champ pour chaque valeur d'un autre champ et trace une ligne reliant les valeurs. Vous pouvez aussi utiliser le noeud d'un graphique tracé pour générer un tracé en ligne. Ce noeud propose quelques options supplémentaires. Pour plus d'informations, voir «Noeud Nuage», à la page 237. *Nécessite* : une paire de champs de chaque type.		**Zone** Calcule une statistique récapitulative pour un champ pour chaque valeur d'un autre champ et trace une surface reliant les valeurs. La différence entre une ligne et une aire est minime du fait qu'une aire ressemble à une ligne dont la zone située en dessous est coloriée. Cependant, si vous utilisez un type esthétique en couleur, le résultat est une scission de la courbe et un empilement de la surface. *Nécessite* : une paire de champs de chaque type.

Tableau 35. Types de graphiques disponibles (suite).

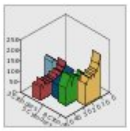
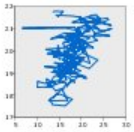
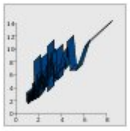
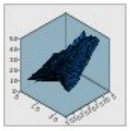
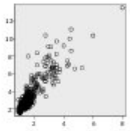
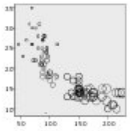
Icône Graphique	Description	Icône Graphique	Description
	**Aire 3D** Affiche les valeurs d'un champ tracé par rapport aux valeurs d'un autre, et le découpe par champ catégoriel⌂. Un élément de surface est tracé pour chaque catégorie. *Nécessite* : un champ catégoriel et une paire de champs de chaque type.		**Path** Affiche les valeurs d'un champ tracées par rapport aux valeurs d'un autre champ, sous la forme d'une courbe reliant les valeurs dans l'ordre où elles apparaissent dans le jeu de données d'origine. L'ordre est la principale différence entre un chemin et une ligne. *Nécessite* : une paire de champs de chaque type.
	**Ribbon** Calcule une statistique récapitulative pour un champ pour chaque valeur d'un autre champ et trace un ruban reliant les valeurs. Un ruban est essentiellement une courbe dotée d'un effet 3-D. Il ne s'agit pas d'un véritable graphique 3-D. *Nécessite* : une paire de champs de chaque type.		**Surface** Affiche les valeurs de trois champs tracés par rapport à leurs valeurs réciproques, sous la forme d'une surface reliant les valeurs. *Nécessite* : trois champs de chaque type.
	**Graphique en nuage de points** Affiche les valeurs d'un champ tracées par rapport aux valeurs d'un autre champ. Ce graphique peut mettre en évidence la relation entre les champs (s'il en existe). Vous pouvez aussi utiliser le noeud d'un graphique tracé pour générer un nuage de points. Ce noeud propose quelques options supplémentaires. Pour plus d'informations, voir «Noeud Nuage», à la page 237. *Nécessite* : une paire de champs de chaque type.		**Graphique en bulles** Tout comme un nuage de points de base, il affiche les valeurs d'un champ tracées par rapport aux valeurs d'un autre champ. La différence tient en ce que les valeurs d'un troisième champ sont utilisées pour faire varier la taille de chaque point. *Nécessite* : trois champs de chaque type.

Tableau 35. Types de graphiques disponibles (suite).

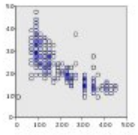
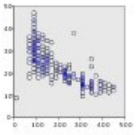
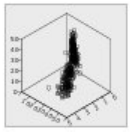
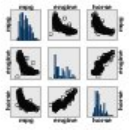
Icône Graphique	Description	Icône Graphique	Description
	**Nuage de points regroupé par casiers** Tout comme un nuage de points de base, il affiche les valeurs d'un champ tracées par rapport aux valeurs d'un autre champ. La différence tient en ce que des valeurs similaires sont regroupées et qu'un type esthétique de couleur ou de taille est utilisé pour indiquer le nombre d'observations de chaque groupe. *Nécessite* : une paire de champs continus.		**nuage de points regroupé par casiers hexagonaux** Voir la description des nuages de points groupés. La différence réside dans la forme des casiers sous-jacents qui ont la forme d'hexagones plutôt que de cercles. Le nuage de points en groupes hexagonaux résultant ressemble au nuage de points groupé. Cependant, le nombre de valeurs de chaque groupe est différent pour chaque graphique à cause de la forme des groupes sous-jacents. *Nécessite* : une paire de champs continus.
	**Nuage de points 3D** Affiche les valeurs de trois champs tracés les uns par rapport aux autres. Ce graphique peut mettre en évidence la relation entre les champs (s'il en existe). Vous pouvez aussi utiliser le noeud d'un graphique tracé pour générer un nuage de points en 3-D. Ce noeud propose quelques options supplémentaires. Pour plus d'informations, voir «Noeud Nuage», à la page 237. *Nécessite* : trois champs de chaque type.		**Matrice de nuage de points (SPLOM)** Affiche les valeurs d'un champ tracées par rapport aux valeurs d'un autre champ pour chaque champ. Une SPLOM est comparable à un tableau de diagrammes de dispersion. Le SPLOM comprend aussi un histogramme de chaque champ. *Nécessite* : au moins deux champs continus.

Tableau 35. Types de graphiques disponibles (suite).

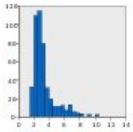
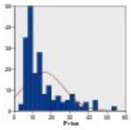
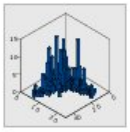
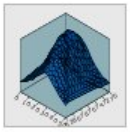
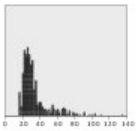
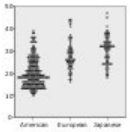
Icône Graphique	Description	Icône Graphique	Description
	**Histogram** Affiche la distribution d'effectifs d'un champ. Un histogramme peut aider à déterminer le type de distribution et voir si la distribution est asymétrique. Vous pouvez aussi utiliser le noeud de graphique d'histogramme pour générer ce graphique. Ce noeud propose quelques options supplémentaires. Pour plus d'informations, voir «Onglet Tracé d'histogramme», à la page 254. *Nécessite* : un champ unique de chaque type.		**Histogramme avec distribution normale** Affiche la distribution d'effectifs d'un champ continu avec une courbe surimposée de la distribution normale. *Nécessite* : un seul champ continu.
	**Histogramme 3D** Affiche la distribution d'effectifs d'une paire de champs continus. *Nécessite* : une paire de champs continus.		**Densité 3D** Affiche la distribution d'effectifs d'une paire de champs continus. Il est similaire à un histogramme 3-D, l'unique différence résidant dans le fait qu'une surface est utilisée à la place des barres pour afficher la distribution. *Nécessite* : une paire de champs continus.
	**Tracé de points** Affiche les observations/lignes individuelles et les empile aux points de données distincts de l'axe *x*. Ce graphique est similaire à un histogramme en ce fait qu'il affiche la distribution des données, mais il affiche chaque observation/ligne plutôt qu'un effectif agrégé d'un groupe spécifique (plage de valeurs). *Nécessite* : un champ unique de chaque type.		**Tracé de points 2-D** Affiche les observations/lignes individuelles et les empile aux points de données distincts de l'axe *y* pour chaque catégorie d'un champ catégoriel⌂. *Nécessite* : un champ catégoriel et un champ continu.

Tableau 35. Types de graphiques disponibles (suite).

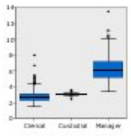
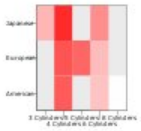
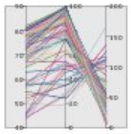
Icône Graphique	Description	Icône Graphique	Description
	**Boxplot** Calcule les cinq statistiques (minimum, premier quartile, médiane, troisième quartile et maximum) d'un champ continu pour chaque catégorie d'un champ catégoriel⌂. Les résultats s'affichent sous la forme d'éléments de boîte à moustaches ou de schéma. Les boîtes à moustaches peuvent vous aider à voir comment la distribution de données continues varie au sein des catégories. *Nécessite* : un champ catégoriel et un champ continu.		**Carte thermique** Calcule la moyenne d'un champ continu pour l'intersection de catégories de deux champs catégoriel⌂s. *Nécessite* : une paire de champs catégoriels et un champ continu.
	**Parallel** Crée des axes parallèles pour chaque champ et trace une ligne à travers la valeur de champ pour chaque ligne/observation des données. *Nécessite* : au moins deux champs continus.		**Choroplèthe d'effectifs** Calcule le total pour chaque catégorie de champ catégoriel (**Légende des données**) et dessine une qui utilise la saturation de couleur pour représenter les totaux dans les fonctions de carte qui correspondent aux catégories. *Nécessite* : un champ catégoriel. Un fichier carte dont la clé correspond aux catégories **Clé de données**.
	**Carte choroplèthe des moyennes/médianes/sommes** Calcule la moyenne/médiane ou la somme d'un champ continu (**Couleur**) pour chaque catégorie de champ catégoriel (**Légende des données**) et dessine une qui utilise la saturation de couleur pour représenter les statistiques calculées dans les fonctions de carte qui correspondent aux catégories. *Nécessite* : un champ catégoriel et un champ continu. Un fichier carte dont la clé correspond aux catégories **Clé de données**.		**Choroplèthe de valeurs** Dessine une carte qui utilise une couleur pour représenter les valeurs d'un champ catégoriel (**Couleur**) pour les fonctions de carte qui correspondent aux valeurs définies par un autre champ catégoriel (**Légende des données**). S'il existe plusieurs valeurs qualitatives du champ Couleur pour chaque fonction, la valeur modale est utilisée. *Nécessite* : une paire de champs catégoriels. Un fichier carte dont la clé correspond aux catégories **Clé de données**.

Tableau 35. Types de graphiques disponibles (suite).

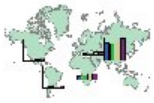
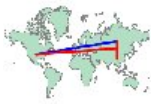
Icône Graphique	Description	Icône Graphique	Description
	**Coordonnées sur une choroplèthe d'effectifs** Similaire au Choroplèthe d'effectifs, excepté qu'il y a deux champs continus supplémentaires (**Longitude** et **Latitude**) qui identifient les coordonnées pour le tracé des points sur la Choroplèthe. *Nécessite* : un champ catégoriel et une paire de champs continus. Un fichier carte dont la clé correspond aux catégories **Clé de données**.		**Coordonnées sur une carte choroplèthe des moyennes/médianes/sommes** Similaire au Choroplèthe des moyennes/médianes/sommes, excepté qu'il y a deux champs continus supplémentaires (**Longitude** et **Latitude**) qui identifient les coordonnées pour le tracé des points sur la Choroplèthe. *Nécessite* : un champ catégoriel et trois champs continus. Un fichier carte dont la clé correspond aux catégories **Clé de données**.
	**Coordonnées sur une Choroplèthe de valeurs** Similaire au Choroplèthe de valeurs, excepté qu'il y a deux champs continus supplémentaires (**Longitude** et **Latitude**) qui identifient les coordonnées pour le tracé des points sur la Choroplèthe. *Nécessite* : une paire de champs catégoriels et une paire dechamps continus. Un fichier carte dont la clé correspond aux catégories **Clé de données**.		**Barres d'effectifs sur une carte** Calcule la proportion de lignes/observations dans chaque catégorie d'un champ catégoriel⌂ (**Catégories**) pour chaque fonction de carte (**Clé de données**) et dessine une carte et les graphiques à barres⌂ au centre de chaque fonction de carte. *Nécessite* : une paire de champs catégoriels. Un fichier carte dont la clé correspond aux catégories **Clé de données**.
	**Barres sur une carte** Calcule une statistique récapitulative pour un champ continu (**Valeurs**) et affiche les résultats pour chacune des catégories d'un champ catégoriel (**Catégories**) pour chaque fonction de carte (**Légende des données**) en tant que graphiques à barres positionnées au centre de chaque fonction de carte. *Nécessite* : une paire de champs catégoriels et un champ continu. Un fichier carte dont la clé correspond aux catégories **Clé de données**.		**Graphique circulaire d'effectifs sur une carte** Affiche la proportion de lignes/observations dans chaque catégorie d'un champ catégoriel⌂ (**Catégories**) pour chaque fonction de carte (**Clé de données**) et dessine une carte et les proportions sous forme de tranches. *Nécessite* : une paire de champs catégoriels. Un fichier carte dont la clé correspond aux catégories **Clé de données**.

Tableau 35. Types de graphiques disponibles (suite).

Icône Graphique	Description	Icône Graphique	Description
	**Graphique circulaire sur une carte** Calcule la somme d'Affiche la somme d'un champ continu (**Valeurs**) dans chaque catégorie d'un champ catégoriel (**Catégories**) pour chaque fonction de carte (**Clé de données**) et dessine une carte et les sommes sous forme de tranches d'un graphique circulaire au centre de chaque fonction de carte. *Nécessite* : une paire de champs catégoriels et un champ continu. Un fichier carte dont la clé correspond aux catégories **Clé de données**.		**Graphique curviligne sur une carte** Calcule une statistique récapitulative pour un champ continu (**Y**) pour chaque valeur d'un autre champ (**X**) pour chaque fonction de carte (**Clé de données**) et dessine une carte et des graphiques curvilignes qui relient les valeurs au centre de chaque fonction de carte. *Nécessite* : un champ catégoriel et une paire de champs de chaque type. Un fichier carte dont la clé correspond aux catégories **Clé de données**.
	**Coordonnées sur une carte de référence** Dessine une carte et des points avec des champs continus (**Longitude** et **Latitude**) qui identifient les coordonnées pour les points. *Nécessite* : une paire de champs d'intervalle. Un fichier carte.		**Flèches sur une carte de référence** Dessine une carte et des flèche avec des champs continus qui identifient les points de départ (**Longitude de départ** et **Latitude de départ**) et les points de fin (**Longitude de fin** et **Latitude de fin**) pour chaque flèche. Chaque enregistrement/observation dans les résultats de données dans une flèche de la carte. *Nécessite* : quatre champs continus. Un fichier carte.
	**Carte de superposition de points** Dessine une carte de référence et la superpose à une autre carte à points avec les points en couleur en fonction du champ catégoriel⌂ (**Couleur**). *Nécessite* : une paire de champs catégoriels. Un fichier carte avec points dont la clé correspond aux catégories **Clé de données**. Un fichier carte de référence.		**Carte de superposition de polygone** Dessine une carte de référence et la superpose à une autre carte polygone avec les polygones en couleur en fonction du champ catégoriel⌂ (**Couleur**). *Nécessite* : une paire de champs catégoriels. Un fichier carte de polygone dont la clé correspond aux catégories **Clé de données**. Un fichier carte de référence.

Tableau 35. Types de graphiques disponibles (suite).

Icône Graphique	Description	Icône Graphique	Description
	**Carte de superposition de ligne** Dessine une carte de référence et la superpose à une autre carte curviligne avec les lignes en couleur en fonction du champ catégoriel⌂ (**Couleur**). *Nécessite* : une paire de champs catégoriels. Un fichier carte avec lignes dont la clé correspond aux catégories **Clé de données**. Un fichier carte de référence.		

Création de visualisations de carte

Pour de nombreuses visualisations, vous n'avez que deux choix à faire : les champs (variables) souhaités et un modèle pour visualiser ces champs. Aucun choix ou action supplémentaire n'est nécessaire. Les visualisations de carte nécessitent au moins une étape supplémentaire : sélectionnez un fichier carte qui définit les informations géographiques pour la visualisation de carte.

Les étapes de base pour créer une carte simple sont les suivantes :

1. Sélectionnez les champs souhaités dans l'onglet de base. Pour des informations sur le type et le nombre de champs nécessaires aux différentes visualisations de carte, voir «Types de visualisation des Représentations graphiques intégrées disponibles», à la page 208.
2. Sélectionnez un modèle de carte.
3. Cliquez sur l'onglet Détaillé.
4. Vérifiez que la **Clé de données** et les autres listes déroulantes nécessaires sont définies sur les champs appropriés.
5. Dans le groupe Fichiers cartes, cliquez sur **Sélectionner un fichier carte**.
6. Utilisez la boîte de dialogue Sélectionner les cartes pour choisir le fichier carte et la clé de carte. Les valeurs de la clé de carte doivent correspondre aux valeurs du champ spécifié par la **clé de données**. Vous pouvez utiliser le bouton **Comparer** pour comparer ces valeurs. Si vous sélectionnez le modèle de carte superposée, vous aurez également besoin de choisir une carte de référence. La carte de référence ne se trouve pas dans les données. Elle sert d'arrière-plan pour la carte principale. Pour des informations supplémentaires sur la boîte de dialogue Sélectionner les cartes, voir «Sélection des fichiers cartes pour les visualisations de carte», à la page 206.
7. Cliquez sur **OK** pour fermer la boîte de dialogue Sélectionner les cartes.
8. Dans le Sélecteur de modèles de représentations graphiques, cliquez sur **Exécuter** pour créer la visualisation de carte.

Représentation graphiqueExemples

Cette section comporte plusieurs exemples différents pour faire une démonstration des options disponibles. Les exemples présentent également des informations relatives à l'interprétation des visualisations finales.

Ces exemples utilisent le flux intitulé *graphboard.str* qui fait référence aux fichiers de données *employee_data.sav*, *customer_subset.sav* et *worldsales.sav*. Ces fichiers sont disponibles à partir du dossier *Demos* de n'importe quelle installation du client IBM SPSS Modeler. Ce dossier est accessible à partir du groupe de programmes IBM SPSS Modeler dans le menu Démarrer de Windows. Le fichier *graphboard.str* se trouve dans le dossier *streams*.

Nous vous conseillons de lire les exemples dans leur ordre de présentation. Les exemples postérieurs s'appuient sur les précédents.

Exemple : Graphique à barres en cluster avec statistique récapitulative

Nous allons créer un diagramme à barres qui résume un champ ou une variable numérique continue pour chaque catégorie d'une variable d'ensemble/catégorielle. En particulier, nous allons créer un graphique à barres⌂ qui indique le salaire moyen des hommes et des femmes.

Cet exemple ainsi que plusieurs des exemples suivants utilisent *Employee data*, un jeu de données fictif contenant des informations sur les employés d'une société.

1. Ajoutez un noeud source Statistics qui pointe vers *employee_data.sav*.
2. Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
3. Dans l'onglet Base, sélectionnez *Sexe* et *Salaire actuel* . (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)
4. Sélectionnez **Barres**.
5. Dans la liste déroulante Récapitulatif, sélectionnez **Moyenne**.
6. Cliquez sur **Exécuter**.
7. Sur l'écran qui s'affiche, cliquez sur le bouton « Afficher les libellés de champ et de valeur » de la barre d'outils (le second bouton du groupe de deux situé au centre de la barre d'outils).

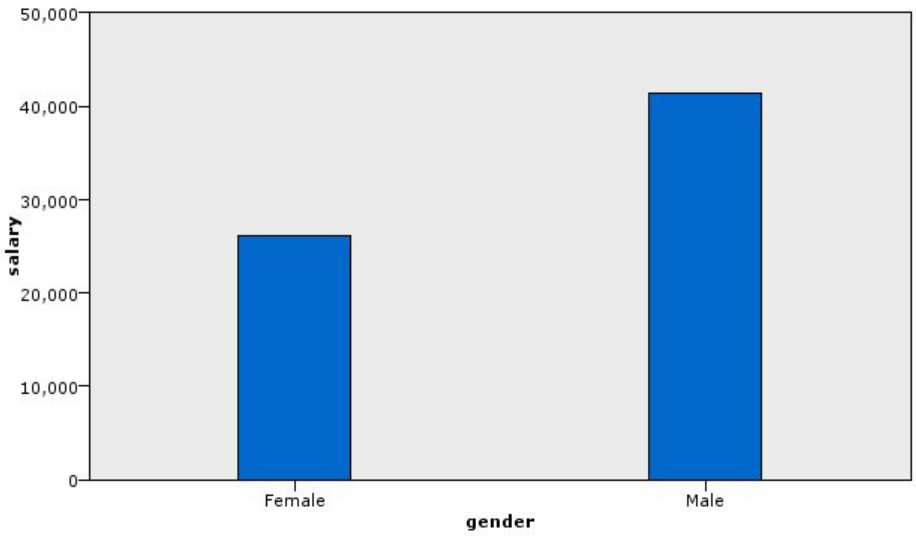


Figure 8. Diagramme à barres avec statistique récapitulative

Nous pouvons remarquer que :

- En fonction de la hauteur des barres, il est clair que le salaire moyen des hommes et supérieur au salaire moyen des femmes.

Exemple : Graphique à barres empilées avec statistique récapitulative

A présent, nous allons créer un graphique à barres empilées pour voir si la différence de salaire moyen entre les hommes et les femmes dépend du type de travail. Les femmes gagnent peut-être plus que les hommes, en moyenne, pour certains types d'emploi.

Remarque : Cet exemple utilise *Employee data*.

1. Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
2. Dans l'onglet Basique, sélectionnez *Employment Category* et *Current Salary*. (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)

3. Sélectionnez **Barres**.
4. Dans la liste Récapitulatif, sélectionnez **Moyenne**.
5. Cliquez sur l'onglet Détaillé. Remarque : vos sélections sur l'onglet précédent se reflètent ici.
6. Dans le groupe Apparences en option, choisissez *sexe* dans la liste déroulante Couleur.
7. Cliquez sur **Exécuter**.

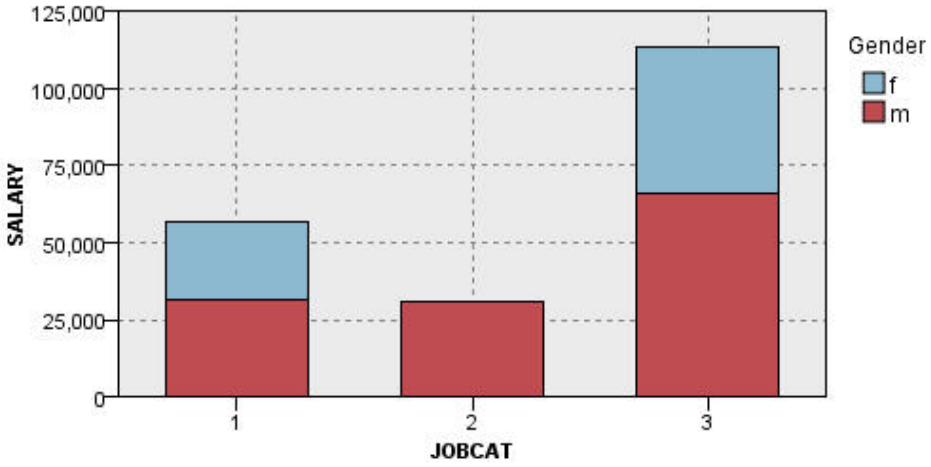


Figure 9. Graphique à barres empilées

Nous pouvons remarquer que :

- La différence entre les salaires moyens pour chaque type d'emploi ne semble pas aussi importante qu'elle l'était dans le graphique à barres⌂ qui comparait les salaires moyens de tous les hommes et femmes. Peut-être y a-t-il un nombre variable d'hommes et de femmes dans chaque groupe. Vous pourriez le vérifier en créant un graphique à barres⌂ de nombres.
- Quelque soit le type d'emploi, le salaire moyen des hommes est toujours supérieur au salaire moyen des femmes.

Exemple : Histogramme panélisé

Nous allons créer un histogramme panélisé par sexe afin de pouvoir comparer les distributions des fréquences de salaire pour les hommes et les femmes. La distribution des fréquences indique combien d'observations/lignes se trouvent à l'intérieur de plages de salaire spécifiques. L'histogramme panélisé peut nous aider à analyser plus en détails la différence de salaire entre les sexes.

Remarque : Cet exemple utilise *Employee data*.

1. Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
2. Dans l'onglet Base, sélectionnez *Salaire actuel* .
3. Sélectionnez **Histogramme**.
4. Cliquez sur l'onglet Détaillé.
5. Dans le groupe Panneaux et Animation, choisissez *sexe* dans la liste déroulante Panel Across.
6. Cliquez sur **Exécuter**.

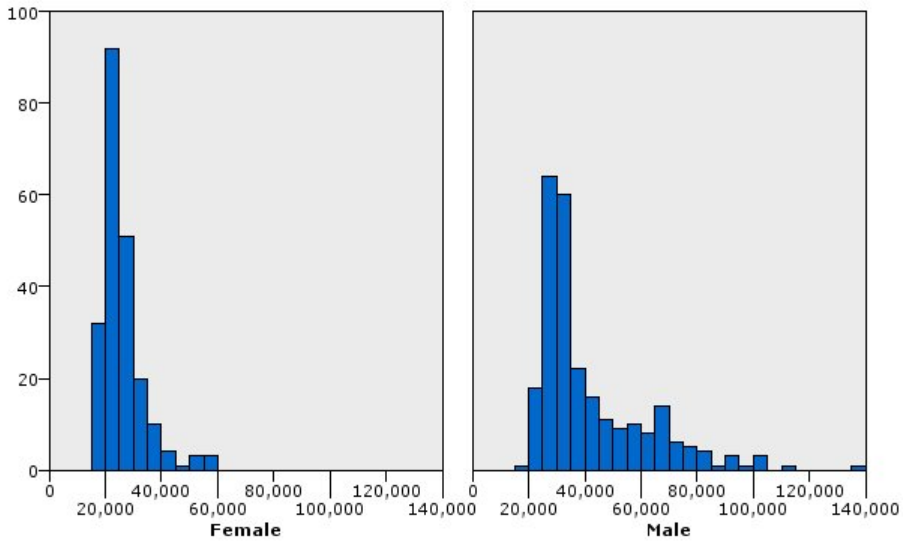


Figure 10. Histogramme panélisé

Nous pouvons remarquer que :

- Aucune distribution de fréquences n'est une distribution normale. Autrement dit, les histogrammes ne ressemblent pas à des courbes en cloche, comme ce serait le cas si les données étaient distribuées normalement.
- Les barres les plus hautes sont situées sur le côté gauche de chaque graphique. Par conséquent, les hommes comme les femmes sont plus nombreux à avoir des salaires plus bas que des salaires plus élevés.
- Les distributions des fréquences de salaire parmi les hommes et les femmes ne sont pas égales. Observez la forme des histogrammes. Il y a plus d'hommes à avoir des salaires plus élevés que de femmes à avoir des salaires plus élevés.

Exemple : Graphique en points panélisé

De même qu'un histogramme, un tracé de points indique la distribution d'une plage numérique continue. A la différence d'un histogramme, qui indique les nombres de plages de données mis en plages, un tracé de points montre toutes les lignes/observations contenues dans les données. Par conséquent, un tracé de points offre une granularité supplémentaire comparé à l'histogramme. En fait, l'utilisation d'un tracé de points peut constituer le point de départ privilégié lors de l'analyse des distributions des fréquences.

Remarque : Cet exemple utilise *Employee data*.

1. Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
2. Dans l'onglet Base, sélectionnez *Salaire actuel* .
3. Sélectionnez **Tracé en points**.
4. Cliquez sur l'onglet Détaillé.
5. Dans le groupe Panneaux et Animation, choisissez *sexe* dans la liste déroulante Panel Across.
6. Cliquez sur **Exécuter**.
7. Agrandissez la fenêtre de sortie afin de voir le tracé plus distinctement.

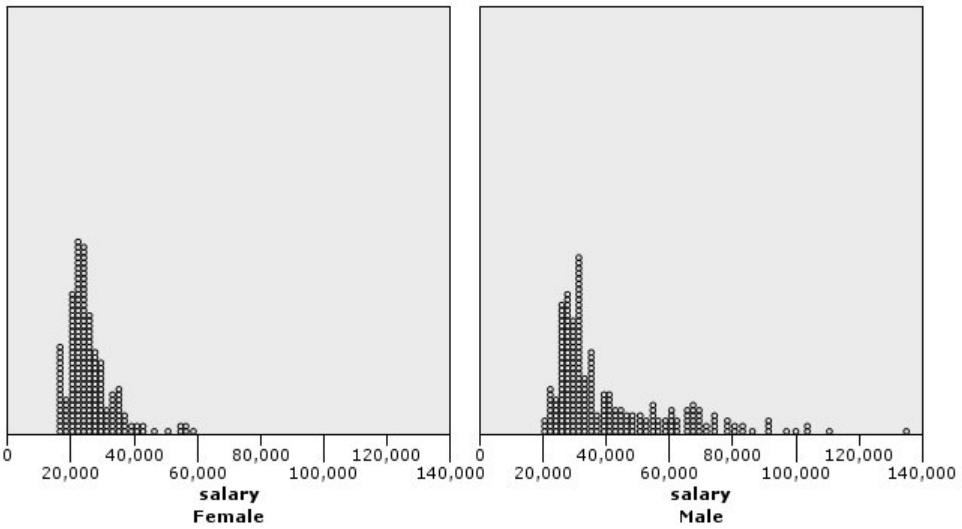


Figure 11. Tracé en points panélisé

Comparé à l'histogramme (voir «Exemple : Histogramme panélisé», à la page 218), nous pouvons observer les éléments suivants :

- Le plus haut niveau à 20 000 qui apparaissait dans l'histogramme pour les femmes est moins sensible dans le tracé de points. Il existe plusieurs observations/lignes concentrées autour de cette valeur, mais la plupart de ces valeurs sont plus proches de 25 000. Ce niveau de granularité n'est pas apparent dans l'histogramme.
- Bien que l'histogramme des hommes suggère que le salaire moyen des hommes diminue progressivement après 40 000, le graphique à points montre que la distribution est plutôt uniforme après cette valeur, jusqu'à 80 000. A chaque valeur de salaire dans cet intervalle, il existe trois hommes ou plus qui gagnent ce salaire en particulier.

Exemple : Boîte à moustaches

Une boîte à moustaches est un autre graphique utile pour la visualisation de la distribution des données. Une boîte à moustaches contient plusieurs mesures statistiques que nous allons explorer après avoir créé la visualisation.

Remarque : Cet exemple utilise *Employee data*.

1. Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
2. Dans l'onglet Base, sélectionnez *Sexe* et *Salaire actuel* . (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)
3. Sélectionnez **Boîte à moustaches**.
4. Cliquez sur **Exécuter**.

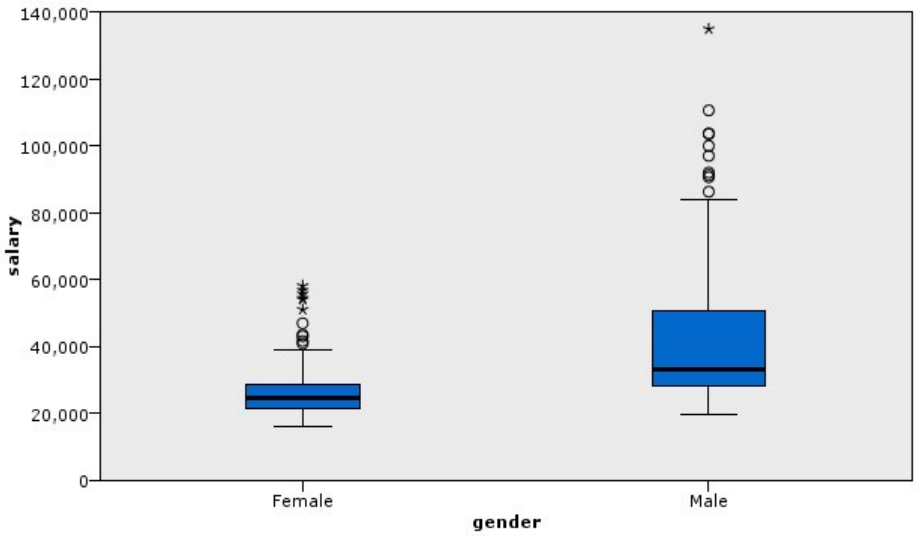


Figure 12. Boîte à moustaches

Etudions les différentes parties de la boîte à moustaches :

- La ligne sombre au milieu des boîtes est la médiane du *salaire*. La moitié des observations /lignes a une valeur supérieure à la médiane, et la moitié a une valeur inférieure. Comme la moyenne, la médiane est une mesure de la tendance centrale. Contrairement à la moyenne, elle est moins influencée par les observations/lignes avec des valeurs extrêmes. Dans cet exemple, la médiane est inférieure à la moyenne (comparez avec «Exemple : Graphique à barres en cluster avec statistique récapitulative», à la page 217). Le fait que la moyenne et la médiane soient différentes indique qu'il existe quelques observations/lignes avec des valeurs extrêmes qui élèvent la moyenne. Autrement dit, il existe quelques employés qui gagnent des salaires élevés.
- Le bas de la boîte indique le 25ème centile. Vingt-cinq pour cent des observations/lignes ont des valeurs au-dessous du 25ème percentile. Le haut de la boîte indique le 75ème centile. Vingt-cinq pour cent des observations/lignes ont des valeurs au-dessus du 75ème percentile. Cela signifie que 50 % des observations/lignes sont situées dans la boîte. La boîte est beaucoup plus petite pour les femmes que pour les hommes. Ceci indique que le *salaire* varie moins pour les femmes que pour les hommes. Le haut et le bas de la boîte sont souvent appelés **charnières**.
- Les barres en T qui partent des boîtes sont appelées **limites internes** ou **moustaches**. Elles s'étendent jusqu'à 1,5 fois la hauteur de la zone ou, si aucune observation/ligne a une valeur comprise dans cet intervalle, jusqu'aux valeurs minimum ou maximum. Si les données sont distribuées normalement, environ 95 % des données doivent être situées entre les limites internes. Dans cet exemple, les limites internes sont moins étendues pour les femmes que pour les hommes, ce qui constitue une autre indication que le *salaire* varie moins pour les femmes que pour les hommes.
- Les points sont des **valeurs extrêmes**. Il s'agit de valeurs qui n'entrent pas dans les limites internes. Les valeurs extrêmes sont des valeurs aberrantes. Les astérisques ou les étoiles sont des **valeurs extrêmes et éloignées**. Elles représentent des observations/lignes qui ont des valeurs égales à plus de trois fois la hauteur des boîtes. Il existe plusieurs valeurs extrêmes pour les femmes et les hommes. Souvenez-vous que la moyenne est supérieure à la médiane. La moyenne est plus élevée du fait de ces valeurs extrêmes.

Exemple : Graphique circulaire

Nous allons désormais utiliser un jeu de données différent pour explorer d'autres types de visualisation. Le jeu de données est *customer_subset*, un fichier de données fictif contenant des informations sur des clients.

Nous allons d'abord créer un graphique circulaire pour vérifier la proportion de clients résidant dans différentes régions géographiques.

1. Ajoutez un noeud source Statistics qui pointe vers *customer_subset.sav*.
2. Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
3. Dans l'onglet Base, sélectionnez >*Indicateur géographique* .
4. Sélectionnez **Graphique circulaire de nombres**.
5. Cliquez sur **Exécuter**.

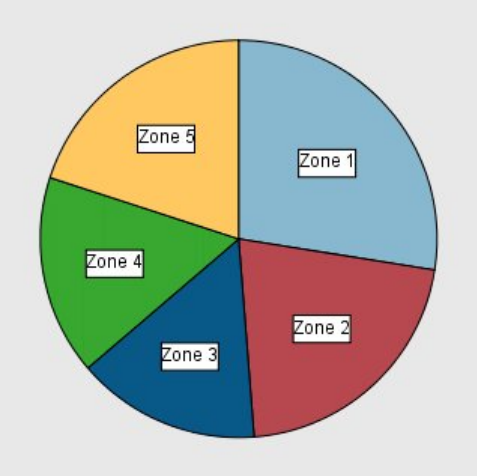


Figure 13. Graphique circulaire

Nous pouvons remarquer que :

- Zone 1 a davantage de clients que chacune des autres zones.
- Les clients sont équitablement répartis entre les autres zones.

Exemple : Carte thermique

Nous allons maintenant créer une carte thermique⌂ catégorielle pour vérifier le revenu moyen des clients de différentes régions géographiques et de différentes tranches d'âge.

Remarque : Cet exemple utilise *customer_subset*.

1. Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
2. Dans l'onglet Base, sélectionnez *Indicateur géographique*, *Catégorie d'âge* et *Revenu du ménage en milliers* , dans cet ordre. (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)
3. Sélectionnez **Carte thermique**⌂.
4. Cliquez sur **Exécuter**.
5. Sur la fenêtre de sortie, cliquez sur le bouton « Afficher les libellés de champ et de valeur » de la barre d'outils (le bouton à droite parmi les deux au centre de la barre d'outils).

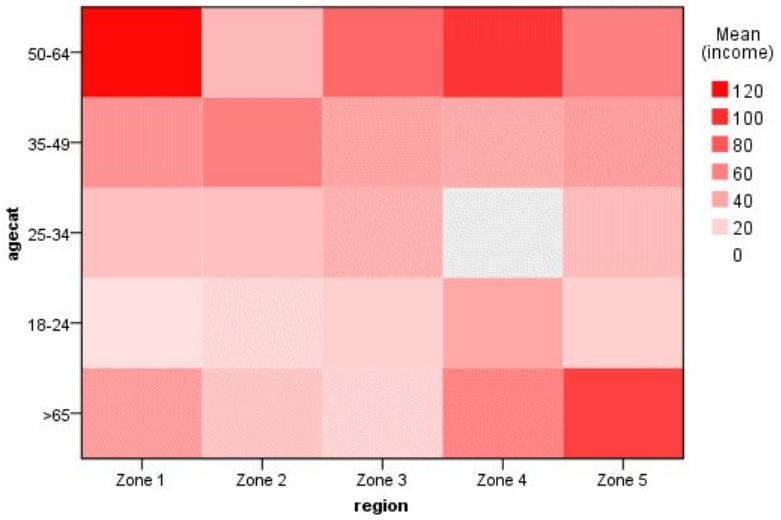


Figure 14. Carte de zones de chaleur catégorielle

Nous pouvons remarquer que :

- Une carte thermique⌂ est semblable à une table utilisant des couleurs plutôt que des numéros pour représenter les valeurs des cellules. Le rouge brillant et profond indique la valeur la plus élevée, tandis que le gris indique une valeur basse. La valeur de chaque cellule est la moyenne du champ ou de la variable continue pour chaque paire de catégories.
- Excepté dans les Zones 2 et 5, le groupe de clients situés dans la tranche d'âge 50-64 ans a un revenu moyen du ménage supérieur à ceux des autres groupes.
- Il n'y a pas de clients situés de la tranche d'âge 25-34 ans dans la Zone 4.

Exemple : Matrice de nuage de points (SPLOM)

Nous allons créer une matrice de diagramme de dispersion de plusieurs variables différentes afin de déterminer s'il existe des relations entre les variables du jeu de données.

Remarque : Cet exemple utilise *customer_subset*.

1. Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
2. Dans l'onglet Base, sélectionnez *Age en années*, *Revenu du ménage en milliers* et *Dette de la carte de crédit en milliers* . (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)
3. Sélectionnez **SPLOM**.
4. Cliquez sur **Exécuter**.
5. Agrandissez la fenêtre de sortie afin de voir la matrice plus distinctement.

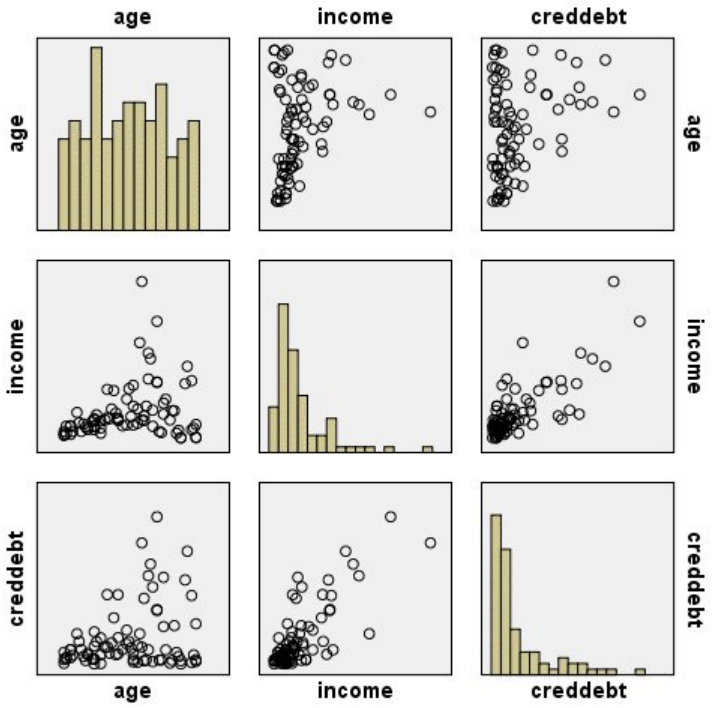


Figure 15. Matrice de nuage de points (SPLOM)

Nous pouvons remarquer que :

- Les histogrammes affichés sur la diagonale montrent la répartition de chaque variable dans la SPLOM. L'histogramme de l'*âge* apparaît dans la cellule en haut à gauche, celui du *revenu* dans la cellule du centre, et celui de la *dettcred* dans la cellule en bas à droite. Aucune des variables ne semble être distribuée normalement. Autrement dit, aucun histogramme ne ressemble à une courbe en cloche. Notez en outre que les histogrammes du *revenu* et de la *dettcred* sont asymétriques.
- Il ne semble pas y avoir de relation entre l'*âge* et les autres variables.
- il existe une relation linaire entre le *revenu* et la *dettcred*. En effet, la *dettcred* augmente à mesure que le *revenu* augmente. Vous pouvez créer des diagrammes de dispersion individuels de ces variables et des autres variables liées pour étudier les relations plus en détails.

Exemple : Choroplèthe (carte de couleur) des sommes

Nous ne créerons pas de visualisation de carte. Par conséquent, dans l'exemple suivant, nous créerons une variation de cette visualisation. Le jeu de données est *ventes mondiales* qui est un fichier de données hypothétiques qui contient les revenus des ventes par continent et par produit.

1. Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
2. Dans l'onglet Base, sélectionnez *Continent* et *Revenus*. (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)
3. Sélectionnez **Choroplèthe des sommes**.
4. Cliquez sur l'onglet Détaillé.
5. Dans le groupe Apparences en option, choisissez *Continent* dans la liste déroulante Libellé de données.
6. Dans le groupe Fichiers cartes, cliquez sur **Sélectionner un fichier carte**.
7. Dans la boîte de dialogue Sélectionner les cartes, vérifiez que **Carte** est défini sur *Continents* et **Clé de carte** est défini sur *CONTINENT*.

8. Dans les groupes Comparer la carte et Valeurs de données, cliquez sur **Comparer** pour vérifier que les clés de carte correspondent aux clés de données. Dans cet exemple, toutes les valeurs de clés de données ont des clés et des fonctions de carte correspondantes. Nous pouvons également voir qu'il n'y a pas de données pour l'Océanie.
9. Dans la boîte de dialogue Sélectionner les cartes, cliquez sur **OK**.
10. Cliquez sur **Exécuter**.

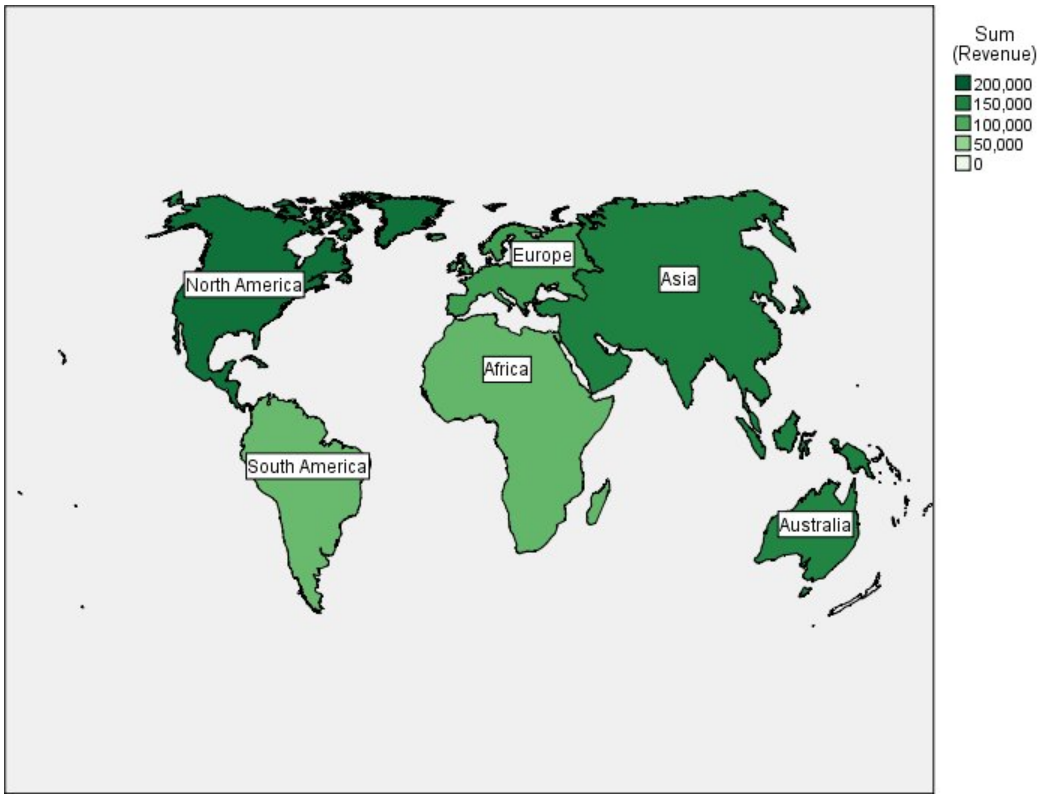


Figure 16. Choroplèthe des sommes

Dans cette visualisation de carte, nous pouvons facilement voir que le revenu est plus élevé en Amérique du Nord qu'en Amérique du Sud et en Afrique. Chaque continent est libellé parce que nous avons utilisé *Continent* comme apparence de libellé de données.

Exemple : Graphiques à barres sur une carte

Cet exemple montre la façon dont les revenus sont divisés en produit dans chaque continent.

Remarque : Cet exemple utilise *ventes mondiales*.

1. Ajoutez un noeud Représentation graphique et ouvrez-le pour le modifier.
2. Dans l'onglet Base, sélectionnez *Continent*, *Produit* et *Revenus*. (Utilisez la combinaison Ctrl+clic pour sélectionner plusieurs champs/variables.)
3. Sélectionnez **Bâtons sur une carte**.
4. Cliquez sur l'onglet Détaillé.

   Lorsque vous utilisez plusieurs champs de type spécifique, il est important de vérifier que chaque champ est affecté à l'emplacement correspondant.
5. Sélectionnez *Produit* dans la liste déroulante Catégories.
6. Sélectionnez *Revenus* dans la liste déroulante Valeurs.
7. Sélectionnez *Continent* dans la liste déroulante Clé de données.

8. Dans la liste déroulante Récapitulatif, sélectionnez *Somme*.
9. Dans le groupe Fichiers cartes, cliquez sur **Sélectionner un fichier carte**.
10. Dans la boîte de dialogue Sélectionner les cartes, vérifiez que **Carte** est défini sur *Continents* et **Clé de carte** est défini sur *CONTINENT*.
11. Dans les groupes Comparer la carte et Valeurs de données, cliquez sur **Comparer** pour vérifier que les clés de carte correspondent aux clés de données. Dans cet exemple, toutes les valeurs de clés de données ont des clés et des fonctions de carte correspondantes. Nous pouvons également voir qu'il n'y a pas de données pour l'Océanie.
12. Dans la boîte de dialogue Sélectionner les cartes, cliquez sur **OK**.
13. Cliquez sur **Exécuter**.
14. Agrandissez la fenêtre de sortie afin de voir l'affichage plus distinctement.

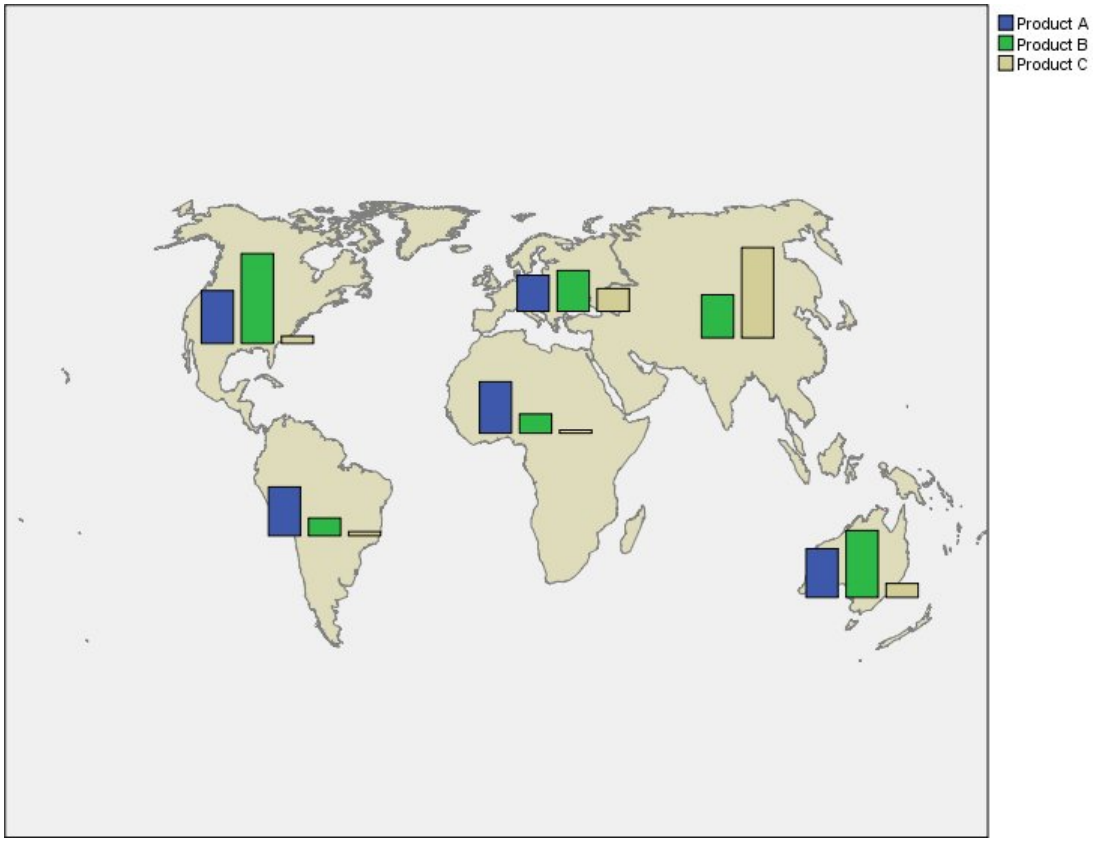


Figure 17. Graphiques à barres sur une carte

Nous pouvons remarquer que :

- La répartition du total des revenus par produit est à peu près la même en Amérique du Sud et en Afrique.
- *Produit C* génère le revenu le moins élevé partout sauf en Asie.
- Il n'y a pas ou quasi pas de revenu du *Produit A* en Asie.

Onglet Apparence du panneau Représentation graphique

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Options générales d'apparence

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous-titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Echantillonnage. Spécifiez la méthode pour les jeux de données volumineux. Vous pouvez spécifier le nombre de modalités maximales des jeux de données ou utiliser le nombre d'enregistrements par défaut. Lorsque vous sélectionnez l'option **Echantillon**, les performances des jeux de données volumineux sont optimisées. Vous pouvez également choisir de représenter tous les points de données en sélectionnant **Utiliser toutes les données**, mais sachez que vous risquez de réduire considérablement les performances du logiciel.

Options d'apparence des feuilles de style

Deux boutons vous permettent de contrôler les modèles de visualisation (et les feuilles de style et les cartes) disponibles :

Gérer. Gérer les modèles de visualisation, les feuilles de style et les cartes sur votre ordinateur. Vous pouvez importer, exporter, renommer et supprimer les modèles de visualisation, les feuilles de style et les cartes depuis votre ordinateur local. Pour plus d'informations, voir «Gérer les modèles, les feuilles de style et les fichiers cartes», à la page 229.

Emplacement. Modifier l'emplacement dans lequel les modèles de visualisation, les feuilles de style et les cartes sont stockés. L'emplacement actuel est indiqué à droite du bouton. Pour plus d'informations, voir la rubrique «Définition de l'emplacement des modèles, des feuilles de style et des cartes.», à la page 228.

L'exemple suivant montre l'emplacement des options d'apparence sur le graphique. (*Remarque* : Tous les graphiques n'utilisent pas toutes ces options.)

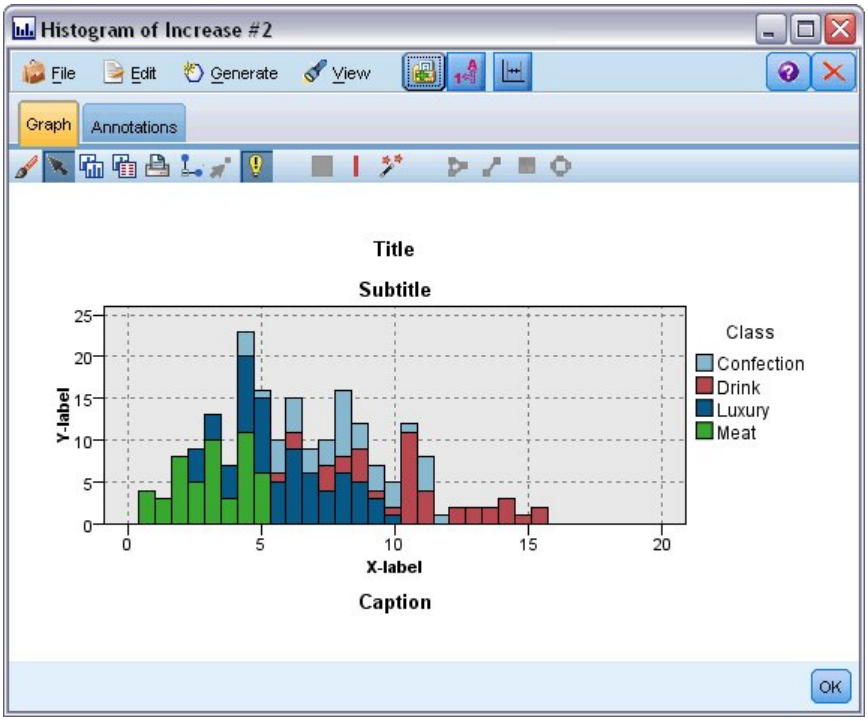


Figure 18. Position des différentes options d'apparence du graphique

Définition de l'emplacement des modèles, des feuilles de style et des cartes.

Les modèles de visualisation, les feuilles de style de visualisation et les fichiers cartes sont stockés dans un dossier local spécifique ou dans le IBM SPSS Collaboration and Deployment Services Repository. Lorsque les modèles, les feuilles de style et les fichiers cartes sont sélectionnés, seuls ceux qui sont intégrés à cet emplacement sont affichés. En conservant tous les modèles, feuilles de style et cartes à un seul endroit, les applications IBM SPSS peuvent facilement y accéder. Pour plus d'informations sur l'ajout de modèles, de feuilles de styles et de fichiers cartes supplémentaires à cet emplacement, consultez «Gérer les modèles, les feuilles de style et les fichiers cartes», à la page 229.

Pour définir l'emplacement des modèles, des feuilles de style et des fichiers cartes

1. Dans une boîte de dialogue de modèle ou de feuilles de style, cliquez sur **Emplacement...** pour afficher la boîte de dialogue Modèles, feuilles de style et cartes.
2. Sélectionnez une option pour l'emplacement par défaut des modèles, des feuilles de style et des fichiers cartes :

   Ordinateur local. Les modèles, les feuilles de style et les fichiers cartes se trouvent dans un dossier spécifique sur votre ordinateur local. Sous Windows XP, ce dossier est *C:\Documents and Settings\<user>\Application Data\SPSSInc\Graphboard*. Il est impossible de modifier ce dossier.

   IBM SPSS Collaboration and Deployment Services Repository. Les modèles, les feuilles de style et les fichiers de carte sont situés dans un dossier spécifié par l'utilisateur dans le IBM SPSS Collaboration and Deployment Services Repository. Pour identifier le dossier spécifique, cliquez sur **Dossier**. Pour plus d'informations, voir «Utilisation du IBM SPSS Collaboration and Deployment Services Repository comme emplacement des modèles, des feuilles de style et des fichiers cartes», à la page 229.
3. Cliquez sur **OK**.

Utilisation du IBM SPSS Collaboration and Deployment Services Repository comme emplacement des modèles, des feuilles de style et des fichiers cartes

Les modèles et les feuilles de style de visualisation peuvent être stockés dans le IBM SPSS Collaboration and Deployment Services Repository. Cet emplacement est un dossier spécifique du IBM SPSS Collaboration and Deployment Services Repository. S'il est défini comme emplacement par défaut, tous les modèles, toutes les feuilles de style et tous les fichiers cartes de cet emplacement peuvent être sélectionnés.

Pour définir un dossier dans IBM SPSS Collaboration and Deployment Services Repository comme emplacement des modèles, des feuilles de style et des fichiers cartes

1. Dans une boîte de dialogue comportant un bouton Emplacement, cliquez sur **Emplacement...**.
2. Sélectionnez IBM SPSS Collaboration and Deployment Services Repository.
3. Cliquez sur **Dossier**.

   Remarque : Si vous n'êtes pas encore connecté au IBM SPSS Collaboration and Deployment Services Repository, le système vous invite à entrer vos informations de connexion.
4. Dans la boîte de dialogue Sélectionner un dossier, sélectionnez le dossier dans lequel les modèles, les feuilles de style et les fichiers cartes sont stockés.
5. Si vous le souhaitez, sélectionnez un libellé depuis l'option **Récupérer l'libellé**. Seuls les modèles, les feuilles de style et les fichiers cartes comportant ce libellé sont affichés.
6. Si vous recherchez un dossier qui contient un modèle ou une feuille de style en particulier, vous pouvez procéder à la recherche du modèle, de la feuille de style ou du fichier carte grâce à l'onglet Rechercher. La boîte de dialogue Sélectionner un dossier sélectionne automatiquement le dossier dans lequel le modèle, la feuille de style ou le fichier de carte est situé.
7. Cliquez sur **Sélectionner un dossier**.

Gérer les modèles, les feuilles de style et les fichiers cartes

Vous pouvez gérer les modèles, les feuilles de style et les fichiers cartes en local sur votre ordinateur à l'aide de la boîte de dialogue Gérer les modèles, les feuilles de style et les cartes. Cette dernière vous permet d'importer, d'exporter, de renommer et de supprimer les modèles de visualisation, les feuilles de style et les fichiers cartes dans l'emplacement local de votre ordinateur.

Cliquez sur **Gérer...** dans l'une des boîtes de dialogue de sélection des modèles, des feuilles de style ou des cartes.

Boîte de dialogue Gérer les modèles, les feuilles de style et les cartes...

L'onglet Modèle répertorie tous les modèles locaux. L'onglet Feuille de style répertorie toutes les feuilles de styles locales, et affiche également des exemples de visualisations avec des données d'échantillon. Vous pouvez sélectionner l'une des feuilles de style pour appliquer ses styles aux exemples de visualisation. Pour plus d'informations, consultez la rubrique «Application de feuilles de style», à la page 313 L'onglet Mappe répertorie tous les fichiers cartes locaux. Cet onglet affiche également les clés de carte qui incluent des valeurs d'échantillon, un commentaire s'il en a été fourni un lors de la création de la carte et un aperçu de la carte.

Les boutons suivants fonctionnent sur tous les onglets activés.

Importer. Importe un modèle de visualisation, une feuille de style ou un fichier carte à partir du fichier système. Le fait d'importer un modèle, une feuille de style ou un fichier carte le rend disponible pour l'application IBM SPSS. Si un autre utilisateur vous a envoyé un modèle, une feuille de style ou un fichier carte, vous l'importez avant de l'utiliser avec votre application.

Exporter. Exporte un modèle de visualisation, une feuille de style ou un fichier carte dans le système de fichiers. Exportez un modèle, une feuille de style ou un fichier carte lorsque vous voulez l'envoyer à un autre utilisateur.

Renommer. Renomme le modèle de visualisation, la feuille de style ou le fichier carte sélectionné. Il est impossible de remplacer un nom par un nom déjà utilisé.

Exporter la clé de carte. Exporter les clés de carte sous la forme d'un fichier csv (valeurs séparées par des virgules). Ce bouton est uniquement activé sur l'onglet Carte.

Supprimer. Supprime le modèle de visualisation, la feuille de style ou le fichier carte sélectionné. Vous pouvez sélectionner plusieurs modèles, feuilles de styles ou fichiers cartes en cliquant tout en maintenant le bouton Ctrl appuyé . Cette action de suppression est irréversible, agissez donc avec précaution.

Conversion et distribution des fichiers de formes Carte

Le sélectionneur de modèles de représentations graphiques vous permet de créer des visualisations de carte en combinant un modèle de visualisation et un fichier SMZ. Les fichiers SMZ sont semblables aux fichiers de formes ESRI (format de fichier SHP). En effet, ils contiennent des informations géographiques permettant de dessiner une carte (par exemple, les frontières d'un pays) mais il sont optimisés pour les visualisations de carte. Le sélectionneur de modèles de représentations graphiques est préinstallé avec un nombre précis de fichiers SMZ. Si vous possédez un fichier de formes ESRI existant que vous souhaitez utiliser pour les visualisations de cartes, vous devez d'abord convertir le fichier de formes en fichier SMZ à l'aide de l'utilitaire de conversion des cartes. L'utilitaire de conversion des cartes prend en charge les fichiers de formes ESRI (types de formes 1, 3 et 5) avec points, polylignes ou polygones contenant une couche unique.

En plus de convertir les fichiers de formes ESRI, l'utilitaire de conversion des cartes vous permet de modifier le niveau de détails des cartes, de modifier les libellés des fonctions, de fusionner et de déplacer les fonctions, en autres options. Vous pouvez également utiliser l'utilitaire de conversion des cartes pour modifier un fichier SMZ existant (y compris ceux déjà installés).

Modification des fichiers SMZ préinstallés

1. Exportez le fichier SMZ du système de gestion. Pour plus d'informations, voir «Gérer les modèles, les feuilles de style et les fichiers cartes», à la page 229.
2. Utilisez l'utilitaire de conversion des cartes pour ouvrir et modifier le fichier SMZ exporté. Il est recommandé d'enregistrer le fichier sous un nom différent. Pour plus d'informations, voir «Utilisation de l'utilitaire de conversion des cartes», à la page 231.
3. Importez le fichier SMZ modifié dans le système de gestion. Pour plus d'informations, voir «Gérer les modèles, les feuilles de style et les fichiers cartes», à la page 229.

Ressources supplémentaires pour les fichiers cartes

Les données géospatiales au format de fichier SHP, qui pourraient être utilisées pour prendre en charge vos besoins de mappage, sont disponibles à partir de nombreuses ressources privées et publiques. Vérifiez les sites Web gouvernementaux locaux pour trouver des données gratuites. De nombreux modèles dans ce produit sont basés sur les données publiques obtenues sur GeoCommons ( ) et auprès du U.S. Census Bureau (http://www.census.gov).

REMARQUE IMPORTANTE : Les informations concernant des produits non IBM ont été obtenues auprès des fournisseurs de ces produits, par l'intermédiaire d'annonces publiques ou via d'autres sources disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances ni leur compatibilité. Elle ne peut recevoir aucune réclamation concernant des produits non IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits. Les références à des sites Web non IBM sont fournies à titre d'information uniquement et n'impliquent en

aucun cas une adhésion aux données qu'ils contiennent. Le matériel de ces sites Web ne fait pas partie du matériel de ce programme IBM, sauf mention contraire dans un fichier Remarques qui accompagne ce programme IBM et l'utilisation du matériel de ces sites se fait à vos propres risques.

Concepts principaux des cartes

Comprendre certains concepts principaux associés aux fichiers de formes vous aidera à utiliser efficacement l'utilitaire de conversion des cartes.

Un **fichier de formes** propose les informations géographiques permettant de dessiner une carte. Il existe trois types de fichiers de formes que l'utilitaire de conversion des cartes prend en charge :

- **Point.** Le fichier de formes identifie les emplacements des points, tels que les villes.
- **Polyligne.** Le fichier de formes identifie les chemins d'accès et leurs emplacements des points, tels que les rivières.
- **Polygone.** Le fichier de formes identifie les contours des régions et leurs emplacements, tels que les pays.

La plupart du temps, vous utiliserez un fichier de formes polygone. Les cartes choroplèthes sont créées à partir des fichiers de formes polygones. Les cartes choroplèthes utilisent la couleur pour représenter une valeur dans des polygones individuels (régions). Les fichiers de formes avec points et polygones sont généralement superposés sur un fichier de formes polygone. Un exemple en est un fichier de formes avec points des villes américaines superposé sur un fichier de formes polygones des Etats américains.

Un fichier de formes est composé de **fonctions**. Les fonctions sont des entités géographiques individuelles. Par exemple, les fonctions peuvent être des pays, des Etats, des villes, etc. Le fichier des formes contient également des données sur les fonctions. Ces données sont stockées dans les **attributs**. Les attributs sont semblables aux champs ou aux variables dans un fichier de données. Il existe toujours au moins un attribut qui est la **clé de carte** de cette fonction. La clé de carte peut être un libellé, comme un pays ou le nom d'une région. La clé de carte est ce que vous relierez à une variable/champ dans un fichier de données pour créer une visualisation de carte.

Veuillez noter que vous pourrez uniquement conserver le ou les attributs clés dans le fichier SMZ. L'utilitaire de conversion des cartes ne prend pas en charge l'enregistrement des attributs supplémentaires. Cela signifie que vous devrez créer plusieurs fichiers SMZ si vous souhaitez effectuer une agrégation à différents niveaux. Par exemple, si vous souhaitez agréger les Etats et les régions américaines, vous aurez besoin de fichiers SMZ distincts : un fichier contenant une clé qui identifie les Etats et un fichier contenant une clé qui identifie les régions.

Utilisation de l'utilitaire de conversion des cartes

Démarrage de l'utilitaire de conversion des cartes

A partir des menus, sélectionnez :

Outils > **Utilitaire de conversion de carte**

Il existe quatre écrans principaux (étapes) dans l'utilitaire de conversion des cartes. Une des étapes inclut également des sous-étapes pour un contrôle plus complet de l'édition des fichiers cartes.

Etape 1 : choisir les fichiers de destination et source

Vous devez d'abord sélectionner un fichier carte source et un fichier de destination pour le fichier carte converti. Vous aurez besoin à la fois du fichier *.shp* et du fichier *.dbf* pour le fichier de formes.

Sélectionnez un fichier .shp (ESRI) ou .smz pour la conversion. Recherchez un fichier carte existant sur votre ordinateur. Il s'agit du fichier que vous convertirez et que vous enregistrerez comme fichier SMZ.

Le fichier *.dbf* pour le fichier de formes *doit* être stocké au même emplacement et avoir un nom de fichier de base qui correspond au fichier *.shp*. Le fichier *.dbf* est nécessaire parce qu'il contient les informations sur les attributs pour le fichier *.shp*.

Définissez une destination et un nom de fichier pour le fichier carte converti. Saisissez un chemin d'accès et un nom de fichier pour le fichier SMZ qui sera créé à partir du fichier source carte d'origine.

- **Importer dans le Sélecteur de modèles.** En plus d'enregistrer un fichier sur le système de fichiers, vous pouvez également ajouter la carte à la liste Gérer du sélectionneur de modèles. Si vous choisissez cette option, la carte sera automatiquement disponible dans le sélectionneur de modèles pour les produits IBM SPSS installés sur votre ordinateur. Si vous ne procédez pas à l'importation dans le Sélectionneur de modèles maintenant, vous devrez effectuer une importation manuelle ultérieurement. Pour plus d'informations sur l'importation des cartes dans le système de gestion Sélecteur de modèle, voir «Gérer les modèles, les feuilles de style et les fichiers cartes», à la page 229.

Etape 2 : choisir une clé de carte

Vous devez maintenant choisir les clés de carte à inclure au fichier SMZ. Vous pourrez ensuite modifier certaines options qui auront un effet sur la représentation de la carte. Les étapes suivantes de l'utilitaire de conversion des cartes contiennent un aperçu de la carte. Les options de représentation que vous choisissez seront utilisées pour générer l'aperçu de la carte.

Choisissez la clé de carte principale. Sélectionnez l'attribut qui sera la clé principale pour identifier et libeller les fonctions de la carte. Par exemple, la clé principale d'une carte du monde pourrait être l'attribut identifiant les noms des pays. La clé principale reliera également vos données aux fonctions de la carte. Par conséquent, vérifiez que les valeurs (libellés) de l'attribut que vous avez choisies correspondent aux valeurs de vos données. Des exemples de libellés sont affichés lorsque vous choisissez un attribut. Si vous avez besoin de modifier ces libellés, vous pourrez le faire dans une autre étape.

Choisir des clés supplémentaires à ajouter. En plus de la clé de carte principale, vérifiez tous les autres attributs clés que vous souhaitez inclure dans le fichier SMZ Généré. Par exemple, certains attributs peuvent contenir des libellés traduits. Si vous savez que vos données sont codées dans d'autres langues, il peut être nécessaire de conserver ces attributs. Veuillez noter que vous pouvez uniquement choisir des clés supplémentaires qui représentent les mêmes fonctions que la clé principale. Par exemple, si la clé principale contient les noms complets des Etats américains, vous pouvez uniquement sélectionner les autres clés qui représentent les Etats américains, comme les abréviations des Etats.

Lisser automatiquement la carte. Les fichiers de formes avec des polygones contiennent généralement trop de points de données et trop de détails pour les visualisations de cartes statistiques. Trop de détails peut avoir un effet distrayant et un impact négatif sur les performances. Vous pouvez réduire le niveau des détails et généraliser la carte grâce au lissage. La carte sera plus claire et sera plus rapidement représentée. Lorsque la carte est automatiquement lissée, l'angle maximum est de 15 degrés et le pourcentage à conserver est de 99. Pour des informations sur ces paramètres, voir «Lisser la carte», à la page 233. Veuillez noter que vous avez la possibilité d'appliquer un lissage supplémentaire ultérieurement dans une autre étape.

Supprimer les bordures entre des polygones attenants dans la même fonction. Certaines fonctions peuvent contenir des sous-fonctions qui contiennent des bordures internes dans les fonctions principales. Par exemple, une carte du monde avec les continents peut contenir des bordures internes des pays qui composent chaque continent. Si vous choisissez cette option, les bordures internes n'apparaîtront pas sur le plan. Dans le cas de la carte du monde avec les continents, choisir cette option supprime les frontières des pays tout en conservant les frontières des continents.

Etape 3 : modifier la carte

Maintenant que vous avez spécifié les options de base de la carte, vous pouvez en ajouter d'autres et les affiner. Ces modifications sont facultatives. Cette étape de l'utilitaire de conversion des cartes vous guide dans les tâches associées et affiche un aperçu de la carte pour que vous puissiez vérifier vos

modifications. Certaines tâches peuvent ne pas être disponibles en fonction du type du fichier de formes (avec points, polylignes ou polygones) et du système de coordonnées.

Chaque tâche possède les contrôle⌂s communs suivants à gauche de l'utilitaire de conversion des cartes.

Afficher les libellés sur la carte. Par défaut, les libellés des fonctions n'apparaissent pas dans l'aperçu. Vous pouvez choisir d'afficher les libellés. Bien que les libellés puissent aider à identifier les fonctions, ils peuvent interférer avec la sélection directe sur l'aperçu de la carte. Activez cette option lorsque vous en avez besoin, par exemple lorsque vous modifiez les libellés des fonctions.

Colorer l'aperçu de la carte. Par défaut, l'aperçu de la carte affiche des zones avec une couleur solide. Toutes les fonctions sont de la même couleur. Vous pouvez choisir d'avoir un assortiment de couleurs attribuées aux fonctions individuelles de la carte. Cette option peut aider à distinguer différentes fonctions de la carte. Cela est particulièrement utile lorsque vous fusionnez des fonctions et que vous souhaitez voir ces nouvelles fonctions représentées dans l'aperçu.

Chaque tâche possède également le contrôle⌂ commun suivant à droite de l'utilitaire de conversion des cartes.

Annuler. Cliquez sur **Annuler** pour revenir à l'état précédent. Vous pouvez annuler un maximum de 100 modifications.

Lisser la carte : Les fichiers de formes avec des polygones contiennent généralement trop de points de données et trop de détails pour les visualisations de cartes statistiques. Trop de détails peut avoir un effet distrayant et un impact négatif sur les performances. Vous pouvez réduire le niveau des détails et généraliser la carte grâce au lissage. La carte sera plus claire et sera plus rapidement représentée. Cette option n'est pas disponible pour les cartes avec points et polygones.

Chiffres maximum. L'angle maximum, qui doit se situer entre une valeur de 1 et 20, spécifie la tolérance pour lisser les ensembles de points qui sont presque linéaires. Une valeur plus importante permet une tolérance plus élevée pour le lissage linéaire et donnera plus de points et par conséquent, une carte plus généralisée. Pour appliquer un lissage linéaire, l'utilitaire de conversion des cartes vérifie l'angle interne formé par chaque ensemble de trois points sur la carte. Si 180 moins l'angle est inférieur à la valeur spécifiée, l'utilitaire de conversion des cartes ignore le point central. Par exemple, l'utilitaire de conversion des cartes vérifie si la ligne formée par les trois points est presque droite. Si c'est le cas, l'utilitaire de conversion des cartes considère la ligne comme une ligne droite entre les points finaux et ignore le point central.

Pourcentage à conserver. Le pourcentage à conserver, qui doit être une valeur comprise entre 90 et 100, détermine la quantité de zone terrestre à conserver lorsque la carte est lissée. Cette option a uniquement un effet sur les fonctions qui contiennent plusieurs polygones comme dans le cas d'une fonction incluant plusieurs îles. Si la zone totale de la fonction moins un polygone est supérieure au pourcentage spécifié de la zone d'origine, l'utilitaire de conversion des cartes ignore le polygone de la carte. L'utilitaire de conversion des cartes ne supprimera jamais tous les polygones de la fonction. C'est-à-dire qu'il y aura au moins un polygone pour la fonction, quelle que soit la quantité de lissage appliqué.

Après avoir choisi un angle et un pourcentage maximum à conserver, cliquez sur **Appliquer**. L'aperçu met à jour les modifications de lissage. Si vous avez besoin de lisser de nouveau la carte, répétez cette action jusqu'au niveau de lissage désiré. Veuillez noter qu'il existe une limite au lissage. Si vous effectuez plusieurs lissages, arrivera un moment où vous ne pourrez plus lisser la carte.

Modifier les libellés des fonctionnalités : Vous pouvez modifier les libellés des fonctions (peut-être pour qu'elles correspondent aux données attendues) et également repositionner les libellés sur la carte. Même si vous ne pensez pas avoir besoin de modifier les libellés, vérifiez-les avant de créer les visualisations à partir de la carte. Parce que les libellés n'apparaissent pas par défaut dans l'aperçu, vous pouvez également sélectionner l'option **Afficher les libellés sur la carte** pour les afficher.

Clés. Sélectionnez la clé contenant les libellés des fonctions à consulter et/ou modifier.

Fonctions. Cette liste affiche les libellés des fonctions contenues dans la clé sélectionnée. Pour modifier le libellé, faites un double clic sur cette liste. Si les libellés apparaissent sur la carte, vous pouvez également faire un double clic sur les libellés des fonctions directement dans l'aperçu de la carte. Si vous souhaitez comparer les libellés à un fichier de données réel, cliquez sur **Comparer**.

X/Y. Ces zones de texte répertorient le point central actuel de la fonctionnalité sur la carte. Les unités apparaissent dans les coordonnées de la carte. Il peut s'agir de coordonnées locales cartésiennes (par exemple, le SPCS - State Plane Coordinate System-) (où **X** est la longitude et **Y** la latitude). Saisissez les coordonnées pour la nouvelle position du libellé. Si les libellés apparaissent, vous pouvez également cliquer et faire glisser un libellé sur la carte pour la déplacer. Les zones de texte seront mises à jour avec la nouvelle position.

Comparaison. Si vous avez un fichier de données qui contient des valeurs de données censées correspondre aux libellés des fonctions pour une clé particulière, cliquez sur **Comparer** pour afficher la boîte de dialogue Comparer à une source de données externe. Dans cette boîte de dialogue, vous pourrez ouvrir le fichier de données et comparer ses valeurs directement avec celles qui se trouvent dans les libellés des fonctions de la clé de carte.

Boîte de dialogue Comparer à une source de données externe : La boîte de dialogue Comparer à une source de données externe vous permet d'ouvrir un fichier de valeurs au format tabulé (avec une extension *.txt*), un fichier de valeurs séparées par des virgules (avec une extension *.csv*) ou un fichier de données mis en forme pour IBM SPSS Statistics (avec une extension *.sav*). Lorsque le fichier est ouvert, vous pouvez sélectionner un champ dans le fichier de données pour comparer les libellés des fonctions dans une clé de carte spécifique. Vous pouvez ensuite corriger les incohérences du fichier carte.

Champs dans le fichier de données. Choisissez le champ dont vous souhaitez comparer les valeurs aux libellés des fonctions. Si la première ligne du fichier *.txt* ou *.csv* contient des libellés de description pour chaque champ, cochez **Utiliser la première ligne comme libellé de colonne**. Sinon, chaque champ sera identifié par sa position dans le fichier de données (par exemple, "Colonne 1", "Colonne 2", etc.).

Clé à comparer. Choisissez la clé de carte dont vous souhaitez comparer les libellés des fonctions aux valeurs des champs des fichier de données.

Comparaison. Cliquez lorsque vous êtes prêt à comparer les valeurs.

Résultats des comparaisons. Par défaut, le tableau Résultats des comparaisons ne répertorie que les valeurs de champ sans correspondance dans le fichier de données. L'application essaie de trouver un libellé de fonction associée, en vérifiant généralement les espaces ajoutés ou manquants. Cliquez sur la liste déroulante dans la colonne *Libellé de carte* pour trouver le libellé de fonction dans le fichier carte correspondant à la valeur de champ affichée. S'il n'existe aucun libellé de fonction correspondante dans votre fichier carte, choisissez l'option *Laisser les libellés sans correspondance*. Si vous souhaitez afficher toutes les valeurs de champ, même celles qui correspondent déjà à un libellé de fonctionnalité, décochez l'option **Afficher uniquement les observations sans correspondance**. Ceci peut être utile pour remplacer une ou plusieurs correspondances.

Chaque fonction ne peut être utilisée qu'une seule fois pour correspondre à une valeur de champ. Si vous souhaitez faire correspondre plusieurs fonctions à une seule valeur de champ, il est possible de fusionner les fonctions puis de faire correspondre la nouvelle fonction fusionnée à la valeur de champ. Pour plus d'informations sur les fonctionnalités de fusion, voir «Fusionner les fonctions».

Fusionner les fonctions : La fusion des fonctions permet de créer de grandes régions sur une carte. Par exemple, si vous convertissez une carte des Etats, vous pouvez fusionner les Etats (les fonctions dans cet exemple) en régions Nord, Sud, Ouest et Est de plus grande taille.

Clés. Sélectionnez la clé de carte contenant les libellés de fonctions qui vous aident à identifier les fonctions à fusionner.

Fonctions. Cliquez sur la première fonction à fusionner. Cliquez sur Ctrl pour sélectionner les autres fonctions à fusionner. Veuillez noter que les fonctionnalités seront également sélectionnées dans l'aperçu de la carte. Vous pouvez cliquer directement sur les fonctions puis sur Ctrl dans l'aperçu de la carte en plus de les sélectionner dans la liste.

Après avoir sélectionné les fonctions à fusionner, cliquez sur **Fusionner** pour afficher la boîte de dialogue Nommer la fonction fusionné dans laquelle vous pourrez appliquer un libellé à la nouvelle fonction. Vous pouvez cocher **Colorer l'aperçu de la carte** après avoir fusionné les fonctions pour vous assurer que les résultats sont corrects.

Après avoir fusionné les fonctions, vous pouvez également déplacer le libellé de la nouvelle fonction. Pour ce faire, utilisez la tâche *Modifier les libellés des fonctions*. Pour plus d'informations, voir «Modifier les libellés des fonctionnalités», à la page 233.

Boîte de dialogue Nommer la fonction fusionnée : La boîte de dialogue Nommer la fonction fusionnée vous permet d'attribuer des libellés à la nouvelle fonction fusionnée.

Le tableau Libellés affiche les informations pour chaque clé dans le fichier carte et vous permet d'attribuer un libellé à chaque clé.

Nouveau libellé. Saisissez un nouveau libellé pour la fonction fusionnée à attribuer à la clé de carte spécifique.

Clé. La clé de carte à laquelle vous attribuez le nouveau libellé.

Anciens libellés. Les libellés des fonctions qui seront fusionnées dans la nouvelle fonction.

Supprimer les bordures entre les polygones attenants. Cocher cette option pour supprimer les bordures des fonctions ayant été fusionnées. Par exemple, si vous fusionnez des Etats dans des zones géographiques, cette option supprime les bordures autour des Etats individuels.

Déplacer les fonctionnalités : Vous pouvez déplacer les fonctions dans la carte. Cette option peut être utile lorsque vous souhaitez rassembler des fonctions, comme le continent et les îles environnantes.

Clés. Sélectionnez la clé de carte contenant les libellés de fonctions qui vous aident à identifier les fonctions à déplacer.

Fonctions. Cliquez sur la première fonction à déplacer. Veuillez noter que la fonction sera sélectionnée dans l'aperçu de la carte. Vous pouvez également cliquer directement sur la fonction dans l'aperçu de la carte.

X/Y. Ces zones de texte répertorient le point central actuel de la fonctionnalité sur la carte. Les unités apparaissent dans les coordonnées de la carte. Il peut s'agir de coordonnées locales cartésiennes (par exemple, le SPCS - State Plane Coordinate System-) (où **X** est la longitude et **Y** la latitude). Saisissez les coordonnées pour la nouvelle position de la fonction. Vous pouvez également cliquer sur une fonction et la déplacer sur la carte. Les zones de texte seront mises à jour avec la nouvelle position.

Supprimer les fonctionnalités : Vous pouvez supprimer les fonctions indésirables de la carte. Cela peut être utile lorsque vous souhaitez supprimer certaines répétitions en supprimant des fonctions qui ne vous intéressent pas dans la visualisation de carte.

Clés. Sélectionnez la clé de carte contenant les libellés des fonctions qui vous aident à identifier les fonctions à supprimer.

Fonctions. Cliquez sur la fonction à supprimer. Si vous souhaitez supprimer plusieurs fonctions en même temps, cliquez sur les fonctions supplémentaires en appuyant sur Ctrl. Veuillez noter que les fonctionnalités seront également sélectionnées dans l'aperçu de la carte. Vous pouvez cliquer directement sur les fonctions puis sur Ctrl dans l'aperçu de la carte en plus de les sélectionner dans la liste.

Supprimer les éléments individuels : En plus de supprimer des fonctions entières, vous pouvez supprimer certains éléments individuels qui composent ces fonctions, comme des lacs et de petites îles. Cette option n'est pas disponible pour les cartes avec points.

Eléments. Cliquez sur les éléments à supprimer. Si vous souhaitez supprimer plusieurs éléments en même temps, cliquez sur les éléments supplémentaires en appuyant sur Ctrl. Veuillez noter que les éléments seront également sélectionnés dans l'aperçu de la carte. Vous pouvez cliquer directement sur les éléments puis sur Ctrl dans l'aperçu de la carte en plus de les sélectionner dans la liste. Parce que la liste des noms d'éléments n'est pas descriptive (chaque élément est doté d'un chiffre dans la fonctionnalité), vérifiez la sélection dans l'aperçu de la carte pour vous assurer que vous avez bien sélectionné les éléments désirés.

Définir la projection :

La projection de carte spécifie la façon dont la terre en trois dimensions est représentée en deux dimensions. Toutes les projections provoquent des distorsions. Cependant, certaines projections sont plus adaptées en fonction du type de carte utilisé : mondiale ou locale. De plus, certaines projections préservent la forme des fonctions d'origine. Les projections qui préservent la forme sont des projections conformes. Cette option est disponible uniquement pour les cartes avec des coordonnées géographiques (longitude et latitude).

Contrairement aux autres options de l'utilitaire de conversion des cartes, la projection peut être modifiée après la création d'une visualisation de carte.

Projection. Sélectionnez une projection de carte. Si vous créez une carte mondiale ou des hémisphères, utilisez les projections *Locale*, de *Mercator* ou *Winkel Tripel*. Pour les zones plus petites, utilisez les projections *Locale*, *conique conforme de Lambert* ou de *Mercator transverse*. Toutes les projections utilisent l'ellipsoïde WGS83 pour les données.

- La projection **Locale** est toujours utilisée lorsque la carte est créée avec un système de coordonnées local, tel que le SPCS (State Plane Coordinate System). Ces systèmes de coordonnées sont définis par des coordonnées cartésiennes plutôt que par des coordonnées géographiques (longitude et latitude). Dans la projection locale, les lignes horizontales et verticales sont espacées de la même manière dans un système de coordonnées cartésiennes. La projection locale n'est pas conforme.
- La projection de **Mercator** est une projection conforme pour les cartes mondiales. Les lignes horizontales et verticales sont droites et toujours perpendiculaires. Veuillez noter que la projection de Mercator s'étend à l'infini en se rapprochant des pôles Nord et Sud et ne peut donc pas être utilisée si votre carte contient le pôle Nord ou le pôle Sud. La distorsion est la plus importante lorsque la carte se rapproche de ces limites.
- La projection de **Winkel Tripel** est une projection non conforme pour les cartes mondiales. Bien qu'elle ne soit pas conforme, elle offre un bon compromis entre la forme et la taille. A l'exception du méridien de l'équateur et du méridien origine, toutes les lignes sont courbes. Si votre carte mondiale contient le pôle Nord ou Sud, cette projection est un choix adapté.
- Comme son nom l'indique, la projection **conique conforme de Lambert** est une projection conforme utilisée pour les cartes des continents ou de zones terrestres plus petites qui sont plus longues à l'Est et à l'Ouest qu'au Nord et au Sud.
- La projection de **Mercator transverse** est une autre projection conforme pour les cartes continentales ou les zones terrestres plus petites. Utilisez cette projection pour les zones terrestres qui sont plus longues au Nord et au Sud qu'à l'Est et à l'Ouest.

Etape 4 : fin

A ce moment, vous pouvez ajouter un commentaire pour décrire le fichier carte et également créer un fichier de données d'échantillon à partir des clés de carte.

Clés de carte. S'il existe plusieurs clés dans le fichier carte, sélectionnez une clé de carte dont vous souhaitez afficher les libellés de fonctions dans l'aperçu. Si vous créez un fichier de données à partir de la carte, ces libellés seront utilisés pour les valeurs des données.

Commentaire. Entrez un commentaire qui décrit la carte ou offre des informations supplémentaires pouvant être utiles à vos utilisateurs, comme les sources des fichiers de formes d'origine. Ce commentaire apparaît dans le système de gestion du sélectionneur de modèles de représentations graphiques.

Créer un jeu de données à partir des libellés de fonctionnalité. Cochez cette option si vous souhaitez créer un fichier de données texte à partir des libellés de fonctions affichés. Lorsque vous cliquez sur **Parcourir...**, vous pouvez spécifier un emplacement et un nom de fichier. Si vous ajoutez une extension *.txt*, le fichier sera enregistré sous la forme d'un fichier avec valeurs séparées par des tabulations. Si vous ajoutez une extension *.csv*, le fichier sera enregistré sous la forme d'un fichier avec valeurs séparées par des virgules. Si vous ajoutez une extension *.sav*, le fichier sera enregistré au format IBM SPSS Statistics. SAV est le format par défaut lorsqu'aucune extension n'est spécifiée.

Distribution des fichiers cartes

Lors de la première étape de l'utilitaire de conversion des cartes, vous avez choisi un emplacement où enregistrer le fichier SMZ converti. Vous pouvez également choisir d'ajouter la carte au système de gestion pour le sélectionneur de modèles de représentations graphiques. Si vous choisissez d'enregistrer dans le système de gestion, la carte sera disponible dans n'importe quel produit IBM SPSS que vous exécutez sur le même ordinateur.

Pour distribuer la carte à d'autres utilisateurs, vous devrez leur envoyer le fichier SMZ. Ces utilisateurs pourront ensuite utiliser le système de gestion pour importer la carte. Vous pouvez simplement envoyer le fichier dont vous avez spécifié l'emplacement à l'étape 1. Si vous souhaitez envoyer un fichier qui se trouve dans le système de gestion, vous devez d'abord l'exporter :

1. Dans le sélectionneur de modèles, cliquez sur **Gérer...**
2. Cliquez sur l'onglet Carte.
3. Sélectionnez la carte à distribuer.
4. Cliquez sur **Exporter...** et choisissez un emplacement où enregistrer le fichier.

Vous pouvez envoyer le fichier carte physique à d'autres utilisateurs. Les utilisateurs devront refaire le processus à l'envers et importer la carte dans le système de gestion.

Noeud Nuage

Les noeuds Nuage montrent les relations existant entre les champs numériques. Vous pouvez créer un graphique Nuage à l'aide de points (on parle alors également de diagramme de dispersion) ou à l'aide de lignes. Vous pouvez créer trois types de nuage de lignes en définissant le mode X dans la boîte de dialogue.

Mode X = Trier

Paramétrez le mode X sur **Trier** pour trier les données du champ représenté sur l'axe *x* par valeur. Une ligne unique allant de gauche à droite apparaît sur le graphique. Si vous utilisez un champ nominal en tant que superposition, vous obtenez plusieurs lignes de différentes nuances, allant de gauche à droite sur le graphique.

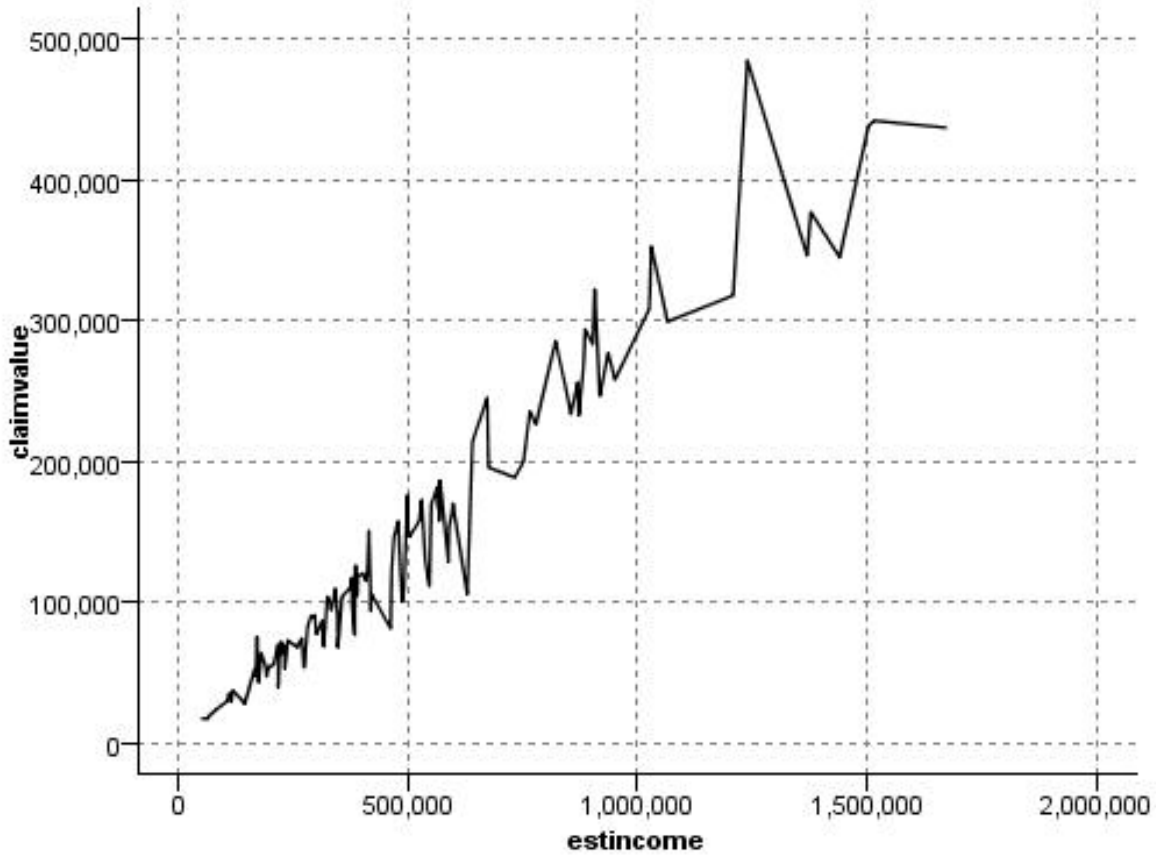


Figure 19. Nuage de lignes avec le mode X paramétré sur Trier

Mode X = Superposer

Paramétrez le mode X sur **Superposer** pour créer plusieurs graphiques de répartition sur le même graphique. Dans un tracé de superposition, les données ne sont pas triées. Tant que les valeurs de l'axe *x* augmentent, les données sont représentées sur une seule ligne. Si les valeurs diminuent, une nouvelle ligne apparaît. Par exemple, si la valeur de *x* s'accroît de 0 à 100, les valeurs *y* sont représentées par une ligne unique. Si la valeur de *x* passe en dessous de 100, une nouvelle ligne est tracée. Le tracé terminé peut contenir plusieurs tracés, ce qui est pratique pour comparer plusieurs séries de valeurs *y*. Ce type de tracé est utile pour les données intégrant une composante temporelle périodique, telle que la consommation électrique sur des périodes successives de vingt-quatre heures.

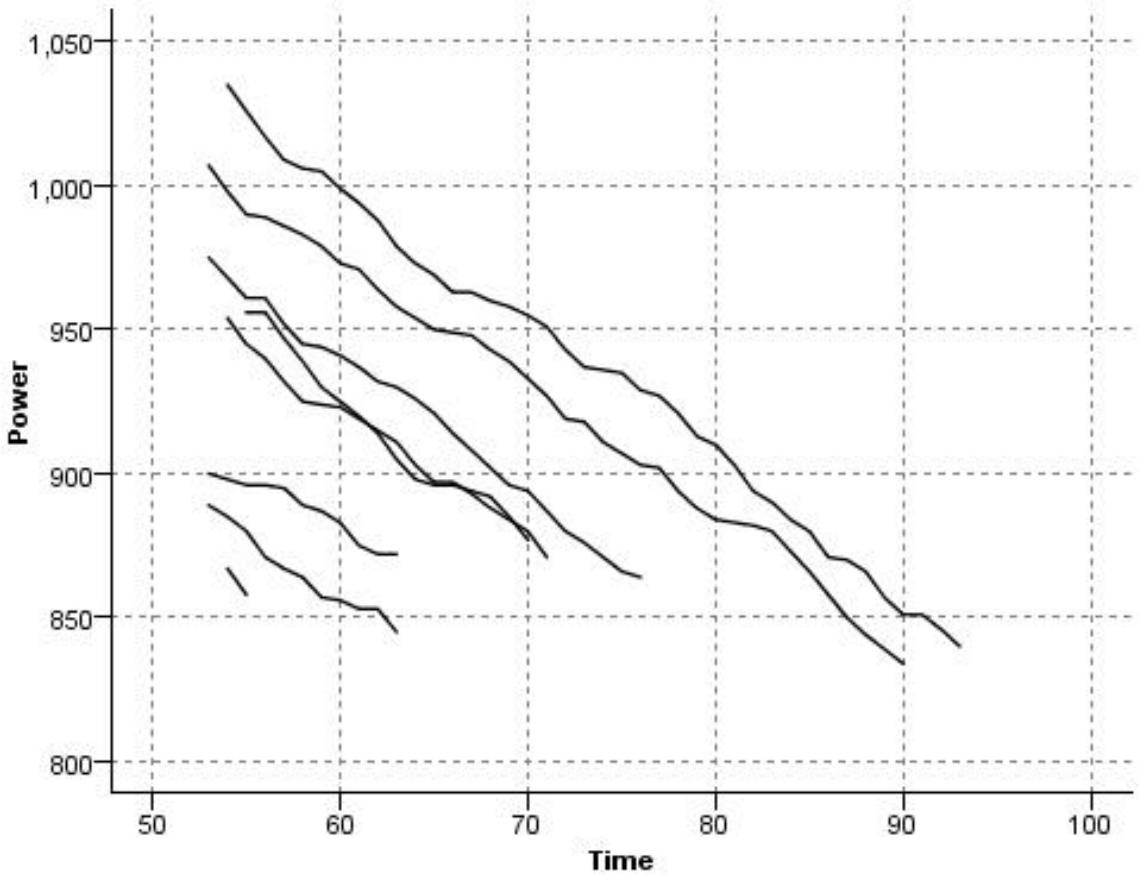


Figure 20. Nuage de lignes avec le mode X paramétré sur Superposer

Mode X = Selon lecture

Paramétrez le mode X sur **Selon lecture** pour représenter les valeurs *x* et *y* telles qu'elles sont lues dans la source de données. Cette option est utile pour les données intégrant une série temporelle et pour lesquelles vous vous intéressez aux tendances ou aux motifs dépendant de l'ordre des données. Il faut parfois trier les données avant de créer ce type de tracé. Il peut également être intéressant de comparer deux nuages similaires dont le mode X est respectivement paramétré sur **Trier** et sur **Selon lecture** afin de déterminer dans quelle mesure le tri influence le motif.

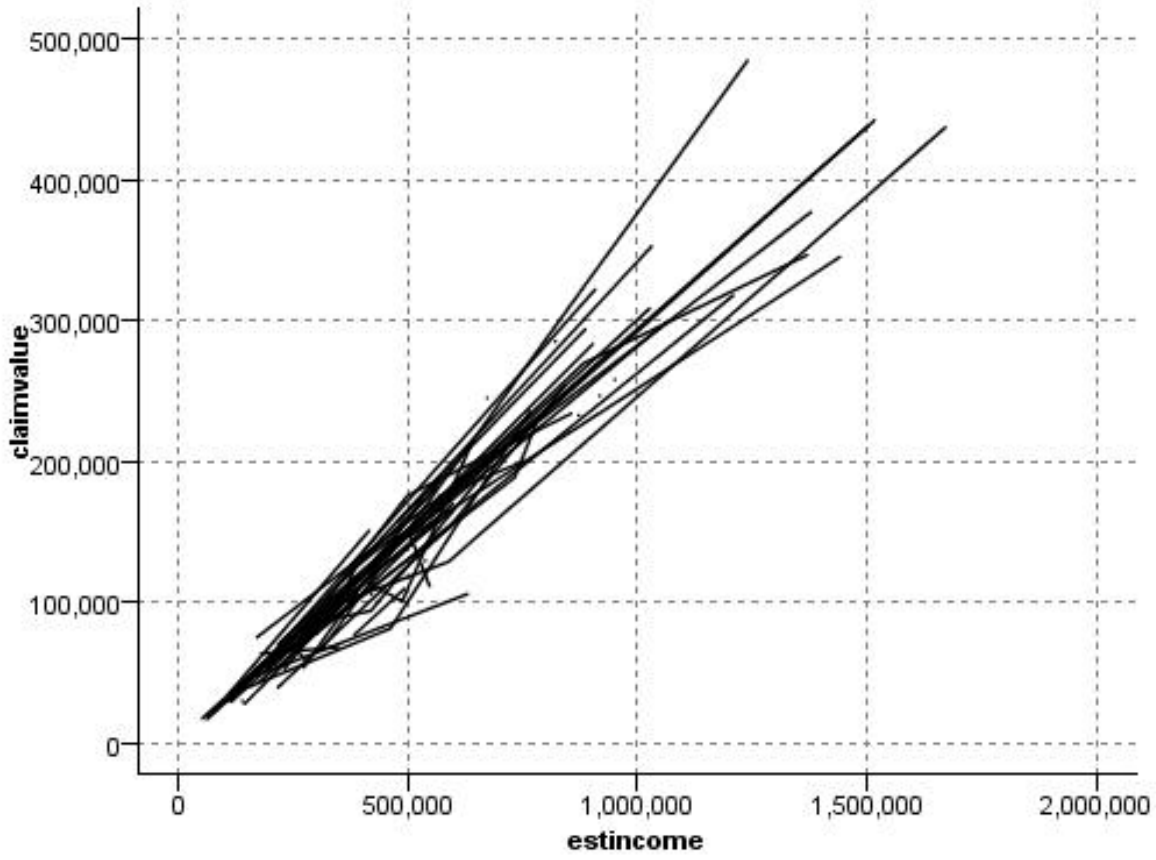


Figure 21. Nuage de lignes affiché précédemment avec le paramétrage Trier, exécuté à nouveau avec le mode X paramétré sur Selon lecture

Vous pouvez aussi utiliser le noeud Représentation graphique pour produire des diagrammes de dispersion et des nuages de lignes. Néanmoins, vous pouvez choisir parmi davantage d'options dans ce noeud. Pour plus d'informations, voir «Types de visualisation des Représentations graphiques intégrées disponibles», à la page 208.

Onglet Noeud nuage

Les graphiques Nuage comparent les valeurs d'un champ *Y* à celles d'un champ *X*. En général, ces champs correspondent respectivement à une variable dépendante et à une variable indépendante.

Champ X. Dans la liste, sélectionnez le champ à afficher sur l'axe *x* horizontal.

Champ Y. Dans la liste, sélectionnez le champ à afficher sur l'axe *y* vertical.

Champ Z. Lorsque vous cliquez sur le bouton 3D du graphiques à barres empilées, vous pouvez ensuite sélectionner un champ dans la liste à afficher sur l'axe *z*.

Superposition. Il existe plusieurs méthodes pour mettre en évidence les catégories des valeurs de données. Par exemple, vous pouvez utiliser *récolteprincipale* en tant que superposition de couleurs afin d'indiquer les valeurs *revenuest* et *valeurréclamation* de la récolte principale cultivée par les demandeurs. Pour plus d'informations, voir «Apparences, superpositions, panneaux et animation», à la page 198.

Type de superposition. Indique si une fonction de superposition ou un lissage apparaît. Les fonctions de lissage et de superposition sont toujours calculées comme fonctions de y.

- **Aucun.** Aucune superposition n'est affichée.
- **Lissage.** Affiche une ligne lissée, calculée à l'aide d'une régression des moindres carrés itérative et robuste pondérée localement (LOESS). Cette méthode permet de calculer efficacement une série de régressions, chacune étant axée sur une petite zone du nuage. Une série de droites de régression "locale" est alors obtenue ; ces droites sont ensuite reliées pour créer une courbe lissée.

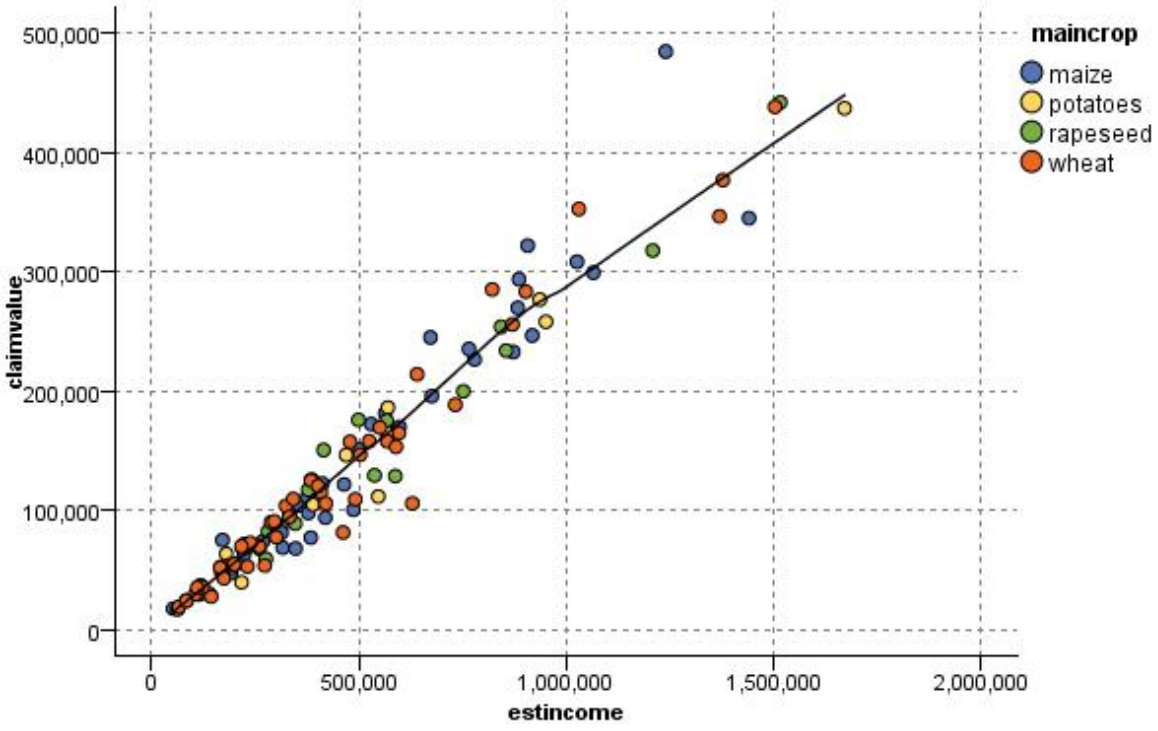


Figure 22. Nuage avec superposition par lissage LOESS

- **Fonction.** Sélectionnez cette option pour indiquer la fonction connue à comparer aux valeurs réelles. Par exemple, pour comparer des valeurs réelles à des valeurs prédites, vous pouvez représenter la fonction y = x sous la forme d'une superposition. Indiquez une fonction **y =** dans la zone de texte. La fonction par défaut est y = x, mais vous pouvez définir toutes sortes de fonctions, telles qu'une fonction quadratique ou une expression arbitraire en termes de x.

  Remarque : les fonctions de superposition ne sont pas disponibles pour les panneaux ou les graphiques animés.

Une fois les options du tracé définies, vous pouvez exécuter le tracé directement à partir de la boîte de dialogue en cliquant sur **Exécuter**. Si vous le souhaitez, vous pouvez cependant utiliser l'onglet Options pour ajouter des spécifications, telles que la création d'intervalles, le mode X et le style.

Onglet Options nuage

Style. Sélectionnez le style de tracé **Point** ou **Ligne**. Si vous sélectionnez **Ligne**, l'option **Mode X** est activée. La sélection de **Point** utilise un symbole plus (+) comme forme de point par défaut. Une fois le graphique créé, vous pouvez modifier la forme des points et leur taille.

Mode X. Pour les nuages de lignes, vous devez choisir un mode X pour définir le style du nuage. Sélectionnez **Trier**, **Superposer** ou **Selon lecture**. Pour **Superposer** ou **Selon lecture**, vous devez spécifier le nombre de modalités maximales des jeux de données à utiliser pour échantillonner les *n* premiers enregistrements. Sinon, les 2,000 enregistrements par défaut sont utilisés.

Amplitude X automatique. Sélectionnez cette option pour utiliser l'intégralité de l'amplitude de valeurs des données sur cet axe. Désélectionnez cette option pour utiliser un sous-ensemble de valeurs explicite déterminé par les valeurs **Minimum** et **Maximum** spécifiées. Vous pouvez entrer les valeurs ou utiliser les flèches. Les amplitudes automatiques sont sélectionnées par défaut pour permettre la création rapide de graphiques.

Amplitude Y automatique. Sélectionnez cette option pour utiliser l'intégralité de l'amplitude de valeurs des données sur cet axe. Désélectionnez cette option pour utiliser un sous-ensemble de valeurs explicite déterminé par les valeurs **Minimum** et **Maximum** spécifiées. Vous pouvez entrer les valeurs ou utiliser les flèches. Les amplitudes automatiques sont sélectionnées par défaut pour permettre la création rapide de graphiques.

Plage Z automatique. Seulement quand un graphique en 3D est spécifié sur l'onglet Nuage. Sélectionnez cette option pour utiliser l'intégralité de l'amplitude de valeurs des données sur cet axe. Désélectionnez cette option pour utiliser un sous-ensemble de valeurs explicite déterminé par les valeurs **Minimum** et **Maximum** spécifiées. Vous pouvez entrer les valeurs ou utiliser les flèches. Les amplitudes automatiques sont sélectionnées par défaut pour permettre la création rapide de graphiques.

Gigue. L'**agitation** est utile pour les nuages de points représentant un jeu de données dans lequel plusieurs valeurs sont récurrentes. Pour obtenir une proportion plus claire des valeurs, vous pouvez utiliser un effet d'agitation afin de distribuer les points de façon aléatoire autour de la valeur réelle.

Remarque à l'attention des utilisateurs des anciennes versions d'IBM SPSS Modeler: la valeur de l'effet d'agitation appliquée dans un tracé utilise une mesure différente dans cette version d'IBM SPSS Modeler. Dans les versions précédentes, la valeur était un nombre réel. Il s'agit désormais d'une proportion de la taille du cadre. Autrement dit, les valeurs d'agitation des anciens flux risquent d'être trop élevées. Dans cette version, toute valeur d'agitation autre que zéro prendra la valeur 0,2.

Nombre maximal d'enregistrements dans le tracé. Spécifiez la méthode de représentation des jeux de données volumineux. Vous pouvez spécifier le nombre de modalités maximales des jeux de données ou utiliser les 2 000 enregistrements par défaut. Lorsque vous sélectionnez les options **Intervalle** ou **Echantillon**, les performances des jeux de données volumineux sont optimisées. Vous pouvez également choisir de représenter tous les points de données en sélectionnant **Utiliser toutes les données**, mais sachez que vous risquez de réduire considérablement les performances du logiciel.

Remarque : lorsque le mode X est paramétré sur **Superposer** ou sur **Selon lecture**, ces options sont désactivées et seuls les *n* premiers enregistrements sont utilisés.

- **Casier.** Sélectionnez cette option pour permettre la création de casiers lorsque le jeu de données contient plus d'enregistrements que le nombre spécifié. La création de casiers applique une fine grille au graphique avant que soit effectué le traçage réel et compte le nombre de points apparaissant dans chacune des cellules de la grille. Dans le graphique final, un point est représenté dans chaque cellule, au niveau du centroïde du casier (moyenne de tous les emplacements de point de casier). La taille des symboles représentés indique le nombre de points dans cette zone (sauf si vous avez utilisé la taille en tant que superposition). L'utilisation du centroïde et de la taille pour représenter le nombre de points fait du nuage mis en casiers un excellent moyen de représenter les jeux de données volumineux. En effet, elle permet d'éviter la superposition des tracés dans les zones denses (masses de couleur impossibles à différencier) et de réduire le nombre d'artefacts de symbole (motifs de densité artificiels). Les artefacts de symbole se produisent lorsque certains symboles (notamment le symbole [+]) entrent en conflit, créant ainsi des zones denses qui n'existaient pas dans les données brutes.

- **Echantillon.** Sélectionnez cette option pour échantillonner de façon aléatoire les données dans le nombre d'enregistrements saisi dans le champ de texte. La valeur par défaut est 2000.

Onglet Apparence tracé

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous-titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Libellé X. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *x* (horizontal), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Libellé Y. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *y* (vertical), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Libellé Z. Disponible uniquement pour les graphiques en 3D. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *z*, soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Utilisation d'un graphique Tracé

Les graphiques Tracé et Courbes sont principalement basés sur la comparaison des valeurs *X* par rapport aux valeurs *Y*. Par exemple, si vous étudiez une fraude potentielle en matière de demande de subventions agricoles, vous pouvez comparer, par le biais d'un réseau de neurones, le revenu réclamé dans la demande à son estimation. L'utilisation d'une superposition, telle que le type de culture, permettra de démontrer s'il existe un lien entre la demande (valeur ou nombre) et le type de culture.

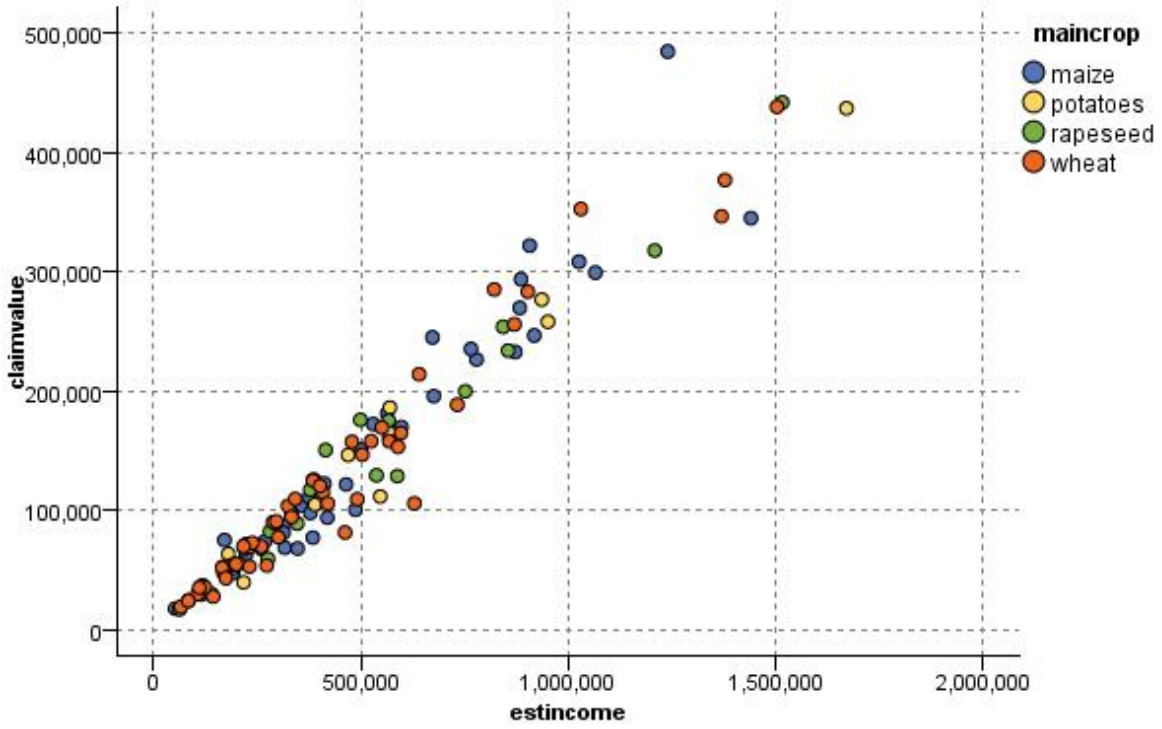


Figure 23. Tracé représentant la relation entre le revenu estimé et la valeur de la demande, le type de culture principale servant de superposition

Les graphiques Nuage, Courbes et Evaluation étant des illustrations en deux dimensions de la comparaison entre *Y* et *X*, il est facile d'interagir avec ceux-ci en définissant des zones, en marquant un élément, ou en traçant des bandes. Vous pouvez également générer des noeuds pour les données représentées par ces zones, bandes, ou éléments. Pour plus d'informations, voir «Exploration de graphiques», à la page 290.

Noeud Courbes

Un graphique Courbes est un type de graphique particulier qui affiche plusieurs champs *Y* pour un seul champ *X*. Les champs *Y* sont représentés par des lignes colorées. Chacun équivaut à un noeud Tracé dont le style est défini sur **Ligne** et le mode X sur **Trier**. Les graphiques Courbes sont utiles lorsque vous avez des données de séquence temporelle et que vous souhaitez explorer les variations de plusieurs variables dans le temps.

Onglet Tracé de courbes

Champ X. Dans la liste, sélectionnez le champ à afficher sur l'axe *x* horizontal.

Champs Y. Sélectionnez dans la liste les champs à afficher sur l'intervalle des valeurs de champ *X*. Utilisez le sélecteur de champs pour sélectionner plusieurs champs. Cliquez sur le bouton de suppression pour supprimer des champs de la liste.

Superposition. Il existe plusieurs méthodes pour mettre en évidence les catégories des valeurs de données. Par exemple, vous pouvez utiliser une superposition animée afin d'afficher plusieurs tracés pour chaque valeur des données. C'est utile pour les ensembles contenant plus de 10 catégories. Lors d'une

utilisation avec des ensembles contenant plus de 15 catégories, vous remarquerez peut-être une baisse des performances. Pour plus d'informations, voir «Apparences, superpositions, panneaux et animation», à la page 198.

Normaliser. Sélectionnez cette option pour mettre toutes les valeurs *Y* à l'échelle sur l'intervalle 0–1 afin de les afficher sur le graphique. La fonction de normalisation vous permet d'explorer les relations existant entre les lignes, relations qui risqueraient sinon d'être occultées en raison des différences au niveau de l'intervalle de valeurs de chaque série ; il est recommandé de l'utiliser lorsque vous représentez plusieurs lignes sur le même graphique ou lorsque vous comparez des graphiques dans des panneaux mitoyens. (Il est inutile d'appliquer une normalisation lorsque toutes les valeurs de données sont comprises dans un même intervalle.)

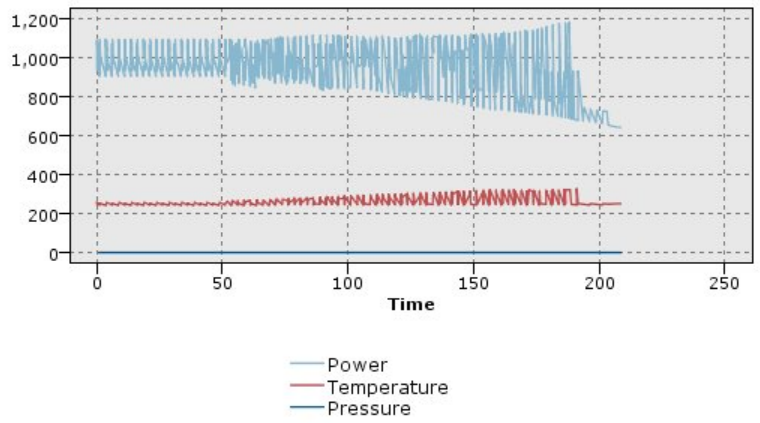


Figure 24. Graphique Courbes standard indiquant les variations de la centrale électrique dans le temps (sans normalisation, il est impossible de visualiser le tracé concernant la pression)

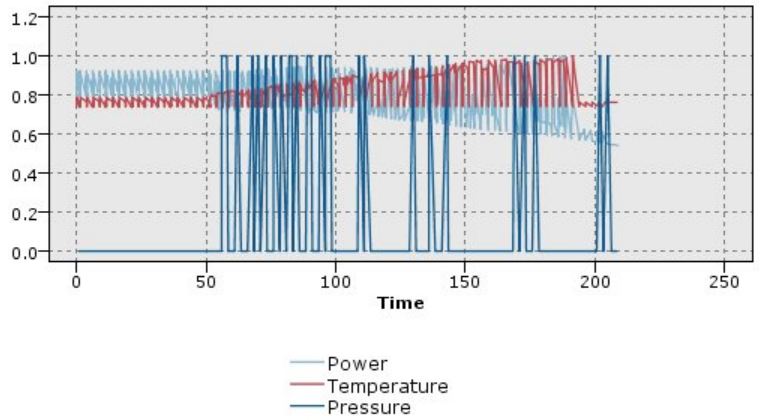


Figure 25. Graphique Courbes normalisé montrant le tracé de la pression

Fonction de superposition. Sélectionnez cette option pour indiquer la fonction connue à comparer aux valeurs réelles. Par exemple, pour comparer des valeurs réelles à des valeurs prédites, vous pouvez représenter la fonction `y = x` sous la forme d'une superposition. Indiquez une fonction **`y =`** dans la zone de texte. La fonction par défaut est `y = x`, mais vous pouvez définir toutes sortes de fonctions, telles qu'une fonction quadratique ou une expression arbitraire en termes de `x`.

Remarque : les fonctions de superposition ne sont pas disponibles pour les panneaux ou les graphiques animés.

Lorsque le nombre d'enregistrements est supérieur à. Spécifiez la méthode de représentation des jeux de données volumineux. Vous pouvez spécifier la taille maximale des jeux de données ou utiliser les 2 000 points par défaut. Lorsque vous sélectionnez les options **Intervalle** ou **Echantillon**, les performances des jeux de données volumineux sont optimisées. Vous pouvez également choisir de représenter tous les points de données en sélectionnant **Utiliser toutes les données**, mais sachez que vous risquez de réduire considérablement les performances du logiciel.

Remarque : lorsque le mode X est paramétré sur **Superposer** ou sur **Selon lecture**, ces options sont désactivées et seuls les *n* premiers enregistrements sont utilisés.

- **Casier.** Sélectionnez cette option pour permettre la création de casiers lorsque le jeu de données contient plus d'enregistrements que le nombre spécifié. La création d'intervalles applique une fine grille au graphique avant que soit effectué le traçage réel et compte le nombre de connexions apparaissant dans chacune des cellules de la grille. Dans le graphique final, une connexion est représentée dans chaque cellule, au niveau du centroïde de l'intervalle (moyenne de tous les emplacements de connexion de l'intervalle).
- **Echantillon.** Sélectionnez cette option pour échantillonner de façon aléatoire les données dans le nombre d'enregistrements spécifié.

Onglet Apparence de courbes

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous-titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Libellé X. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *x* (horizontal), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Libellé Y. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *y* (vertical), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Utilisation d'un graphique Courbes

Les graphiques Tracé et Courbes sont principalement basés sur la comparaison des valeurs *X* par rapport aux valeurs *Y*. Par exemple, si vous étudiez une fraude potentielle en matière de demande de subventions agricoles, vous pouvez comparer, par le biais d'un réseau de neurones, le revenu réclamé dans la demande à son estimation. L'utilisation d'une superposition, telle que le type de culture, permettra de démontrer s'il existe un lien entre la demande (valeur ou nombre) et le type de culture.

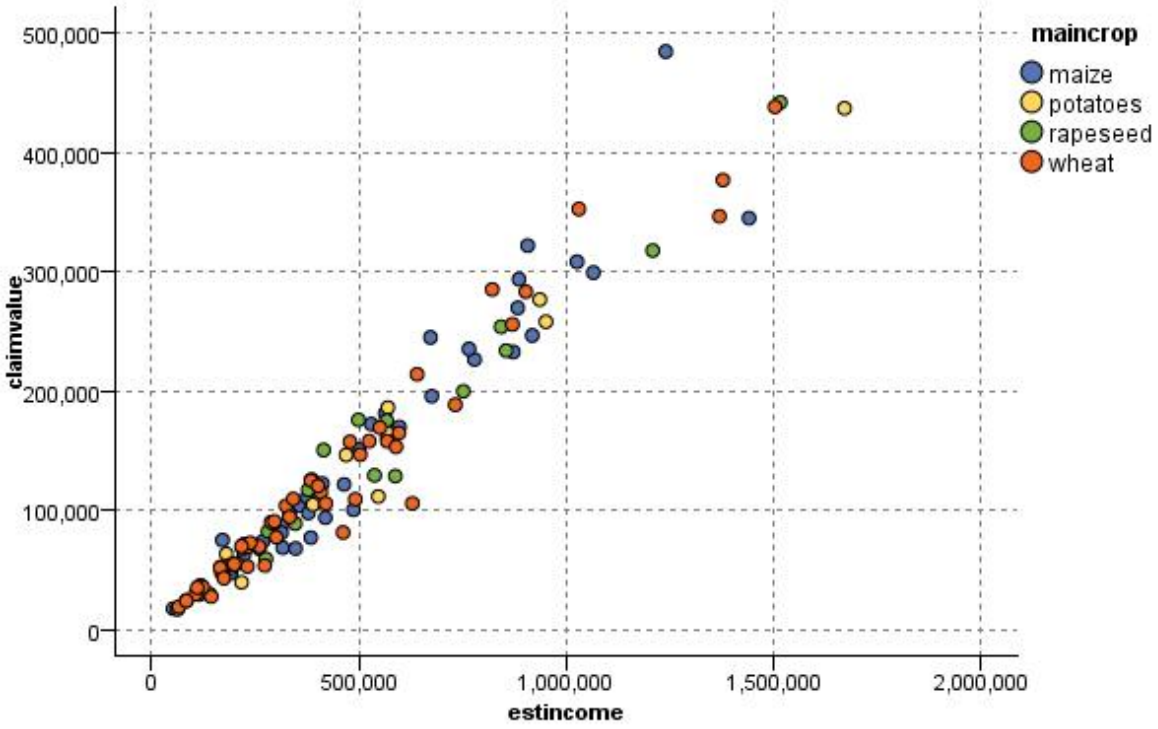


Figure 26. Tracé représentant la relation entre le revenu estimé et la valeur de la demande, le type de culture principale servant de superposition

Les graphiques Nuage, Courbes et Evaluation étant des illustrations en deux dimensions de la comparaison entre *Y* et *X*, il est facile d'interagir avec ceux-ci en définissant des zones, en marquant un élément, ou en traçant des bandes. Vous pouvez également générer des noeuds pour les données représentées par ces zones, bandes, ou éléments. Pour plus d'informations, voir «Exploration de graphiques», à la page 290.

Noeud Tracé horaire

Les noeuds Tracé horaire vous permettent de visualiser la représentation d'une ou de plusieurs séries temporelles au fil du temps. Les séries représentées doivent contenir des valeurs numériques et sont supposées avoir lieu sur une durée au sein de laquelle les périodes sont uniformes.

Dans SPSS Modeler version 17.1 et les versions antérieures, vous utilisez généralement un noeud Intervalle de temps avant un noeud Tracé horaire pour créer un champ *TimeLabel*, lequel est utilisé par défaut pour désigner l'axe *x* des graphiques.

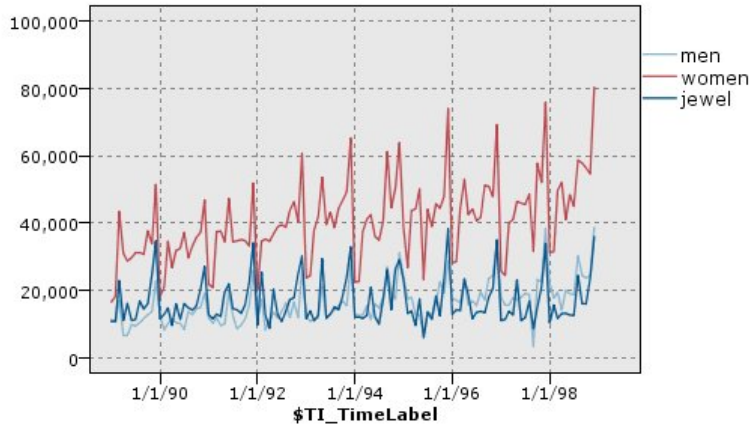


Figure 27. Représentation des ventes de vêtements et de bijoux homme et femme dans le temps

Création d'interventions et d'événements

Vous pouvez créer des champs d'événement et d'intervention à partir du Tracé horaire en générant un noeud Calculer (indicateur ou nominal) à partir des menus contextuels. Par exemple, vous pourriez créer un champ d'événements en cas de grève des chemins de fer, où les conditions de circulation sont True (vrai) si l'événement est survenu et False (faux) dans le cas contraire. Pour un champ d'intervention, une augmentation de prix par exemple, vous pouvez utiliser un noeud de calcul pour identifier la date de l'augmentation, avec 0 pour l'ancien prix et 1 pour le nouveau. Pour plus d'informations, voir «Noeud Calculer», à la page 162.

Onglet Tracé horaire

Tracé. Permet de choisir comment tracer les séries temporelles.

- **Champs sélectionnés.** Trace des valeurs pour les séries temporelles sélectionnées. Si vous sélectionnez cette option lors du tracé des intervalles de confiance, désélectionnez la case **Normaliser**.
- **Modèles de série temporelle sélectionnés**. Utilisée en association avec un modèle de séries temporelles, cette option trace tous les champs liés (valeurs réelles et prédites, et intervalles de confiance) pour une ou plusieurs séries temporelles sélectionnées. Cette option désactive d'autres options de la boîte de dialogue. Il s'agit de l'option recommandée pour le tracé d'intervalles de confiance.

Série. Sélectionnez un ou plusieurs champs contenant des séries temporelles à représenter. Il doit s'agir de données numériques.

Libellé de l'axe X. Choisissez le libellé par défaut ou un champ unique à utiliser en tant que libellé de l'axe *x* dans les graphiques Tracé. Si vous choisissez Par défaut, le système utilise le champ TimeLabel créé à partir d'un noeud Intervalles de temps en amont (pour les flux créés dans SPSS Modeler version 17.1 et versions antérieures), ou d'entiers séquentiels s'il n'existe pas de noeud Intervalles de temps en amont.

Afficher les séries dans des panneaux distincts. Indique si chaque série apparaît dans un panneau distinct. Si vous ne choisissez pas cette option, toutes les séries temporelles sont représentées sur le même graphique et les lissages ne sont pas disponibles. Dans le cas d'une représentation sur un même graphique, chaque série arbore une couleur différente.

Normaliser. Sélectionnez cette option pour mettre toutes les valeurs *Y* à l'échelle sur l'intervalle 0–1 afin de les afficher sur le graphique. La fonction de normalisation vous permet d'explorer les relations existant entre les lignes, relations qui risqueraient sinon d'être occultées en raison des différences au niveau de

l'intervalle de valeurs de chaque série ; il est recommandé de l'utiliser lorsque vous représentez plusieurs lignes sur le même graphique ou lorsque vous comparez des graphiques dans des panneaux mitoyens. (Il est inutile d'appliquer une normalisation lorsque toutes les valeurs de données sont comprises dans un même intervalle.)

Afficher. Sélectionnez un ou plusieurs éléments à afficher dans votre graphique Nuage. Vous avez le choix entre des lignes, des points et des lissages (LOESS). Les lissages sont disponibles uniquement si vous affichez les séries dans des panneaux distincts. Par défaut, l'élément ligne est sélectionné. Veillez à sélectionner au moins un élément de graphique Tracé avant d'exécuter le noeud Graphique, sinon le système renvoie une erreur indiquant que vous n'avez sélectionné aucun élément à représenter.

Limiter le nombre d'enregistrements. Sélectionnez cette option si vous souhaitez limiter le nombre d'enregistrements représentés. Spécifiez le nombre d'enregistrements, lus à partir du début de votre fichier de données, qui seront représentés dans l'option **Nombre maximal d'enregistrements à représenter graphiquement**. Ce nombre est défini sur 2 000 par défaut. Pour représenter les *n* derniers enregistrements de votre fichier, vous pouvez utiliser un noeud Trier avant ce noeud pour organiser les enregistrements dans l'ordre temporel décroissant.

Onglet Apparence du tracé horaire

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous-titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Libellé X. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *x* (horizontal), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Libellé Y. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *y* (vertical), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Présentation. Pour les tracés horaires uniquement, vous pouvez indiquer si les valeurs temporelles doivent être représentées sur un axe horizontal ou sur un axe vertical.

Utilisation d'un graphique Tracé horaire

Une fois que vous avez créé un graphique Tracé horaire, vous disposez de plusieurs options pour ajuster l'affichage du graphique et générer des noeuds pour une analyse plus approfondie. Pour plus d'informations, voir «Exploration de graphiques», à la page 290.

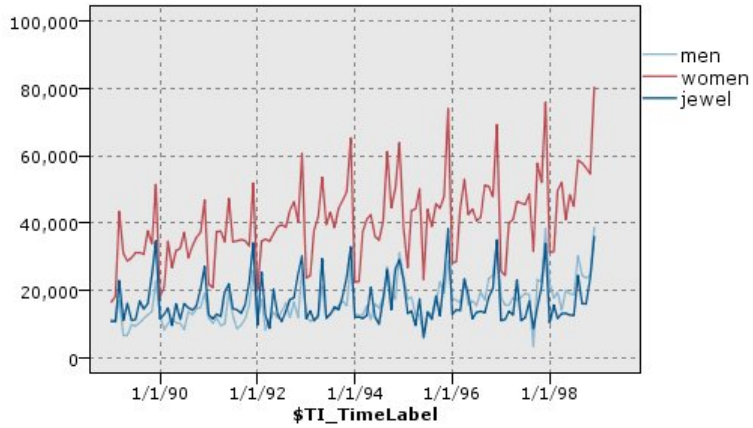


Figure 28. Représentation des ventes de vêtements et de bijoux homme et femme dans le temps

Après avoir créé un graphique Tracé horaire, défini des bandes et examiné les résultats, vous pouvez utiliser les options du menu Générer et du menu contextuel pour créer des noeuds Sélectionner ou Calculer. Pour plus d'informations, voir «Génération de noeuds à partir de graphiques», à la page 298.

Noeud Proportion

Les graphiques ou les tableaux Proportion montrent l'occurrence, dans un jeu de données, de valeurs symboliques (non numériques) comme un type de prêt hypothécaire ou le sexe d'un individu. Les noeuds distribution servent souvent à montrer les déséquilibres des données, déséquilibres pouvant être rectifiés grâce à l'utilisation d'un noeud Equilibrer avant la création d'un modèle. Vous pouvez générer automatiquement un noeud Equilibrer à l'aide du menu Générer de la fenêtre d'un graphique ou d'un tableau Proportion.

Vous pouvez aussi utiliser le noeud Représentation graphique pour produire des graphiques à barres. Néanmoins, vous pouvez choisir parmi davantage d'options dans ce noeud. Pour plus d'informations, voir «Types de visualisation des Représentations graphiques intégrées disponibles», à la page 208.

Remarque : pour montrer l'occurrence de valeurs numériques, utilisez de préférence un noeud Histogramme.

Onglet Nuage de proportion

Tracé. Sélectionnez le type de proportion. Sélectionnez **Champs sélectionnés** pour afficher la proportion du champ sélectionné. Sélectionnez **Tous les booléens (valeurs vraies)** pour afficher la proportion des valeurs true (vrai) des champs booléens du jeu de données.

Champ. Sélectionnez le champ nominal ou le champ indicateur dont vous souhaitez montrer la proportion des valeurs. Seuls les champs n'ayant pas été explicitement définis comme numériques sont répertoriés dans la liste.

Superposer. Sélectionnez le champ nominal ou le champ indicateur à utiliser en tant que superposition de couleurs, illustrant la proportion de ses valeurs au sein de chaque valeur du champ spécifié. Par exemple, vous pouvez utiliser la réponse à une campagne de marketing (*pep*) comme superposition au nombre d'enfants (*enfant*) afin d'illustrer la réactivité en fonction de la taille de la famille. Pour plus d'informations, voir «Apparences, superpositions, panneaux et animation», à la page 198.

Normaliser par couleur. Sélectionnez cette option pour mettre les barres à l'échelle de sorte que toutes les barres occupent la totalité du graphique. Les valeurs de superposition correspondent à une proportion de chaque barre, ce qui facilite les comparaisons entre les catégories.

Trier. Sélectionne la méthode utilisée pour afficher les valeurs sur le graphique de la distribution. Sélectionnez **Alphabétique** pour utiliser le classement par ordre alphabétique ou **Par effectif** pour répertorier les valeurs par ordre décroissant d'occurrence.

Echelle proportionnelle. Sélectionnez cette option pour mettre la proportion des valeurs à l'échelle de sorte que la valeur la plus représentée occupe la totalité du tracé. Toutes les autres barres sont mises à l'échelle en fonction de cette valeur. Désélectionnez cette option pour mettre les barres à l'échelle en fonction du total de chaque valeur.

Onglet Apparence de proportion

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous-titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Libellé X. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *x* (horizontal), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Libellé Y. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *y* (vertical), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Utilisation d'un noeud Proportion

Les noeuds Proportion sont utilisés pour montrer la proportion des valeurs symboliques dans un jeu de données. Ils sont fréquemment utilisés avant les noeuds de manipulation pour explorer les données et corriger les déséquilibres. Par exemple, si les instances des personnes sans enfant interrogées sont beaucoup plus nombreuses que celles des autres types de personne interrogée, vous souhaiterez peut-être réduire le nombre de ces instances afin de pouvoir générer une règle plus utile pour vos opérations d'exploration de données ultérieures. Les noeuds Proportion vous aident à examiner ces déséquilibres et à prendre des décisions les concernant.

Le noeud Proportion est particulier car il produit à la fois un graphique et un tableau pour analyser vos données.

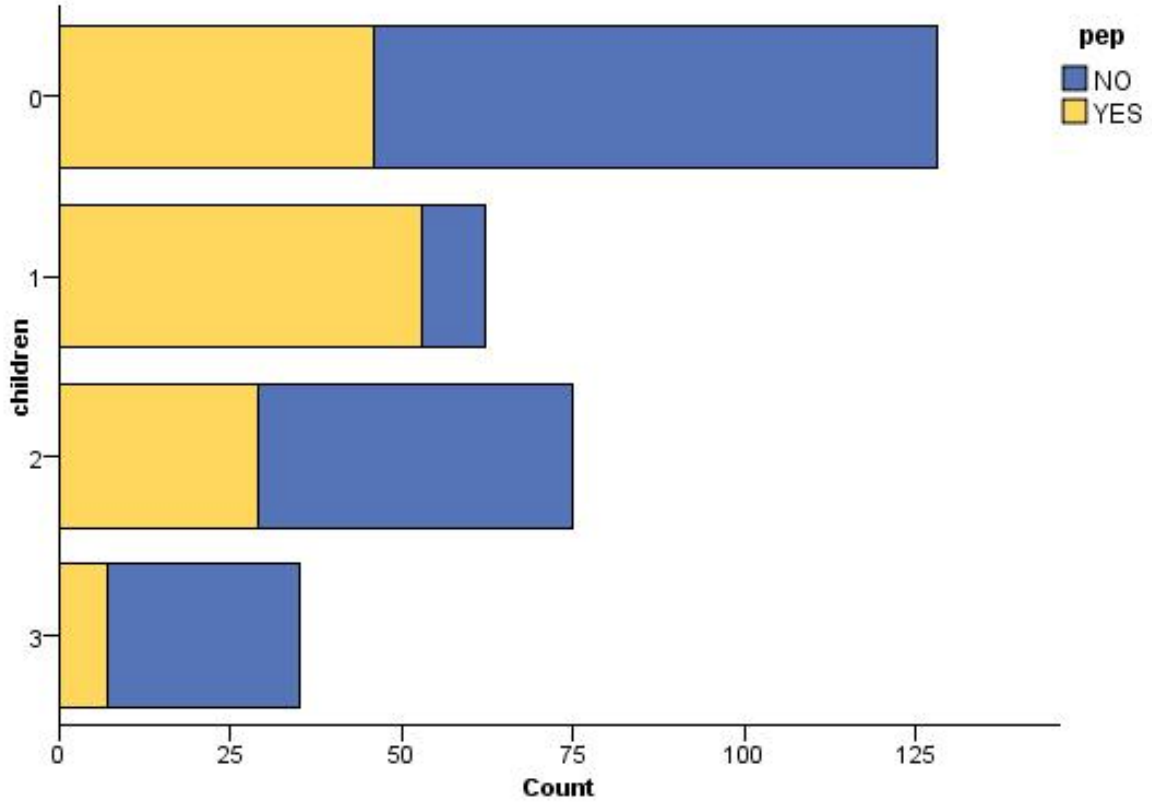


Figure 29. Graphique Proportion affichant le nombre de personnes avec ou sans enfants qui ont répondu à une campagne de marketing

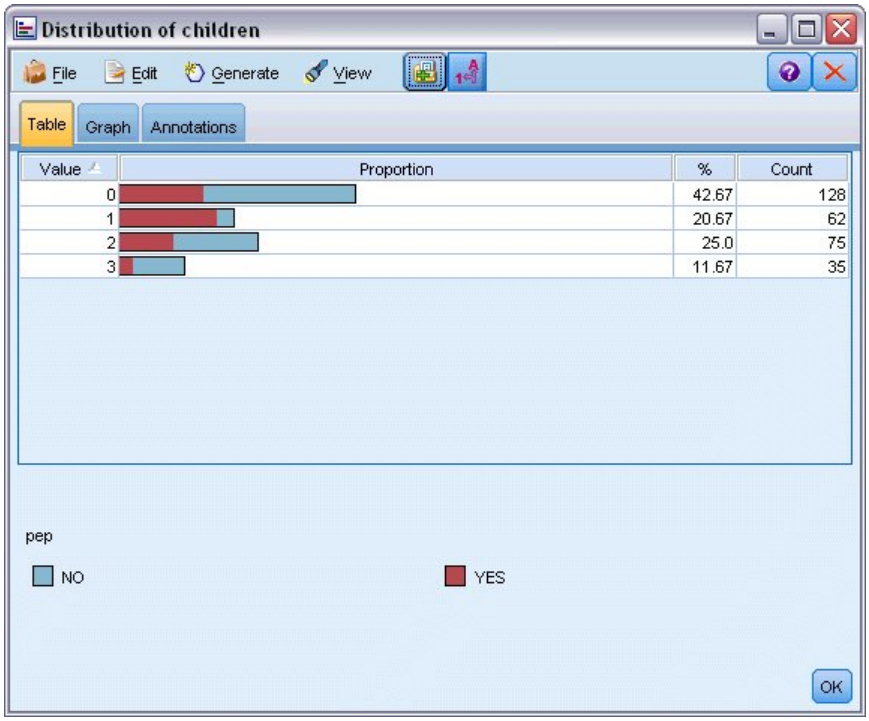


Figure 30. Tableau Proportion affichant la proportion de personnes avec ou sans enfants qui ont répondu à une campagne de marketing

Une fois que vous avez créé un graphique et un tableau Proportion et examiné les résultats, vous pouvez utiliser les options des menus pour regrouper et copier les valeurs, ainsi que pour générer un certain nombre de noeuds pour la préparation des données. En outre, vous pouvez copier ou exporter les informations du graphique et du tableau pour les utiliser dans d'autres applications, par exemple MS Word ou MS PowerPoint. Pour plus d'informations, voir «Impression, enregistrement, copie et exportation de graphiques», à la page 314.

Pour sélectionner et copier des valeurs à partir d'un tableau Proportion

1. Cliquez sur les lignes tout en maintenant le bouton de la souris enfoncé afin de sélectionner un ensemble de valeurs. Vous pouvez aussi utiliser l'option **Sélectionner tout** du menu Edition pour sélectionner toutes les valeurs.
2. Dans le menu Edition, sélectionnez **Copier la table** ou **Copier la table (avec les noms de champ)**.
3. Collez les valeurs dans le Presse-papiers ou dans l'application souhaitée.

   Remarque : Les barres ne sont pas copiées directement. A la place, ce sont les valeurs du tableau qui sont copiées. Autrement dit, les valeurs superposées ne figurent pas dans le tableau copié.

Pour regrouper des valeurs à partir d'un tableau Proportion

1. Sélectionnez les valeurs à regrouper à l'aide de la méthode Ctrl+clic.
2. Dans le menu Edition, sélectionnez **Associer**.

Remarque : Lorsque vous associez et dissociez des valeurs, le graphique de l'onglet Graphique est automatiquement redessiné pour refléter les modifications.

Vous pouvez également :

- Dissocier les valeurs d'un groupe en sélectionnant le nom de ce groupe dans la liste des proportions et en choisissant **Dissocier** dans le menu Edition.

- Editer un groupe en sélectionnant son nom dans la liste des proportions et en choisissant **Editer le groupe** dans le menu Edition. Dans la boîte de dialogue qui apparaît, vous pouvez ajouter des valeurs au groupe ou les en supprimer.

Options du menu Générer

Vous pouvez utiliser les options du menu Générer pour sélectionner un sous-ensemble de données, calculer un champ booléen, regrouper des valeurs, reclassifier des valeurs ou équilibrer les données d'un graphique ou d'un tableau. Ces opérations génèrent un noeud Préparation des données et le placent dans l'espace de travail de flux. Pour utiliser le noeud généré, connectez-le à un flux existant. Pour plus d'informations, voir «Génération de noeuds à partir de graphiques», à la page 298.

Noeud Histogramme

Les noeuds Histogramme montrent l'occurrence des valeurs des champs numériques. Ils sont souvent utilisés pour explorer les données avant toute génération de modèle ou manipulation. Semblables au noeud distribution, les noeuds Histogramme servent souvent à montrer les déséquilibres des données. Vous pouvez utiliser le noeud Représentation graphique pour produire un histogramme, mais vous pouvez également sélectionner davantage d'options dans ce noeud. Pour plus d'informations, voir «Types de visualisation des Représentations graphiques intégrées disponibles», à la page 208.

Remarque : pour montrer l'occurrence des valeurs des champs symboliques, utilisez un noeud distribution.

Onglet Tracé d'histogramme

Champ. Sélectionnez le champ numérique dont vous souhaitez montrer la proportion des valeurs. Seuls les champs n'ayant pas été explicitement définis comme symboliques (catégoriels) sont répertoriés.

Superposition. Sélectionnez un champ symbolique afin de montrer les catégories de valeurs du champ spécifié. La sélection d'un champ de superposition transforme le graphique Histogramme en un graphique aux valeurs superposées. Les couleurs servent à représenter les différentes catégories du champ de superposition. Le noeud Histogramme met à disposition trois types de superpositions : couleur, panneau et animation. Pour plus d'informations, voir «Apparences, superpositions, panneaux et animation», à la page 198.

Onglet Options d'histogramme

Amplitude X automatique. Sélectionnez cette option pour utiliser l'intégralité de l'amplitude de valeurs des données sur cet axe. Désélectionnez cette option pour utiliser un sous-ensemble de valeurs explicite déterminé par les valeurs **Minimum** et **Maximum** spécifiées. Vous pouvez entrer les valeurs ou utiliser les flèches. Les amplitudes automatiques sont sélectionnées par défaut pour permettre la création rapide de graphiques.

Casiers. Sélectionnez soit **Par nombre** soit **Par largeur**.

- Sélectionnez **Par nombre** pour afficher un nombre fixe de barres dont la largeur dépend de l'amplitude et du nombre de casiers spécifiés. Indiquez le nombre de casiers à utiliser dans le graphique dans l'option **Nbre de casiers**. Utilisez les flèches pour rectifier le nombre.
- Sélectionnez **Par largeur** pour créer un graphique formé de barres de largeur fixe. Le nombre d'intervalles dépend de la largeur indiquée et de l'amplitude de valeurs. Indiquez la largeur des barres dans l'option **Largeur de casier**.

Normaliser par couleur. Sélectionnez cette option pour attribuer la même hauteur à toutes les barres, les valeurs superposées étant affichées sous la forme d'un pourcentage de la totalité des observations dans chaque barre.

Afficher la courbe normale. Sélectionnez cette option pour ajouter une courbe normale au graphique affichant la moyenne et la variance des données.

Bandes séparées pour chaque couleur. Sélectionnez cette option pour afficher chaque valeur superposée sous la forme d'une bande distincte sur le graphique.

Onglet Apparence d'histogramme

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous-titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Libellé X. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *x* (horizontal), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Libellé Y. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *y* (vertical), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Utilisation des histogrammes

Les graphiques Histogramme montrent la proportion des valeurs d'un champ numérique dont les valeurs sont situées sur l'axe *x*. Les histogrammes fonctionnent de la même manière que les graphiques Résumés. Les graphiques Résumé montrent la proportion des valeurs d'un champ numérique *par rapport à celles d'un autre*, plutôt que l'occurrence de valeurs d'un champ unique.

Lorsque vous avez créé un graphique, vous pouvez observer les résultats et déterminer des bandes pour diviser les valeurs le long de l'axe *x* ou définir des régions. Vous pouvez également marquer des éléments dans le graphique. Pour plus d'informations, voir «Exploration de graphiques», à la page 290.

Vous pouvez utiliser les options du menu Générer pour créer des noeuds Equilibrer, Sélectionner ou Calculer à l'aide des données du graphique ou plus spécifiquement dans les bandes, régions ou les éléments marqués. Ce type de graphique est souvent utilisé avant les noeuds de manipulation pour explorer les données et corriger les éventuels déséquilibres en générant un noeud Equilibrer à partir du graphique à utiliser dans le flux. Vous pouvez également générer un noeud booléen Calculer pour ajouter un champ montrant à quelle bande appartient chaque enregistrement ou un noeud Sélectionner pour sélectionner tous les enregistrements appartenant à un ensemble ou à une amplitude spécifique de valeurs. Ces opérations vous aident à vous concentrer sur un sous-ensemble particulier de données pour procéder à une exploration plus approfondie. Pour plus d'informations, voir «Génération de noeuds à partir de graphiques», à la page 298.

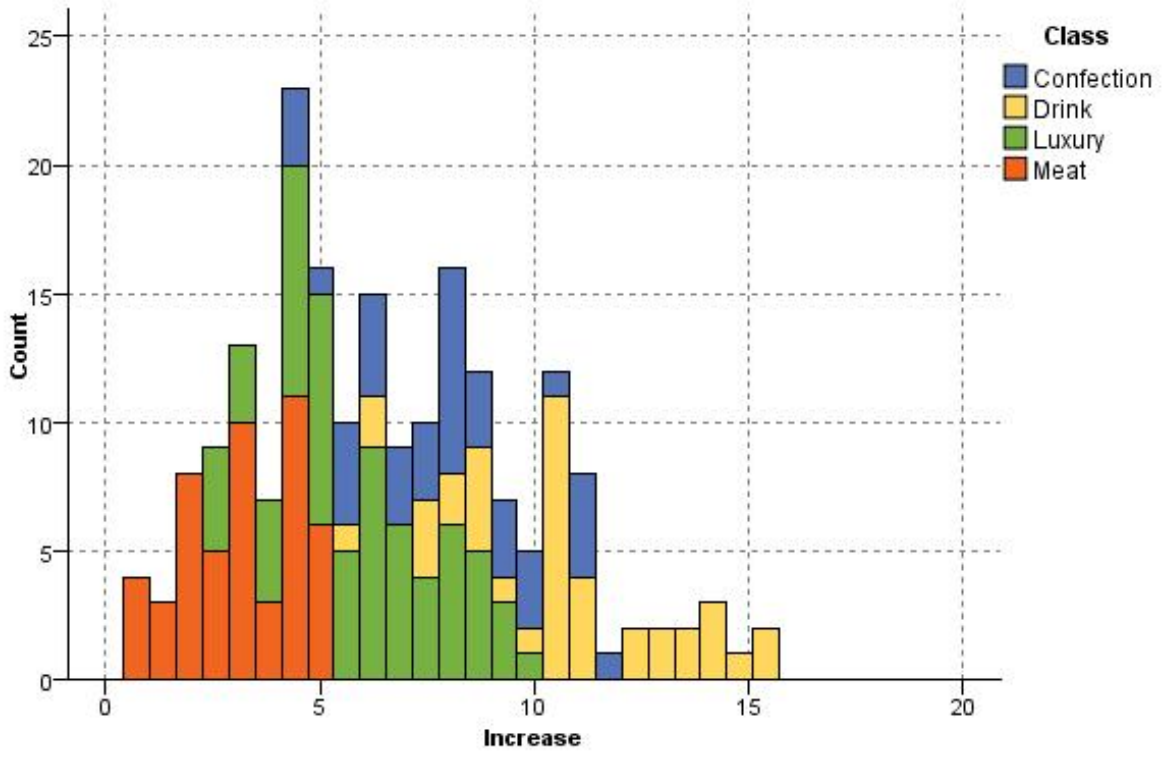


Figure 31. Graphique Histogramme montrant la proportion par catégorie de la hausse d'achat due à une promotion

Noeud Résumé

Les graphiques Résumé sont semblables aux graphiques Histogramme, excepté que les graphiques Résumé montrent la proportion des valeurs d'un champ numérique par rapport à celles d'un autre, plutôt que l'occurrence de valeurs d'un champ unique. Les graphiques Résumé sont utiles pour illustrer une variable ou un champ dont les valeurs changent avec le temps. Grâce à la représentation graphique en 3D, vous pouvez en outre inclure un axe symbolique affichant les proportions par catégorie. Des Résumés bidimensionnels sont affichés sous forme de graphiques à barres empilées, avec des superpositions le cas échéant. Pour plus d'informations, voir «Apparences, superpositions, panneaux et animation», à la page 198.

Onglet nuage de Résumé

Résumé. Sélectionnez le champ dont les valeurs doivent être résumées et affichées en fonction de l'amplitude de valeurs du champ défini dans **Sur**. Seuls les champs n'ayant pas été indiqués comme symboliques sont répertoriés.

Sur. Sélectionnez le champ dont les valeurs doivent être utilisées pour afficher le champ défini dans **Résumé**.

Par. Activée lors de la création de graphiques en 3D, cette option vous permet de sélectionner le champ nominal ou le champ indicateur utilisé pour l'affichage du champ de résumé par catégorie.

Opération. Permet de sélectionner ce que représente chaque barre du graphique Résumé. Les options disponibles sont **Somme**, **Moyenne**, **Maximum**, **Minimum** et **Ecart type**.

Superposition. Sélectionnez un champ symbolique afin de montrer les catégories de valeurs du champ sélectionné. La sélection d'un champ de superposition transforme le graphique Résumé et crée plusieurs barres de différentes couleurs pour chaque catégorie. Ce noeud dispose de trois types de superposition : couleur, panneau et animation. Pour plus d'informations, voir «Apparences, superpositions, panneaux et animation», à la page 198.

Onglet Options de résumé

Amplitude X automatique. Sélectionnez cette option pour utiliser l'intégralité de l'amplitude de valeurs des données sur cet axe. Désélectionnez cette option pour utiliser un sous-ensemble de valeurs explicite déterminé par les valeurs **Minimum** et **Maximum** spécifiées. Vous pouvez entrer les valeurs ou utiliser les flèches. Les amplitudes automatiques sont sélectionnées par défaut pour permettre la création rapide de graphiques.

Casiers. Sélectionnez soit **Par nombre** soit **Par largeur**.

- Sélectionnez **Par nombre** pour afficher un nombre fixe de barres dont la largeur dépend de l'amplitude et du nombre de casiers spécifiés. Indiquez le nombre de casiers à utiliser dans le graphique dans l'option **Nbre de casiers**. Utilisez les flèches pour rectifier le nombre.
- Sélectionnez **Par largeur** pour créer un graphique formé de barres de largeur fixe. Le nombre d'intervalles dépend de la largeur indiquée et de l'amplitude de valeurs. Indiquez la largeur des barres dans l'option **Largeur de casier**.

Onglet Apparence de résumé

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous-titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Sur le libellé. Vous pouvez soit accepter le libellé généré automatiquement, soit sélectionner **Personnalisé** pour indiquer un libellé.

Collecte des libellés. Vous pouvez soit accepter le libellé généré automatiquement, soit sélectionner **Personnalisé** pour indiquer un libellé.

Par libellé. Vous pouvez soit accepter le libellé généré automatiquement, soit sélectionner **Personnalisé** pour indiquer un libellé.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Les exemples suivants montrent où sont placées les options d'apparence dans une version 3D du graphique.

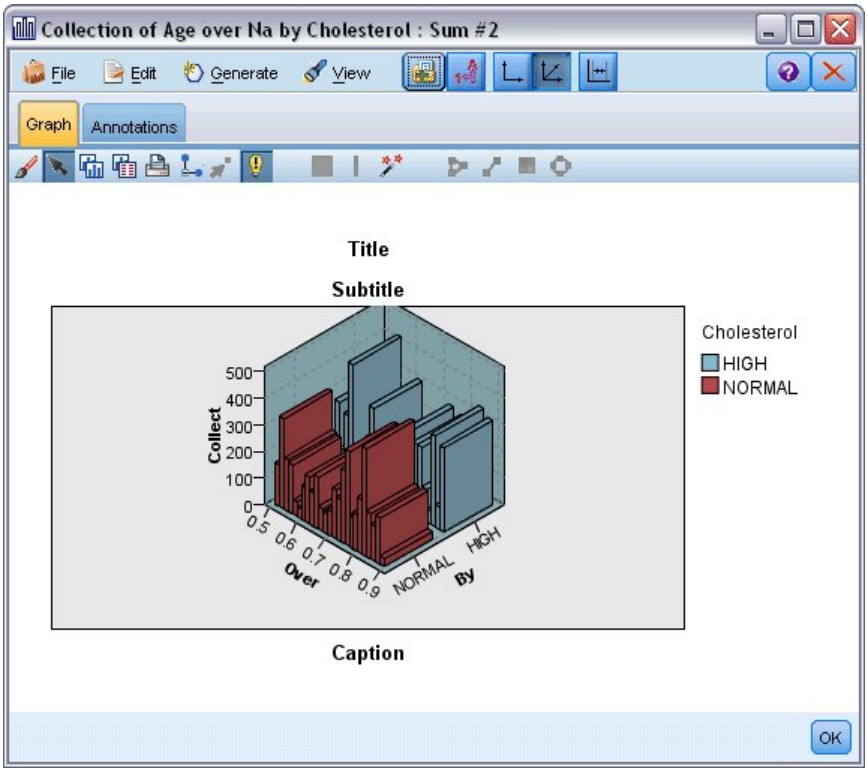


Figure 32. Position des options d'apparence du graphique sur un graphique résumé en 3D

Utilisation d'un graphique Résumé

Les graphiques Résumé montrent la proportion des valeurs d'un champ numérique *par rapport à celles d'un autre*, plutôt que l'occurrence de valeurs d'un champ unique. Les histogrammes fonctionnent de la même manière que les graphiques Résumés. Les graphiques Histogramme montrent la proportion des valeurs d'un champ numérique dont les valeurs sont situées sur l'axe *x*.

Lorsque vous avez créé un graphique, vous pouvez observer les résultats et déterminer des bandes pour diviser les valeurs le long de l'axe *x* ou définir des régions. Vous pouvez également marquer des éléments dans le graphique. Pour plus d'informations, voir «Exploration de graphiques», à la page 290.

Vous pouvez utiliser les options du menu Générer pour créer des noeuds Equilibrer, Sélectionner ou Calculer à l'aide des données du graphique ou plus spécifiquement dans les bandes, régions ou les éléments marqués. Ce type de graphique est souvent utilisé avant les noeuds de manipulation pour explorer les données et corriger les éventuels déséquilibres en générant un noeud Equilibrer à partir du graphique à utiliser dans le flux. Vous pouvez également générer un noeud booléen Calculer pour ajouter un champ montrant à quelle bande appartient chaque enregistrement ou un noeud Sélectionner pour sélectionner tous les enregistrements appartenant à un ensemble ou à une amplitude spécifique de valeurs. Ces opérations vous aident à vous concentrer sur un sous-ensemble particulier de données pour procéder à une exploration plus approfondie. Pour plus d'informations, voir «Génération de noeuds à partir de graphiques», à la page 298.

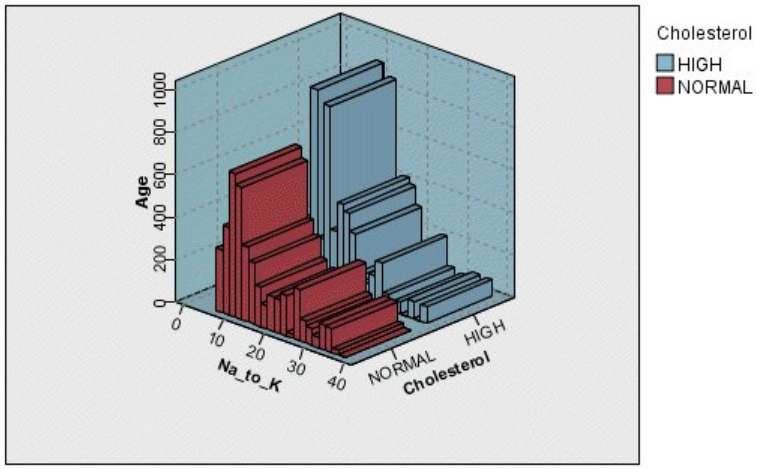


Figure 33. Graphique Résumé en 3D montrant la somme Na_sur_K par rapport à l'âge pour les niveaux de cholestérol élevés et normaux

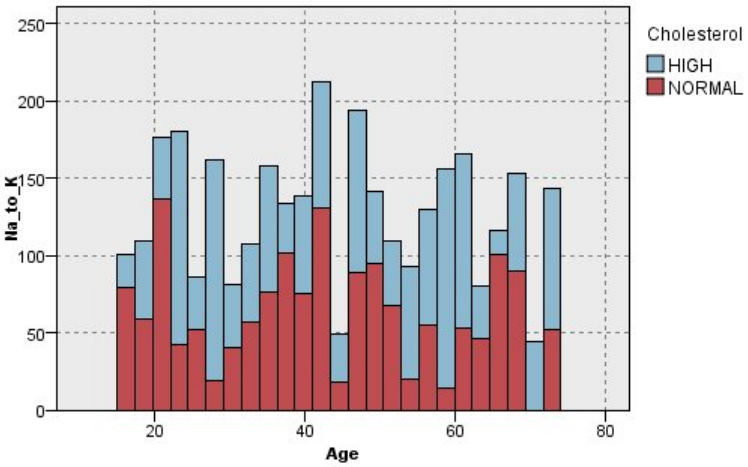


Figure 34. Graphique Résumé sans axe z, mais avec le taux de cholestérol affiché sous la forme d'une superposition de couleurs

Noeud Relations

Les noeuds Relations montrent la force des relations existant entre les valeurs de plusieurs champs symboliques. Le graphique affiche les connexions à l'aide de divers types de ligne indiquant la force de la connexion. Par exemple, vous pouvez utiliser un noeud Relations pour explorer la relation qui existe entre différents articles achetés sur un site de commerce électronique ou dans un magasin classique de vente au détail.

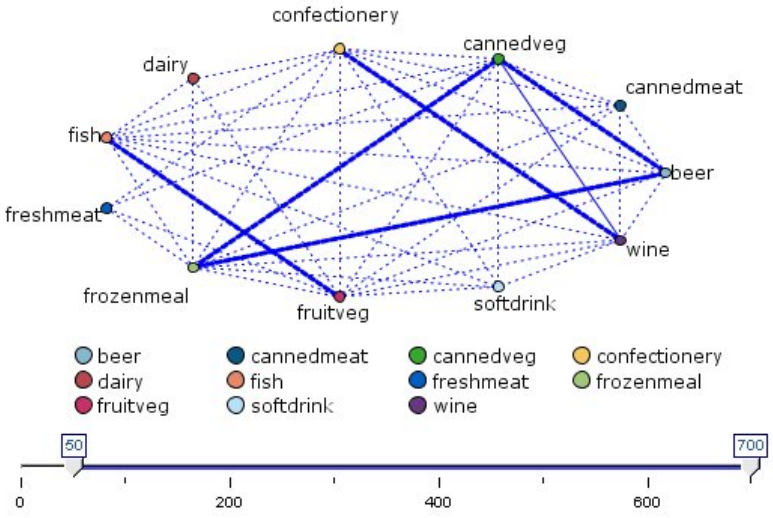


Figure 35. Graphique Relations montrant les relations entre les articles d'épicerie achetés

Relations orientées

Les noeuds Relations orientées sont semblables aux noeuds Relations car ils montrent la force des relations entre des champs symboliques. Cependant, les graphiques Relations orientées affichent uniquement les connexions d'un ou de plusieurs champs A partir de vers un seul champ Vers. Les connexions sont unidirectionnelles car ce sont des connexions unilatérales.

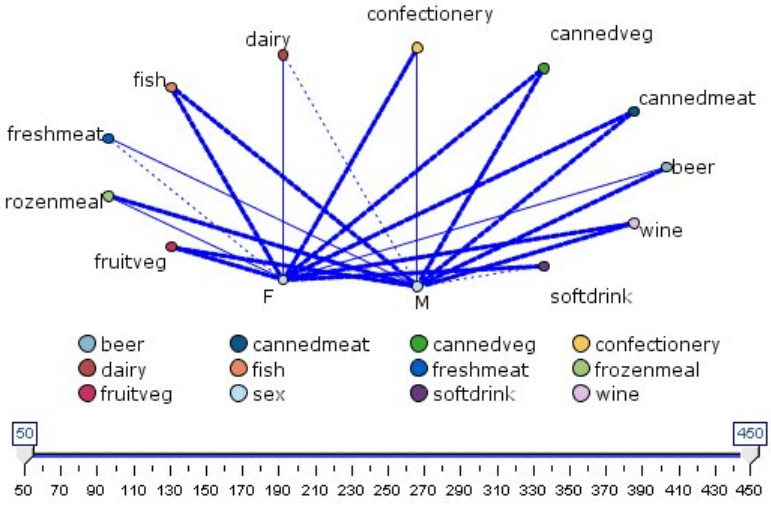


Figure 36. Graphique Relations orientées montrant la relation entre les articles d'épicerie achetés et le sexe de l'individu

Comme pour les noeuds Relations, le graphique affiche les connexions à l'aide de divers types de ligne indiquant la force de la connexion. Par exemple, vous pouvez utiliser un noeud Relations orientées pour explorer la relation existant entre le sexe de l'individu et la propension à acheter certains articles.

Onglet Graphique relations

Web. Sélectionnez cette option pour créer un graphique Relations illustrant la force des relations entre tous les champs spécifiés.

Relations orientées. Sélectionnez cette option pour créer un graphique Relations orientées illustrant la force des relations entre plusieurs champs et les valeurs d'un champ, comme le sexe d'un individu ou la religion. Lorsque cette option est sélectionnée, Champ Vers est activé et la commande Champs plus bas est renommée Champs A partir de pour plus de clarté.

Champ Vers (relations orientées uniquement). Sélectionnez un champ indicateur ou un champ nominal utilisé pour une relation orientée. Seuls les champs n'ayant pas été explicitement définis comme numériques sont répertoriés.

Champs/Champs A partir de. Sélectionnez les champs permettant de créer le graphique Relations. Seuls les champs n'ayant pas été explicitement définis comme numériques sont répertoriés. Utilisez le sélecteur de champs pour sélectionner plusieurs champs ou sélectionner les champs par type.

Remarque : dans les relations orientées, cette commande sert à sélectionner les champs A partir de.

Afficher uniquement les indicateur ayant une valeur vraie. Sélectionnez cette option pour afficher uniquement les booléens ayant une valeur true (vrai) pour un champ booléen. Cette option simplifie l'affichage des relations et est souvent utilisée avec les données pour lesquelles l'occurrence des valeurs positives est particulièrement élevée.

Valeurs des lignes. Sélectionnez le type de seuil dans la liste déroulante.

- L'option **Absolus** définit les seuils en fonction du nombre d'enregistrements contenant chaque paire de valeurs.
- L'option **Pourcentages globaux** indique le nombre absolu d'observations représentées par le lien sous la forme d'une proportion de toutes les occurrences de chaque paire de valeurs représentée dans le graphique Relations.
- Les options **Pourcentages de la plus petite valeur/du plus petit champ** et **Pourcentages de la plus grande valeur/du plus grand champ** indiquent le champ/la valeur à utiliser pour l'évaluation des pourcentages. Supposons, par exemple, que 100 enregistrements comportent la valeur *drugY* dans le champ *Médicament* et que seuls 10 enregistrements comportent la valeur *FAIBLE* dans le champ *BP*. Si 7 enregistrements comportent les deux valeurs *drugY* et *FAIBLE*, ce pourcentage est égal à 70 % ou à 7 %, selon le champ référencé, le plus petit (*BP*) ou le plus grand (*Drug*).

Remarque : avec les graphiques Relations orientées, les troisième et quatrième options mentionnées ci-dessus ne sont pas disponibles. En revanche, vous pouvez sélectionner **Pourcentage de valeur/champ Vers** et **Pourcentage de valeur/champ A partir de**.

Les liens forts sont plus volumineux. Sélectionnée par défaut, cette option correspond à la méthode standard d'affichage des liens entre les champs.

Les liens faibles sont plus volumineux. Sélectionnez cette option pour inverser la signification des liens représentés par des lignes en gras. Cette option est fréquemment utilisée pour détecter des fraudes ou examiner des valeurs éloignées.

Onglet Options de relations

L'onglet Options des noeuds Relations contient plusieurs options supplémentaires permettant de personnaliser le graphique de sortie.

Nombres de liens. Les options suivantes permettent de déterminer le nombre de liens affichés dans le graphique de sortie. Certaines de ces options, telles que **Liens faibles au-dessus de** et **Liens forts au-dessus de**, sont également disponibles dans la fenêtre du graphique de sortie. Dans le graphique final, vous pouvez également utiliser un curseur de défilement pour rectifier le nombre de liens affichés.

- **Nombre maximal de liens à afficher.** Choisissez un chiffre indiquant le nombre maximal de liens à afficher dans le graphique de sortie. Utilisez les flèches pour rectifier la valeur.

- **Afficher uniquement les liens au-dessus de.** Choisissez un chiffre indiquant la valeur minimale pour laquelle afficher une connexion sur le Web. Utilisez les flèches pour rectifier la valeur.
- **Afficher tous les liens.** Sélectionnez cette option pour afficher tous les liens sans tenir compte des valeurs minimale ou maximale. L'activation de cette option peut accroître le temps de traitement si le nombre de champs est important.

Supprimer si enregistrements peu nombreux. Sélectionnez cette option pour ignorer les connexions qui ne comportent que peu d'enregistrements. Définissez le seuil de cette option en entrant un nombre dans le champ **Enregistrements/ligne min.**.

Supprimer si enregistrements trop nombreux. Sélectionnez cette option pour ignorer les connexions comportant un nombre élevé d'enregistrements. Entrez un nombre dans le champ **Enregistrements/ligne max.**.

Liens faibles en dessous de. Choisissez un chiffre indiquant le seuil entre les connexions faibles (lignes en pointillé) et les connexions standard (lignes normales). Toutes les connexions au-dessous de cette valeur sont considérées comme faibles.

Liens forts au-dessus de. Choisissez un chiffre indiquant le seuil entre les connexions fortes (lignes en gras) et les connexions standard (lignes normales). Toutes les connexions au-dessus de cette valeur sont considérées comme fortes.

Taille du lien. Choisissez les options permettant de déterminer la taille des liens :
- **La taille des liens varie en permanence.** Sélectionnez cette option pour afficher une amplitude de tailles de lien reflétant la variation des forces de connexion en fonction des valeurs de données réelles.
- **Les liens sont classés en trois catégories de taille : fort/normal/faible.** Sélectionnez cette option pour afficher trois forces de connexion : forte, normale et faible. Les points de césure de ces catégories peuvent être définis grâce aux options ci-avant, ainsi que dans le graphique final.

Affichage des relations. Sélectionnez le type d'affichage des relations :
- **Présentation en cercle.** Sélectionnez cette option pour utiliser l'affichage standard des relations.
- **Présentation en réseau.** Sélectionnez cette option pour utiliser un algorithme afin de regrouper les liens les plus forts. L'objectif est de mettre en évidence les liens forts en utilisant la différenciation spatiale et l'épaisseur des lignes.
- **Présentation orientée.** Sélectionnez cette option pour créer un affichage des relations orientées utilisant la sélection **Champ Vers** de l'onglet Tracé comme cible de la direction.
- **Présentation de grille.** Sélectionnez cette option pour créer un affichage des relations présenté sur une grille régulière.

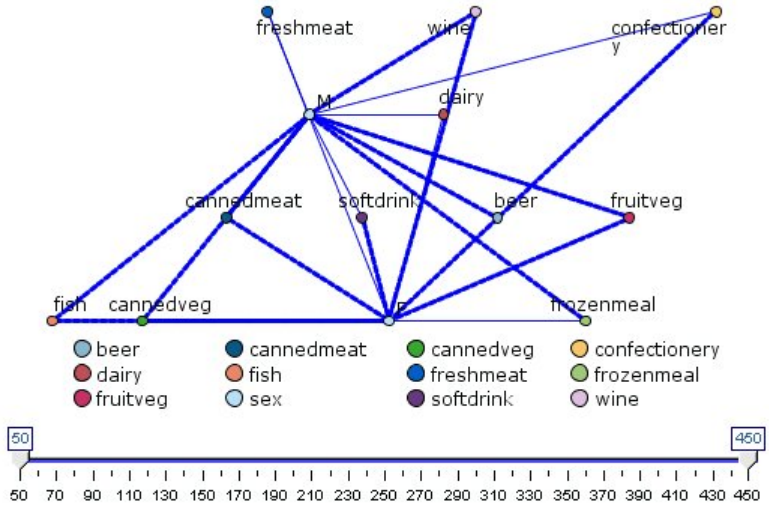


Figure 37. Graphique Relations montrant des connexions fortes entre les surgelés et les conserves de légumes d'une part, et les autres articles d'épicerie d'autre part

Remarque : Lors du filtrage des liens affichés (à l'aide du curseur du graphique Relations ou du contrôle **Afficher uniquement les liens au-dessus de** de l'onglet Options du noeud Relations), vous pouvez vous retrouver dans une situation où tous les liens qui restent à afficher possèdent la même valeur (en d'autres termes, ce sont tous des liens faibles, des liens normaux ou des liens forts, tels que définis par les contrôles **Liens faibles en dessous de** et **Liens forts au-dessus de** dans l'onglet Options du noeud Relations). Si cela se produit, tous les liens sont affichés avec la ligne d'épaisseur moyenne dans la sortie graphique Relations.

Onglet Apparence relations

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous-titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Afficher la légende. Vous pouvez spécifier si la légende apparaît. Masquer la légende peut améliorer l'apparence des tracés comportant de nombreux champs.

Utiliser les libellés en tant que noeuds. Vous pouvez insérer le texte du libellé dans chaque noeud au lieu d'afficher les libellés côte à côte. Pour les tracés dotés de peu de champs, cela risque de rendre le graphique plus lisible.

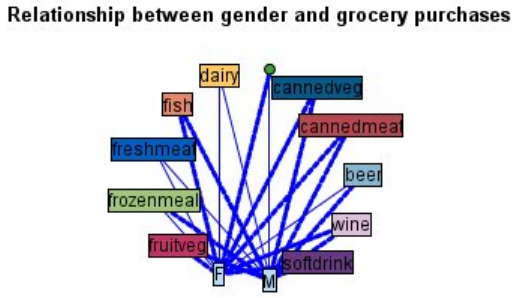


Figure 38. Graphique Relations affichant les libellés en tant que noeuds

Utilisation d'un graphique Relations

Les noeuds Relations sont utilisés pour montrer la force des liens entre les valeurs de plusieurs champs symboliques. Les connexions sont affichées dans un graphique composé de différents types de ligne servant à indiquer la force croissante des connexions. Vous pouvez, par exemple, utiliser un noeud Relations pour explorer le lien entre les niveaux de cholestérol et de tension artérielle, et le médicament le plus efficace pour traiter la maladie d'un patient.

- Les connexions fortes sont représentées à l'aide de lignes en gras. Ceci indique que les deux valeurs sont fortement liées et nécessitent une attention particulière.
- Les connexions moyennes sont représentées par des lignes d'épaisseur normale.
- Les connexions faibles sont représentées par des lignes en pointillé.
- Si aucune ligne n'apparaît entre deux valeurs, cela signifie soit que les deux valeurs n'apparaissent jamais dans le même enregistrement, soit que le nombre d'enregistrements contenant cette combinaison est inférieur au seuil défini dans la boîte de dialogue du noeud Relations.

Une fois que vous avez créé un noeud Relations, vous disposez de plusieurs options pour ajuster l'affichage du graphique et générer des noeuds pour une analyse plus approfondie.

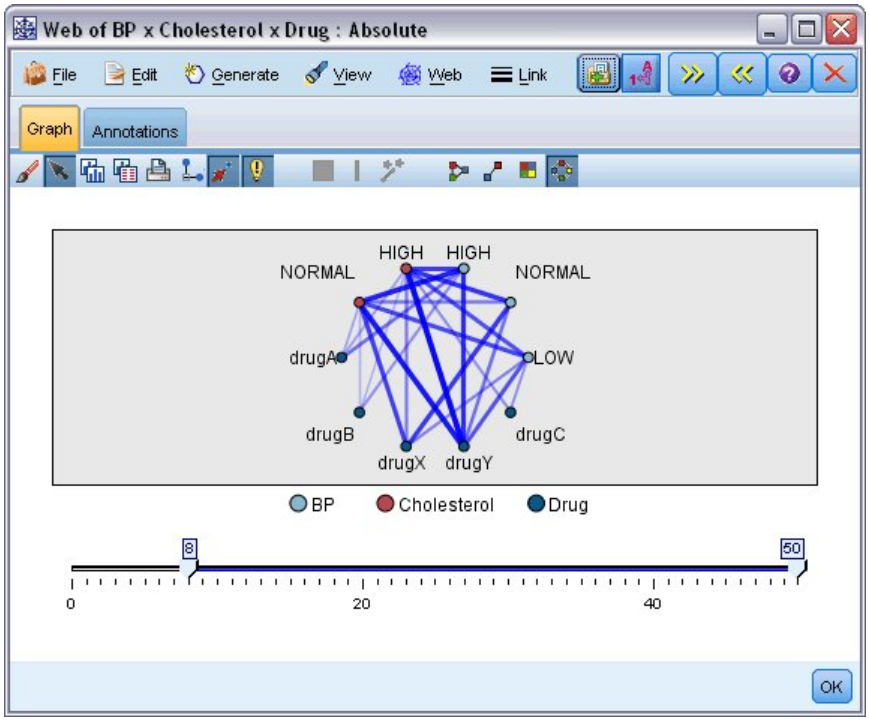


Figure 39. Graphique Relations indiquant un certain nombre de relations fortes, comme celle entre la tension artérielle normale et le médicament MédX, ou entre le taux de cholestérol élevé et MédY

Pour les noeuds Relations et Relations orientées, vous pouvez :

- Modifier la présentation de l'affichage des relations.
- Masquer des points pour simplifier l'affichage.
- Modifier les seuils qui gèrent les styles de ligne.
- Mettre en surbrillance des lignes entre certaines valeurs pour indiquer qu'il s'agit d'une relation "sélectionnée".
- Générer un noeud Sélectionner pour un ou plusieurs enregistrements "sélectionnés", ou un noeud booléen Calculer associé à une ou plusieurs relations du graphique Relations

Ajustement des points

- **Déplacez** les points en cliquant dessus à l'aide de la souris et en les faisant glisser jusqu'à l'emplacement voulu. Le graphique sera redessiné pour faire apparaître le nouvel emplacement.
- **Masquez** les points en cliquant dessus à l'aide du bouton droit de la souris et en sélectionnant **Masquer** ou **Masquer et redessiner** dans le menu contextuel. L'option **Masquer** masque uniquement le point sélectionné et toute ligne associée à ce point. **Masquer et redessiner** redessine le graphique, en tenant compte de vos modifications. Toutes les modifications manuelles sont annulées.
- **Affichez** tous les points masqués en sélectionnant **Tout afficher** ou **Tout afficher et redessiner** dans le menu Relations de la fenêtre du graphique. Sélectionnez **Tout afficher et redessiner** pour redessiner le graphique et effectuer les ajustements nécessaires pour inclure tous les points précédemment masqués, ainsi que leurs connexions.

Sélection ou "mise en évidence" de lignes

Les lignes sélectionnés sont surlignées en rouge.

1. Pour sélectionner une seule ligne, cliquez dessus avec le bouton gauche.
2. Pour sélectionnez plusieurs lignes, effectuez l'une des actions suivantes :

- A l'aide du curseur, dessinez un cercle autour des points des lignes que vous souhaitez sélectionnez.

- Maintenez la touche CTRL enfoncée et cliquez sur les lignes à sélectionner avec le bouton gauche.

Vous pouvez désélectionner toutes les lignes sélectionnées en cliquant sur l'arrière-plan du graphique, ou en choisissant **Effacer la sélection** dans le menu Relations de la fenêtre du graphique.

Affichage de la relation à l'aide d'une autre présentation

Dans le menu Relations, choisissez **Présentation en cercle**, **Présentation en réseau**, **Présentation orientée** ou **Présentation de grille** pour modifier la présentation du graphique.

Activation ou désactivation du curseur des liens

Dans le menu Affichage, choisissez **Curseur des liens**.

Sélection ou marquage d'enregistrements pour une unique relation

1. Cliquez avec le bouton droit de la souris sur la ligne représentant la relation voulue.
2. Dans le menu contextuel, choisissez **Générer le noeud Sélectionner pour le lien** ou **Générer le noeud Calculer pour le lien**.

Un noeud Sélectionner ou Calculer est automatiquement ajouté dans l'espace de travail du flux avec les options et les conditions appropriées définies :

- Le noeud Sélectionner sélectionne tous les enregistrements de la relation donnée.
- Le noeud Calculer génère un booléen qui indique si la relation sélectionnée est valide pour tous les enregistrements du jeu de données. Le nom du champ indicateur correspond à l'association (à l'aide d'un trait de soulignement) des deux valeurs constituant la relation, comme `FAIBLE_MédC` ou `MédC_FAIBLE`.

Sélection ou marquage d'enregistrements pour un groupe de relations

1. Sélectionnez dans le graphique Relations les lignes représentant les relations voulues.
2. Dans le menu Générer de la fenêtre du graphique, sélectionnez **Noeud Sélectionner (Et)**, **Noeud Sélectionner (Ou)**, **Noeud Calculer (Et)** ou **Noeud Calculer (Ou)**.

- Les noeuds "ou" donnent la disjonction des conditions. Autrement dit, le noeud est appliqué aux enregistrements pour lesquels l'une des relations sélectionnées est valide.
- Les noeuds "et" donnent la conjonction des conditions. Autrement dit, le noeud est appliqué uniquement aux enregistrements pour lesquels toutes les relations sélectionnées sont valides. Une erreur se produit si certaines des relations sélectionnées s'excluent mutuellement.

Une fois votre sélection effectuée, un noeud Sélectionner ou Calculer est automatiquement ajouté dans l'espace de travail de flux avec les options et les conditions appropriées définies.

Remarque : Lors du filtrage des liens affichés (à l'aide du curseur du graphique Relations ou du contrôle **Afficher uniquement les liens au-dessus de** de l'onglet Options du noeud Relations), vous pouvez vous retrouver dans une situation où tous les liens qui restent à afficher possèdent la même valeur (en d'autres termes, ce sont tous des liens faibles, des liens normaux ou des liens forts, tels que définis par les contrôles **Liens faibles en dessous de** et **Liens forts au-dessus de** dans l'onglet Options du noeud Relations). Si cela se produit, tous les liens sont affichés avec la ligne d'épaisseur moyenne dans la sortie graphique Relations.

Ajustement des seuils des graphiques Relations

Une fois le graphique Relations créé, vous pouvez ajuster les seuils qui gèrent les styles des lignes à l'aide du curseur de la barre d'outils pour modifier la ligne visible minimale. Vous pouvez également afficher d'autres options de seuil en cliquant sur le bouton de la barre d'outils représentant une double-flèche jaune, afin d'agrandir la fenêtre du graphique Relations. Cliquez ensuite sur l'onglet **Commandes** pour

afficher les options supplémentaires.

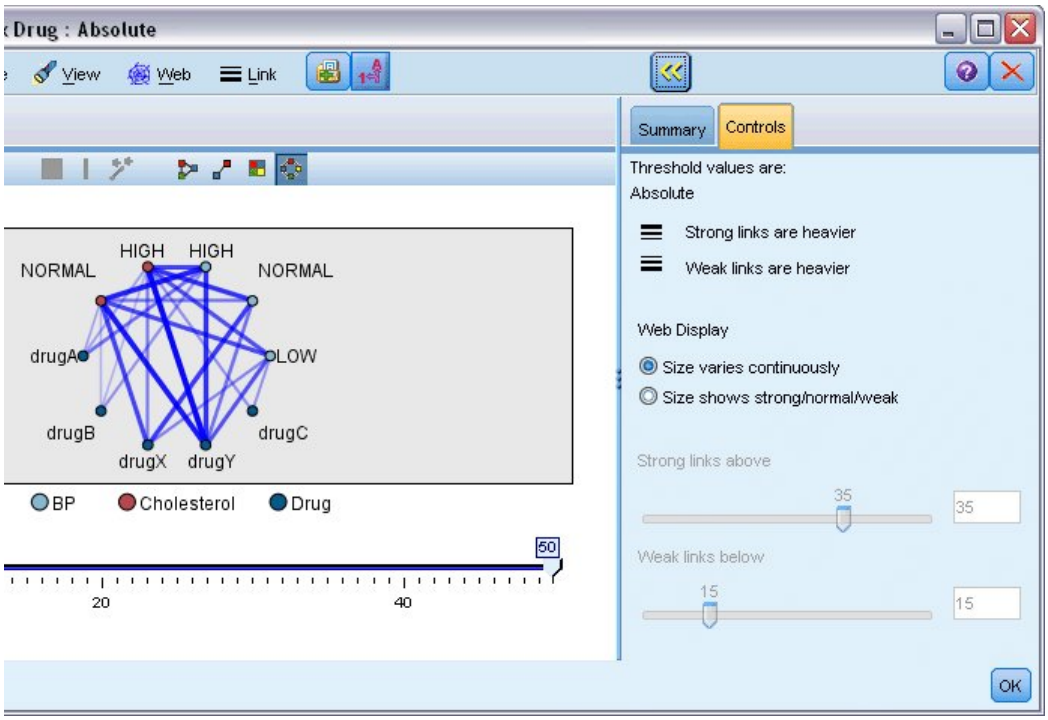


Figure 40. Fenêtre agrandie comportant les options d'affichage et de seuil

Les valeurs de seuil sont. Affiche le type de seuil sélectionné lors de la création dans la boîte de dialogue du noeud Relations.

Les liens forts sont plus volumineux. Sélectionnée par défaut, cette option correspond à la méthode standard d'affichage des liens entre les champs.

Les liens faibles sont plus volumineux. Sélectionnez cette option pour inverser la signification des liens représentés par des lignes en gras. Cette option est fréquemment utilisée pour détecter des fraudes ou examiner des valeurs éloignées.

Affichage des relations. Choisissez les options permettant de déterminer la taille des liens du graphique de sortie :

- **La taille varie en permanence.** Sélectionnez cette option pour afficher une amplitude de tailles de lien reflétant la variation des forces de connexion en fonction des valeurs de données réelles.
- **La valeur de la taille est fort/normal/faible.** Sélectionnez cette option pour afficher trois forces de connexion : forte, normale et faible. Les points de césure de ces catégories peuvent être définis grâce aux options ci-avant, ainsi que dans le graphique final.

Liens forts au-dessus de. Choisissez un chiffre indiquant le seuil entre les connexions fortes (lignes en gras) et les connexions standard (lignes normales). Toutes les connexions au-dessus de cette valeur sont considérées comme fortes. Utilisez le curseur pour rectifier la valeur ou saisissez un chiffre dans le champ.

Liens faibles en dessous de. Choisissez un chiffre indiquant le seuil entre les connexions faibles (lignes en pointillé) et les connexions standard (lignes normales). Toutes les connexions au-dessous de cette valeur sont considérées comme faibles. Utilisez le curseur pour rectifier la valeur ou saisissez un chiffre dans le champ.

Après avoir ajusté les seuils du graphique Relations, vous pouvez réorganiser ou redessiner l'affichage des relations en utilisant les nouvelles valeurs des seuils à travers le menu Relations situé sur la barre d'outils du graphique Relations. Une fois que vous avez trouvé les paramètres révélant les motifs les plus significatifs, vous pouvez mettre à jour les paramètres d'origine du noeud Relations (également appelé noeud Relations parent) en sélectionnant **Mettre à jour le noeud parent** dans le menu Relations de la fenêtre du graphique.

Création d'un récapitulatif Web

Vous pouvez créer un récapitulatif des relations répertoriant les liens forts, moyens et faibles en cliquant sur le bouton de la barre d'outils représentant une double-flèche jaune, afin d'agrandir la fenêtre du graphique Relations. Cliquez ensuite sur l'onglet **Récapitulatif** pour afficher les tableaux de chaque type de lien. Vous pouvez agrandir ou réduire les tableaux en utilisant le bouton bascule correspondant.

Pour imprimer le récapitulatif, choisissez l'option suivante dans le menu de la fenêtre du graphique Relations :

Fichier > **Imprimer le récapitulatif**

Noeud Evaluation

Le noeud Evaluation permet d'évaluer et de comparer facilement des modèles prédictifs afin de choisir celui le mieux adapté à l'application. Les graphiques Evaluation montrent l'aptitude des modèles à prédire des résultats spécifiques. Ils trient les enregistrements en fonction de la valeur prédite et de la confiance dans cette prévision, divisent les enregistrements en groupes de taille égale (**quantiles**), puis reportent la valeur du critère traité pour chaque quantile, du plus élevé au plus faible. Les divers modèles apparaissent sous forme de lignes dans le graphique.

Les résultats sont traités grâce à la définition d'une valeur ou d'une amplitude de valeurs spécifique en tant qu'**occurrence**. Les correspondances indiquent généralement une réussite (telle qu'une vente conclue avec un client) ou un événement intéressant (tel qu'un diagnostic médical spécifique). Vous pouvez définir des critères d'occurrence dans l'onglet Options de la boîte de dialogue. Vous pouvez également utiliser les critères d'occurrence par défaut suivants :

- Les champs de sortie **indicateurs** sont simples ; les correspondances renvoient à des valeurs *vraies*.
- En ce qui concerne les champs de sortie **nominaux**, c'est la première valeur de l'ensemble qui définit une correspondance.
- Pour les champs de sortie **continus**, les correspondances sont les valeurs supérieures à la moitié de l'intervalle du champ.

Il existe six types de graphique Evaluation, chacun mettant en valeur un critère d'évaluation différent :

Graphiques de gains

Les gains sont définis comme la proportion du nombre total de occurrences représentée dans chaque quantile. Les gains sont calculés de la façon suivante : (nombre de correspondances dans le quantile / nombre total de correspondances) x 100 %.

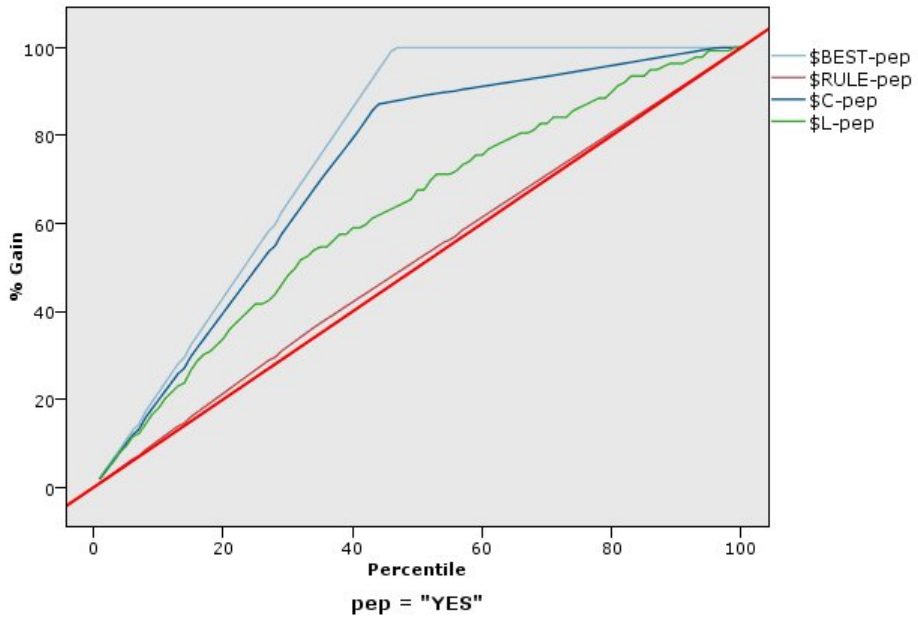


Figure 41. Graphique de gains (cumulatif) avec affichage de la ligne de référence, de la meilleure ligne et de la règle de marché

Graphiques de lift

Ces graphiques comparent le pourcentage d'enregistrements dans chaque quantile qui se sont traduits par des correspondances et le pourcentage total de correspondances dans les données d'apprentissage. Le calcul s'effectue de la façon suivante : (occurrences dans le quantile / enregistrements dans le quantile) / (nombre total d'occurrences / nombre total d'enregistrements).

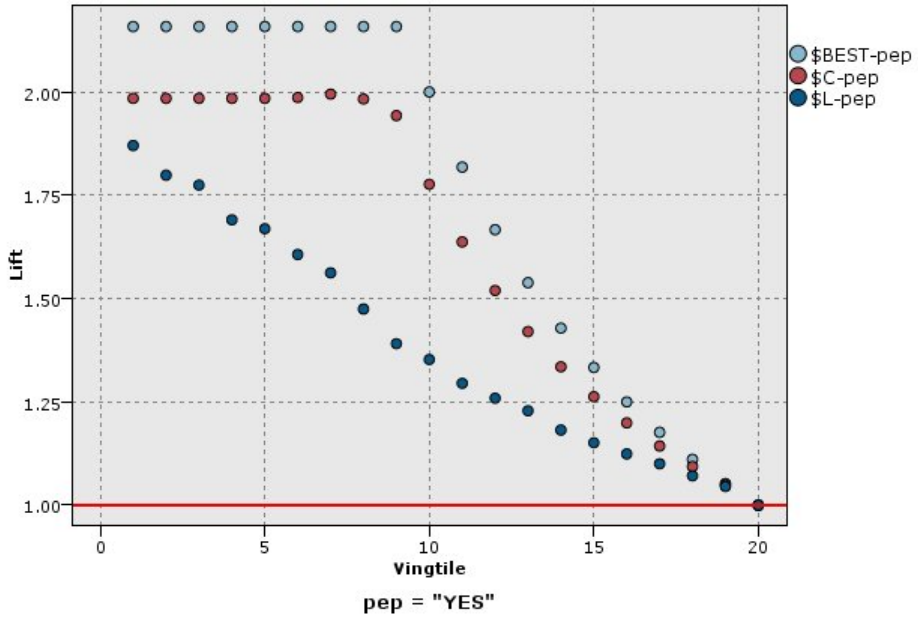


Figure 42. Graphique de lift (cumulatif) utilisant des points et la meilleure ligne

Graphiques de réponses

La réponse correspond tout simplement au pourcentage d'enregistrements dans le quantile qui sont des occurrences. La réponse se calcule de la façon suivante : (nombre de correspondances dans le quantile / enregistrements dans le quantile) x 100 %.

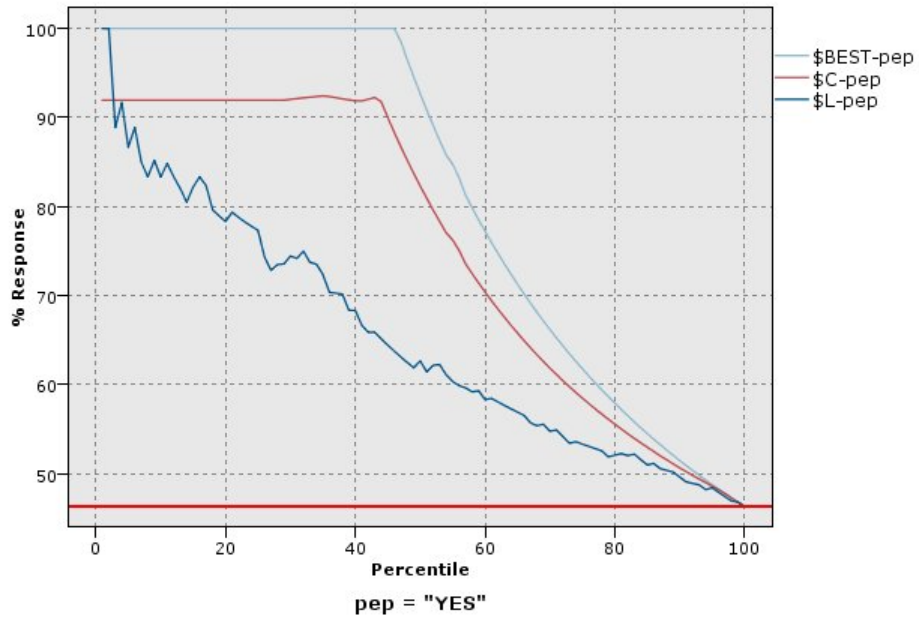


Figure 43. Graphique de réponses (cumulatif) affichant la meilleure ligne

Graphiques de profits

Le profit est égal au **revenu** de chaque enregistrement moins le **coût** de l'enregistrement. Les profits d'un quantile correspondent à la somme des profits de tous ses enregistrements. Les revenus sont supposés ne s'appliquer qu'aux occurrences, mais les coûts s'appliquent à tous les enregistrements. Les profits et les coûts peuvent être fixes ou peuvent être déterminés par les champs des données. Les profits sont calculés de la façon suivante : (somme des revenus des enregistrements du quantile somme des coûts des enregistrements du quantile).

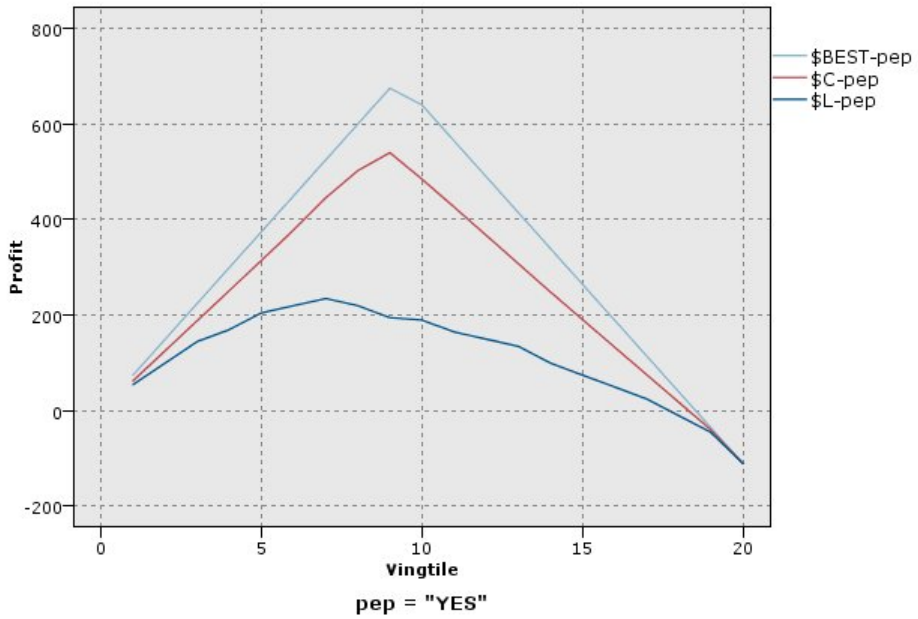


Figure 44. Graphique de profits (cumulatif) affichant la meilleure ligne

Graphiques de retour sur investissement

Le retour sur investissement est semblable au profit dans le sens où il s'agit de définir des revenus et des coûts. Le retour sur investissement compare les profits du quantile à ses coûts. Le retour sur investissement se calcule de la façon suivante : (profits du quantile / coûts du quantile) x 100 %.

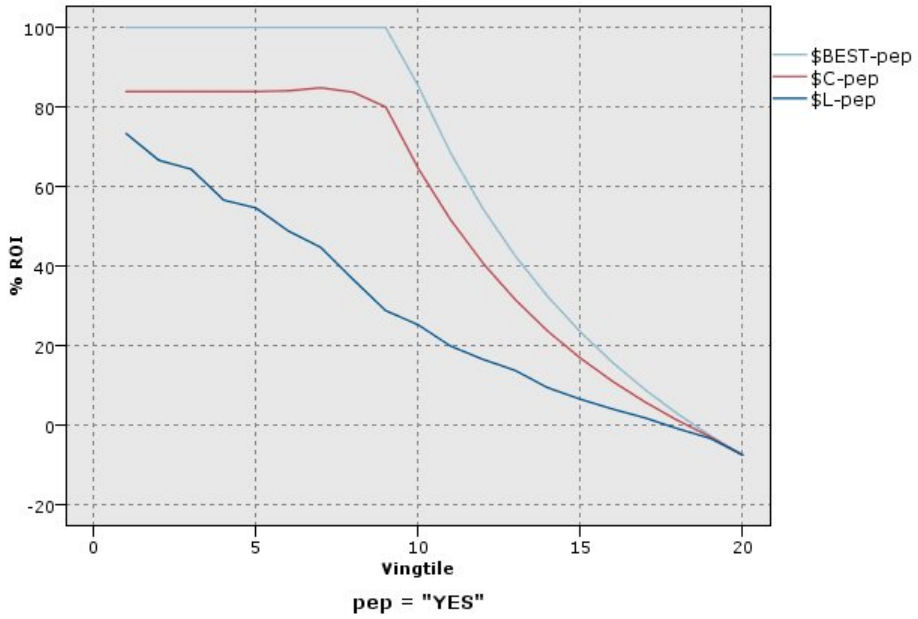


Figure 45. Graphique de retour sur investissement (cumulatif) affichant la meilleure ligne

Graphiques ROC

ROC (Receiver Operator Characteristic) ne peut être utilisé qu'avec des discriminants binaires. ROC peut être utilisé pour visualiser, organiser et sélectionner des discriminants en fonction de leurs performances. Un graphique ROC représente le taux de vrai positif (ou sensibilité) par rapport au taux de faux positif du discriminant. Il décrit les compromis relatifs entre les bénéfices (vrais positifs) et les coûts (faux positifs). Un vrai positif est une instance qui est une occurrence classée en tant que telle. Le taux de vrai positif est donc calculé sous la forme nombre de vrais positifs / nombre d'instances qui sont réellement des occurrences. Un faux positif est une instance qui est une occurrence manquée classée en tant qu'occurrence. Le taux de faux positif est donc calculé sous la forme nombre de faux positifs / nombre d'instances qui sont en fait des occurrences manquées.

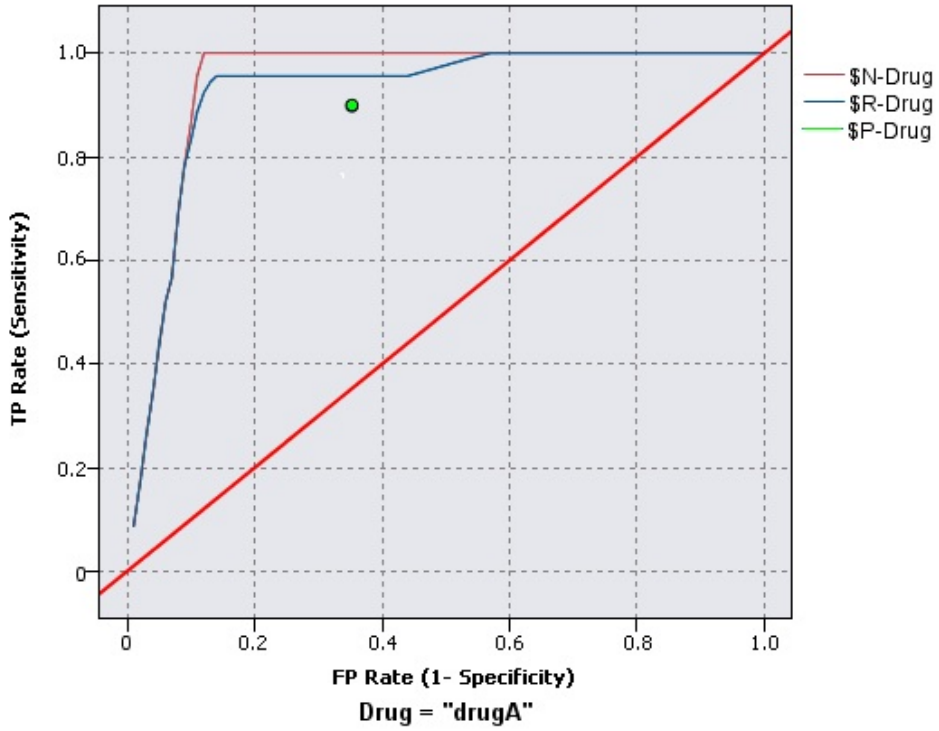


Figure 46. Graphique ROC affichant la meilleure ligne

Les graphiques Evaluation peuvent également être cumulatifs. Ainsi, chaque point est égal à la valeur du quantile correspondant, plus celle de tous les quantiles supérieurs. Les graphiques cumulatifs soulignent mieux la performance globale des modèles, alors que les graphiques non cumulatifs permettent de mettre en valeur les zones problématiques des modèles.

Onglet Tracé d'évaluation

Type de graphique. Sélectionnez l'un des types suivants : **Gains**, **Réponse**, **Lift**, **Profit**, **ROI** (retour sur investissement) ou **ROC** (Receiver Operator Characteristic).

Tracé cumulatif. Cochez cette case pour créer un graphique cumulatif. Dans les graphiques cumulatifs, les valeurs reportées correspondent à celles de chaque quantile, plus celles de tous les quantiles supérieurs. (L'option **Tracé cumulatif** n'est pas disponible pour les graphiques ROC.)

Inclure la ligne de référence. Sélectionnez cette option pour inclure une ligne de référence dans le graphique, qui indique une proportion de correspondances parfaitement aléatoire où la confiance devient inutile. (L'option **Inclure la ligne de référence** n'est pas disponible pour les graphiques de profits et de retour sur investissement.)

Inclure la meilleure ligne. Sélectionnez cette option pour inclure une meilleure ligne dans le graphique, qui indique une confiance parfaite (où les correspondances équivalent à 100 % des observations). (L'option **Inclure la meilleure ligne** n'est pas disponible pour les graphiques ROC.)

Utiliser les critères de profit pour tous les types de graphique. Sélectionnez cette option pour utiliser les critères de profit (coût, revenu et pondération) lors du calcul des mesures d'évaluation, au lieu du nombre normal d'occurrences. Concernant les modèles ayant certaines cibles numériques (par exemple, le modèle qui permet de prévoir les recettes obtenues auprès d'un client en réponse à une offre), la valeur du champ cible donne une meilleure évaluation des performances du modèle que le nombre de correspondances. Cette option permet d'activer les champs **Coûts**, **Revenu** et **Pondération** pour les graphiques Gains, Réponse et Lift. Pour utiliser les critères de profit pour ces trois types de graphique, nous vous recommandons de définir **Revenu** comme champs cible, **Coût** sur 0.0 afin que le profit soit égal au revenu, et de spécifier une condition de correspondance de l'utilisateur de valeur "true" afin que tous les enregistrements soient comptés comme des correspondances. (L'option **Utiliser les critères de profit pour tous les types de graphique** n'est pas disponible pour les graphiques ROC.)

Rechercher les champs prédits/de prédicteur avec. Sélectionnez soit **Modéliser les métadonnées de champ de sortie** pour rechercher les champs prédits dans le graphique en utilisant leurs métadonnées, soit **Format de nom de champ** pour les rechercher par nom.

Champs de score du tracé. Cochez cette case pour activer le sélecteur de champs de score. Puis sélectionnez un ou plusieurs champs de score à intervalles ou continus. Ces champs ne sont pas strictement des modèles prédictifs mais peuvent être utiles pour classer des enregistrements selon leur propension à être une occurrence. Le noeud Evaluation peut comparer toute combinaison d'un ou de plusieurs champs de score avec un ou plusieurs modèles prédictifs. Un exemple typique peut être de comparer plusieurs champs RFM avec votre meilleur modèle prédictif.

Cible. Sélectionnez le champ cible à l'aide du sélecteur de champ. Choisissez un champ nominal ou indicateur instancié comportant deux valeurs ou plus.

Remarque : Ce champ cible s'applique seulement aux champs de score (les modèles prédictifs déterminent leurs propres cibles) et est ignoré si un critère de correspondance personnalisé est défini sur l'onglet Options.

Diviser par partition. Si un champ de partition permet de diviser des enregistrements en échantillons d'apprentissage, de test et de validation, sélectionnez cette option pour afficher un graphique Evaluation distinct pour chaque partition. Pour plus d'informations, voir «Noeud Partitionner», à la page 185.

Remarque : lorsque vous divisez des enregistrements par partition, ceux dont le champ de partition contient des valeurs nulles sont exclus de l'évaluation. Ce problème ne se pose jamais si un noeud Partitionner est utilisé, car ce type de noeud ne génère aucune valeur nulle.

Tracé. Dans la liste déroulante, sélectionnez la taille des quantiles à représenter sur le graphique. Les options disponibles sont **Quartiles**, **Quintiles**, **Déciles**, **Vingtiles**, **Centiles** et **1000-tiles**. (L'option **Tracé** n'est pas disponible pour les graphiques ROC.)

Style. Sélectionnez **Ligne** ou **Point**.

Pour tous les types de graphique, à l'exception des graphiques ROC, vous pouvez en outre préciser les coûts, le revenu et la pondération.

- **Coûts.** Précisez le coût associé à chaque enregistrement. Vous pouvez sélectionner **Fixe** ou **Variable**. Pour les coûts fixes, précisez la valeur du coût. Pour les coûts variables, cliquez sur le sélecteur de champs pour sélectionner un champ comme champ de coût. (L'option **Coûts** n'est pas disponible pour les graphiques ROC.)

- **Recette.** Précisez le revenu associé à chaque enregistrement représentant une occurrence. Vous pouvez sélectionner **Fixe** ou **Variable**. Pour les revenus fixes, précisez la valeur du revenu. Pour les revenus variables, cliquez sur le sélecteur de champs pour sélectionner un champ comme champ de revenu. (L'option **Recette** n'est pas disponible pour les graphiques ROC.)
- **Pondération.** Si les enregistrements de vos données représentent plusieurs unités, vous pouvez utiliser les pondérations de fréquence pour ajuster les résultats. Indiquez la pondération associée à chaque enregistrement, à l'aide des options **Fixe** ou **Variable**. Pour une pondération fixe, précisez la valeur de la pondération (nombre d'unités par enregistrement). Pour une pondération variable, cliquez sur le sélecteur de champs pour sélectionner un champ comme champ de pondération. (L'option **Pondération** n'est pas disponible pour les graphiques ROC.)

Onglet Options d'évaluation

L'onglet Options des graphiques Evaluation permet de définir facilement les occurrences, les règles de marché et les critères d'évaluation affichés dans les graphiques. Vous pouvez également définir des options pour exporter les résultats de l'évaluation du modèle.

Correspondance définie par l'utilisateur. Sélectionnez cette option pour spécifier la condition personnalisée utilisée pour indiquer une occurrence. Cette option permet de définir les résultats qui vous intéressent au lieu de les déduire du type de champ cible et de l'ordre des valeurs.

- **Condition.** Lorsque l'option **Correspondance définie par l'utilisateur** est sélectionnée, vous devez indiquer l'expression CLEM de la condition de correspondance. Par exemple, `@TARGET = "YES"` est une condition valide qui indique que la valeur *Oui* du champ cible sera considérée comme une occurrence lors de l'évaluation. La condition indiquée sera utilisée pour tous les champs cible. Pour créer une condition, entrez une valeur dans le champ ou utilisez le Générateur de formules pour générer une expression de condition. Si les données sont instanciées, vous pouvez insérer des valeurs directement à partir du Générateur de formules.

Score défini par l'utilisateur. Sélectionnez cette option pour indiquer une condition servant à évaluer les observations avant de les affecter à des quantiles. Le score par défaut est calculé à partir de la valeur prédite et de la fiabilité. Utilisez le champ Expression pour créer une expression d'évaluation personnalisée.

- **Expression.** Indiquez l'expression CLEM utilisée pour l'évaluation. Par exemple, si une sortie numérique de l'intervalle 0–1 est triée afin que les valeurs inférieures soient meilleures que les valeurs supérieures, vous pouvez définir une correspondance supérieure `@TARGET < 0,5`, ainsi que le score associé `(1  @PREDICTED)`. L'expression du score doit correspondre à une valeur numérique. Pour créer une condition, entrez une valeur dans le champ ou utilisez le Générateur de formules pour générer une expression de condition.

Inclure une règle métier. Sélectionnez cette option pour indiquer une condition de règle reflétant les critères intéressants. Par exemple, vous pouvez afficher une règle pour toutes les observations où `mortgage = "Y" and income >= 33 000`. Les règles de marché apparaissent sur le graphique et sont appelées *Règle* dans la clé. (L'option **Inclure une règle métier** n'est pas prise en charge pour les graphiques ROC.)

- **Condition.** Indiquez l'expression CLEM utilisée pour définir une règle métier dans le graphique de sortie. Entrez une valeur dans le champ ou utilisez le Générateur de formules pour générer une expression de condition. Si les données sont instanciées, vous pouvez insérer des valeurs directement à partir du Générateur de formules.

Exporter les résultats dans un fichier. Sélectionnez cette option pour exporter les résultats de l'évaluation du modèle dans un fichier texte délimité. Vous pouvez lire ce fichier pour réaliser des analyses spécifiques des valeurs calculées. Pour l'exportation, définissez les options suivantes :

- **Nom de fichier.** Entrez le nom du fichier de sortie. Utilisez le bouton **...** pour accéder au dossier voulu.
- **Délimiteur**. Entrez le caractère, tel qu'une virgule ou un espace, à utiliser comme séparateur de champ.

Inclure les noms des champs. Sélectionnez cette option pour inclure les noms de champ sur la première ligne du fichier de sortie.

Nouvelle ligne après chaque enregistrement. Sélectionnez cette option pour commencer chaque enregistrement sur une nouvelle ligne.

Onglet Apparence de l'évaluation

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous-titre du graphique.

Texte. Vous pouvez soit accepter le libellé de texte généré automatiquement, soit sélectionner **Personnalisé** pour indiquer un libellé.

Libellé X. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *x* (horizontal), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Libellé Y. Vous pouvez soit accepter le libellé généré automatiquement pour l'axe *y* (vertical), soit sélectionner **Personnalisé** pour indiquer un libellé personnalisé.

Afficher le quadrillage. Sélectionnée par défaut, cette option affiche un quadrillage derrière le nuage ou le graphique, vous permettant de déterminer plus facilement les points de césure des zones et des bandes. Les quadrillages sont toujours de couleur blanche, sauf si l'arrière-plan du graphique est blanc ; dans ce cas, ils sont de couleur grise.

Lecture des résultats d'une évaluation de modèle

L'interprétation d'un graphique Evaluation dépend dans une certaine mesure du type du graphique, mais il existe cependant des caractéristiques communes à tous les graphiques Evaluation. Sur les graphiques cumulatifs, les lignes les plus hautes indiquent les modèles mieux adaptés, tout particulièrement sur la gauche du graphique. Souvent, lors de la comparaison de plusieurs modèles, les lignes se croisent, indiquant qu'un modèle est meilleur sur une partie du graphique et un autre modèle sur une autre partie. Dans ce cas, vous devez prendre en considération la portion de l'échantillon qui vous intéresse (ce qui revient à définir un point sur l'axe *x*) lors du choix du modèle.

La plupart des graphiques non cumulatifs sont très similaires. Dans les modèles satisfaisants, les graphiques non cumulatifs sont hauts sur la gauche et bas sur la droite du graphique. (Si un graphique non cumulatif affiche un motif en dents de scie, vous pouvez le rendre plus régulier en réduisant le nombre de quantiles à reporter et en réexécutant le graphique.) La présence de lignes basses sur la gauche du graphique ou de lignes hautes sur la droite indiquent parfois des zones où les prévisions du modèle sont médiocres. Une ligne droite sur l'ensemble du graphique indique que le modèle ne fournit aucune information.

Graphiques de gains. Les graphiques de gains cumulatifs commencent toujours à 0 % sur la gauche et finissent toujours à 100 % sur la droite. Les graphique de gain des bons modèles présentent une hausse rapide en direction de la valeur 100 %, puis se stabilisent. Un modèle ne fournissant aucune information suit une trajectoire en diagonale du coin inférieur gauche au coin supérieur droit (affiché sur le graphique si l'option **Inclure la ligne de référence** est sélectionnée).

Graphiques Lift. Les graphiques Lift cumulatifs commencent au-dessus de 1 à gauche, puis baissent progressivement jusqu'à atteindre 1 à droite. Le bord droit du graphique représente l'intégralité du jeu de données, donc le rapport entre les occurrences dans les quantiles cumulatifs et les occurrences dans les données est égal à 1,0. Dans les modèles satisfaisants, le graphique de lift commence bien au-dessus de 1,0 sur la gauche, reste à un niveau élevé à mesure que vous avancez vers la droite, puis baisse

rapidement vers 1,0 sur la droite du graphique. Si le modèle ne fournit aucune information, la ligne reste autour de 1 sur la totalité du graphique. (Si l'option **Inclure la ligne de référence** est sélectionnée, une ligne de référence horizontale correspondant à la valeur 1 figure sur le graphique.)

Graphiques de réponses. Les graphiques de réponses cumulatifs sont semblables aux graphiques de lift, à l'exception de la mise à l'échelle. Les graphiques de réponses commencent autour de 100 %, puis baissent progressivement jusqu'à atteindre le taux de réponse global (nombre total de correspondances / nombre total d'enregistrements), à droite. Dans les modèles satisfaisants, la ligne commence autour de 100 % ou à 100 % (sur la gauche), reste à un niveau élevé à mesure que vous avancez vers la droite, puis baisse rapidement vers le taux de réponse global sur la droite du graphique. Si le modèle ne fournit aucune information, la ligne reste autour du taux de réponse global sur la totalité du graphique. (Si l'option **Inclure la ligne de référence** est sélectionnée, une ligne de référence horizontale correspondant au taux de réponse global figure sur le graphique.)

Graphiques de profits. Les graphiques de profits cumulatifs montrent la somme des profits à mesure que vous augmentez la taille de l'échantillon sélectionné (de gauche à droite). Les graphiques de profits commencent généralement autour de 0, augmentent régulièrement à mesure que vous avancez vers la droite jusqu'à atteindre un pic ou un plateau au centre du graphique, puis baissent vers le bord droit du graphique. Dans les modèles satisfaisants, les profits affichent un pic bien défini au centre du graphique. Si le modèle ne fournit aucune information, la ligne est relativement droite et peut augmenter, diminuer ou se stabiliser en fonction de la structure coût/revenu utilisée.

Graphiques de retour sur investissement. Les graphiques de retour sur investissement cumulatifs sont semblables aux graphiques de réponses et aux graphiques Lift, à l'exception de la mise à l'échelle. Les graphiques de retour sur investissement commencent généralement au-dessus de 0 %, puis baissent progressivement jusqu'à atteindre le retour sur investissement global de l'intégralité du jeu de données (qui peut être un nombre négatif). Dans les modèles satisfaisants, la ligne commence bien au-dessus de 0 %, reste à un niveau élevé à mesure que vous avancez vers la droite, puis baisse assez rapidement vers le retour sur investissement global sur la droite du graphique. Si le modèle ne fournit aucune information, la ligne reste autour de la valeur du retour sur investissement global.

Graphiques ROC. Les courbes ROC prennent généralement la forme d'un graphique des gains cumulés. La courbe démarre à la coordonnée (0,0) et se termine à la coordonnée (1,1) en partant de gauche à droite. Un graphique qui présente une hausse rapide vers la coordonnée (0,1) puis se stabilise indique un discriminant de bonne qualité. Un modèle qui classe les instances de façon aléatoire en tant qu'occurrences ou occurrences manquées suit une trajectoire en diagonale du coin inférieur gauche au coin supérieur droit (affiché sur le graphique si l'option **Inclure la ligne de référence** est sélectionnée). Si aucun champ de confiance n'est fourni pour un modèle, ce dernier est représenté sous forme de point unique. Le discriminant doté du meilleur seuil de classification est celui qui se trouve le plus près de la coordonnée (0,1), ou dans l'angle supérieur gauche, du graphique. Cet emplacement représente un nombre élevé d'instances correctement classées en tant qu'occurrences et un nombre faible d'instances classées incorrectement en tant qu'occurrences. Les points situés au dessus de la diagonale représentent les bons résultats de classification. Les points situés en dessous de la diagonale représentent les mauvais résultats de classification, pires que si les instances avaient été classées de façon aléatoire.

Utilisation d'un graphique Evaluation

Comme dans les graphiques Histogramme et Résumé, vous pouvez utiliser la souris pour explorer les graphiques Evaluation. L'axe *x* représente les scores des modèles dans les quantiles indiqués (vingtiles ou déciles).

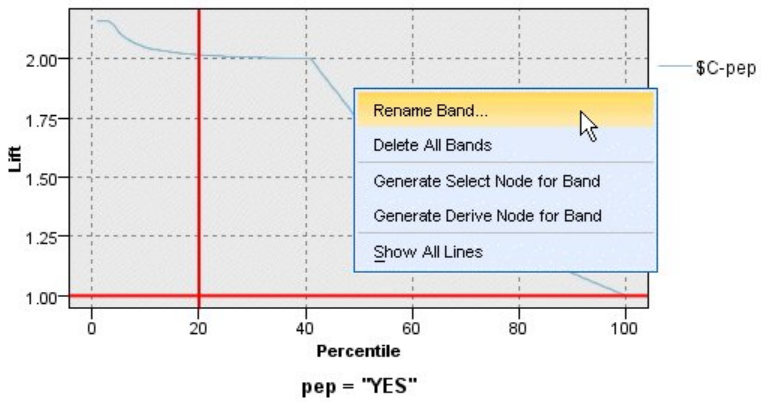


Figure 47. Utilisation d'un graphique Evaluation

Vous pouvez partitionner l'axe *x* en bandes (comme pour un graphique Histogramme) en utilisant l'icône de fractionnement pour afficher les options permettant de fractionner automatiquement l'axe en bandes égales. Pour plus d'informations, voir «Exploration de graphiques», à la page 290. Vous pouvez éditer manuellement les limites des bandes en sélectionnant **Bandes graphiques** dans le menu Edition.

Après avoir créé un graphique Evaluation, défini des bandes et examiné les résultats, vous pouvez utiliser les options du menu Générer et du menu contextuel pour créer automatiquement des noeuds basés sur les sélections du graphique. Pour plus d'informations, voir «Génération de noeuds à partir de graphiques», à la page 298.

Lorsque vous générez des noeuds à partir d'un graphique Evaluation, vous êtes invité à sélectionner un modèle parmi tous ceux disponibles dans le graphique.

Sélectionnez un modèle et cliquez sur **OK** pour générer le nouveau noeud dans l'espace de travail de flux.

Noeud Visualisation de carte

Le noeud Visualisation de carte peut accepter plusieurs connexions d'entrée et afficher des données géospatiales sur une carte sous forme de série de couches. Chaque couche est un champ géospatial unique. Par exemple, la couche de base peut être la carte d'un pays ; sur cette couche, il peut y avoir une couche pour les routes, une couche pour les rivières et une couche pour les villes.

Bien que la plupart des ensembles de données géospatiales contiennent un champ géospatial unique, lorsqu'il existe plusieurs champs géospatiaux dans une entrée unique, vous pouvez choisir les champs à afficher. Deux champs provenant de la même connexion d'entrée ne peuvent pas être affichés simultanément. Toutefois, vous pouvez copier et coller la connexion entrante et afficher un champ différent depuis chaque connexion.

Onglet Tracé de visualisation de carte

Couches

Cette table affiche des informations sur les entrées du noeud de carte. L'ordre des couches dicte l'ordre dans lequel les couches sont affichées dans l'aperçu de carte et dans la sortie graphique lorsque le noeud est exécuté. La première ligne de la table est la couche 'supérieure' et la dernière ligne est la couche 'de base' ; en d'autres termes, chaque couche est affichée sur la carte avant la couche qui se trouve directement sous elle dans la table.

Remarque : Lorsqu'une couche dans la table contient un champ géospatial tridimensionnel, l'axe des X et l'axe des Y seulement sont représentés. L'axe des Z est ignoré.

Nom Les noms sont créés automatiquement pour chaque couche au format suivant : `balise[noeud source:noeud connecté]`. Par défaut, la balise est un nombre, 1 représentant la première entrée connectée, 2 la deuxième entrée, etc. Si nécessaire, cliquez sur le bouton **Editer la couche** pour changer la balise dans la boîte de dialogue Change Map Layer Options. Par exemple, vous pouvez remplacer la balise par "routes" ou "villes" pour refléter l'entrée de données.

Type Affiche l'icône de type de mesure du champ géospatial qui est sélectionné comme couche. Si les données d'entrée contiennent plusieurs champs dont le type de mesure est géospatial, la sélection par défaut utilise l'ordre de tri suivant :

1. Point
2. Chaîne
3. Polygone
4. Multipoint
5. Multichaîne
6. Multipolygone

Remarque : Si deux champs possèdent le même type de mesure, le premier champ (par ordre alphabétique) est sélectionné par défaut.

Symbole

Remarque : Cette colonne est remplie pour les champs Point et Multipoint seulement.
Affiche le symbole qui est utilisé pour les champs Point ou Multipoint. Si nécessaire, cliquez sur le bouton **Editer la couche** pour changer le symbole dans la boîte de dialogue Change Map Layer Options.

Couleur

Affiche la couleur qui est sélectionnée pour représenter la couche sur la carte. Si nécessaire, cliquez sur le bouton **Editer la couche** pour changer la couleur dans la boîte de dialogue Change Map Layer Options. La couleur est appliquée aux différents éléments selon le type de mesure.

- Pour les points ou les multipoints, la couleur est appliquée au symbole pour la couche.
- Pour les chaînes et les polygones, la couleur est appliquée à la forme entière. Les polygones ont toujours un contour noir ; la couleur qui est affichée dans la colonne est celle qui remplit la forme.

Aperçu

Ce panneau affiche un aperçu de la sélection en cours des entrées dans la table **Couches**. L'aperçu prend en compte l'ordre des couches, le symbole, la couleur et d'autres paramètres d'affichage associés aux couches et, si possible, met à jour l'affichage à chaque fois que les paramètres changent. Si vous changez des détails ailleurs dans votre flux, par exemple les champs géospatiaux à utiliser comme couches, ou si vous modifiez des détails tels que les fonctions d'agrégation associées, il peut être nécessaire de cliquer sur le bouton **Actualiser les données** pour mettre à jour l'aperçu.

Utilisez l'**aperçu** pour définir vos paramètres d'affichage avant d'exécuter votre flux. Pour éviter les retards pouvant être causés par l'utilisation d'un jeu de données volumineux, l'aperçu échantillonne chaque couche et crée un affichage à partir des 100 premiers enregistrements.

Modification des couches de carte

Vous pouvez utiliser la boîte de dialogue Change Map Layer Options pour modifier divers détails pour chaque couche affichée dans l'onglet **Tracé** du noeud Visualisation de carte.

Détails d'entrée

Balise Par défaut, la balise est un nombre que vous pouvez remplacer par une chaîne plus significative pour identifier la couche sur la carte. Par exemple, il peut s'agir du nom de l'entrée de données, comme "Villes".

Zone Couche

S'il existe plusieurs champs géospatiaux dans vos données d'entrée, utilisez cette option pour sélectionner le champ à afficher comme couche sur la carte.

Par défaut, les couches que vous pouvez choisir sont triées dans l'ordre suivant :

- Point
- Chaîne
- Polygone
- Multipoint
- Multichaîne
- Multipolygone

Paramètres d'affichage

Utiliser le compartimentage hexagonal

Remarque : Cette option affecte les champs Point et Multipoint seulement.

Le compartimentage hexagonal combine des points proches (en fonction de leurs coordonnées x et y) en un point unique à afficher sur la carte. Le point unique est affiché sous forme d'hexagone mais est en réalité rendu sous forme de polygone.

Etant donné que l'hexagone est rendu sous forme de polygone, les champs de points pour lesquels le compartimentage hexagonal est activé sont traités comme des polygones. Cela signifie que si vous choisissez **Classer par type** dans la boîte de dialogue de noeud de carte, les couches des points auxquelles le compartimentage hexagonal est appliqué sont rendus au-dessus des couches des polygones mais sous les couches des chaînes et des points.

Si vous utilisez le compartimentage hexagonal pour un champ multipoint, le champ est d'abord converti en champ point en compartimentant les valeurs multipoints afin de calculer le point central. Les points centraux sont utilisés pour calculer les compartiments hexagonaux.

Agrégation

Remarque : Cette colonne est disponible uniquement lorsque vous sélectionnez la case à cocher **Utiliser le compartimentage hexagonal** ainsi qu'un champ **Superposition**.

Si vous sélectionnez un champ **Superposition** pour une couche des points qui utilise le compartimentage hexagonal, toutes les valeurs de ce champ doivent être agrégées pour tous les points dans l'hexagone. Spécifiez une fonction d'agrégation pour les champs de superposition à appliquer à la carte. Les fonctions d'agrégation disponibles dépendent du type de mesure.

- Fonctions d'agrégation pour un type de mesure Continu, avec stockage Réel ou Entier :
  - Somme
  - Moyenne
  - Min
  - Max
  - Médiane
  - Premier quartile
  - Troisième quartile

- Fonctions d'agrégation pour un type de mesure Continu, avec stockage Heure, Date ou Horodatage :
  - Moyenne
  - Min
  - Max
- Fonctions d'agrégation pour les types de mesure Nominal ou Catégorielle :
  - Mode
  - Min
  - Max
- Fonctions d'agrégation pour un type de mesure Indicateur :
  - True, si un élément a pour valeur true
  - False, si un élément a pour valeur false

Couleur

Utilisez cette option pour choisir une couleur standard à appliquer à toutes les fonctions du champ géospatial ou un champ de superposition, qui colore les fonctions selon les valeurs d'un autre champ dans les données.

Si vous sélectionnez **Standard**, vous pouvez choisir une couleur dans la palette de couleurs qui apparaît dans le panneau **Ordre des couleurs des graphiques** dans l'onglet Afficher de la boîte de dialogue Options utilisateur.

Si vous sélectionnez **Superposition**, vous pouvez choisir un champ depuis la source de données contenant le champ géospatial qui a été sélectionné comme **Zone Couche**.

- Pour les champs de superposition Nominal ou Catégorielle, la palette de couleurs dans laquelle vous pouvez choisir une couleur est la même que celle qui est proposée pour les options de couleur **Standard**.
- Pour les champs de superposition Continu et Ordinal, une deuxième liste déroulante est affichée, dans laquelle vous pouvez sélectionner une couleur. Lorsque vous sélectionnez une couleur, la superposition est appliquée en variant la saturation de cette couleur selon les valeurs qui figurent dans le champ Continu ou Ordinal. La valeur la plus élevée utilise la couleur choisie dans la liste déroulante et les valeurs inférieures sont affichées par saturations inférieures correspondantes.

Symbole

Remarque : Activé pour les types de mesure Point et Multipoint seulement.

Utilisez cette option pour choisir un symbole **Standard**, qui est appliqué à tous les enregistrements de votre champ géospatial, ou un symbole **Superposition**, qui change l'icône de symbole pour les points en fonction des valeurs d'un autre champ dans vos données.

Si vous sélectionnez **Standard**, vous pouvez choisir l'un des symboles par défaut dans une liste déroulante pour représenter les données de point sur la carte.

Si vous sélectionnez **Superposition**, vous pouvez choisir un champ Nominal, Ordinal ou Catégorielle dans la source de données contenant le champ géospatial qui a été sélectionné comme **Zone Couche**. Pour chaque valeur dans le champ de superposition, un symbole différent est affiché sur la carte.

Par exemple, vos données peuvent contenir un champ de points qui représente les emplacements des magasins et la superposition peut être un champ de type de magasin. Dans cet exemple, tous les restaurants sont identifiés sur la carte par une croix et tous les magasins d'électronique par un carré.

Taille

Remarque : Activé uniquement pour les types de mesure Point, Multipoint, Chaîne et Multichaîne.

Utilisez cette option pour choisir une taille **Standard**, qui est appliquée à tous les enregistrements de votre champ géospatial ou une taille **Superposition**, qui change la taille de l'icône de symbole ou l'épaisseur de la ligne en fonction des valeurs d'un autre champ dans vos données.

Si vous sélectionnez **Standard**, vous pouvez choisir une valeur de largeur en pixels. Les options disponibles sont 1, 2, 3, 4, 5, 10, 20 ou 30.

Si vous sélectionnez **Superposition**, vous pouvez choisir un champ depuis la source de données contenant le champ géospatial qui a été sélectionné comme **Zone Couche**. L'épaisseur de la ligne ou du point varie selon la valeur du champ choisi.

Transparence

Utilisez cette option pour choisir une transparence **Standard**, qui est appliquée à tous les enregistrements de votre champ géospatial, ou une transparence **Superposition**, qui change la transparence du symbole, de la ligne ou du polygone en fonction des valeurs d'un autre champ dans vos données.

Si vous sélectionnez **Standard**, vous pouvez effectuer votre choix parmi plusieurs niveaux de transparence, de 0 % (opaque) puis par incrément de 10 % jusqu'à 100 % (transparent).

Si vous sélectionnez **Superposition**, vous pouvez choisir un champ depuis la source de données contenant le champ géospatial qui a été sélectionné comme **Zone Couche**. Un niveau de transparence différent est affiché sur la carte pour chaque valeur dans le champ de superposition. La transparence est appliquée à la couleur choisie dans la liste déroulante des couleurs pour le point, la ligne ou le polygone.

Libellé des données

Remarque : Cette option n'est pas disponible si vous sélectionnez la case à cocher **Utiliser le compartimentage hexagonal**.

Utilisez cette option pour sélectionner un champ à utiliser comme libellé de données sur la carte. Par exemple, s'il est appliqué à une couche polygone, le libellé de données peut être le champ de nom contenant le nom de chaque polygone. Si vous sélectionnez le champ de nom, ces noms sont affichés sur la carte.

Onglet Apparence de la visualisation de carte

Vous pouvez spécifier les options d'apparence avant de créer le graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous-titre du graphique.

Légende. Saisissez le texte à utiliser comme légende du graphique.

Noeud t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE)© est un outil permettant de visualiser des données en grande dimension. Il convertit les analogies de points de données en probabilités. Les analogies de l'espace initial sont représentées par des probabilités jointes de Gauss et les analogies de l'espace intégré sont représentées par des lois t de Student. Cela rend t-SNE particulièrement sensible à la structure locale et présente d'autres avantages par rapport aux techniques existantes : [1]

- Indique la structure à plusieurs échelles sur une seule mappe
- Révèle les données stockées dans plusieurs collecteurs ou clusters différents
- Réduit l'agglomération de points au centre

Le noeud t-SNE de SPSS Modeler est mis en oeuvre dans Python et nécessite la bibliothèque Python `scikit-learn`©. Pour plus d'informations sur t-SNE et la bibliothèque scikit-learn, voir :

- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE
- https://scikit-learn.org/stable/modules/manifold.html#t-sne

L'onglet Python sur la Palette de noeuds contient ce noeud et d'autres noeuds Python. Le noeud t-SNE est également disponible dans l'onglet Graphiques.

[1] Références :

van der Maaten, L.J.P.; Hinton, G. "Visualizing High-Dimensional Data using t-SNE." Journal of Machine Learning Research. 9:2579-2605, 2008.

van der Maaten, L.J.P. "t-Distributed Stochastic Neighbor Embedding."

van der Maaten, L.J.P. "Accelerating t-SNE using Tree-Based Algorithms." Journal of Machine Learning Research. 15(Oct):3221-3245, 2014.

Options expert du noeud t-SNE

Sélectionnez le mode **Simple** ou **Expert** en fonction des options que vous souhaitez définir pour le noeud t-SNE.

Type de visualisation. Sélectionnez **2D** ou **3D** pour indiquer s'il faut tracer le graphique en deux ou trois dimensions.

Méthode : Sélectionnez **Barnes Hut** ou **Exact**. Par défaut, l'algorithme de calcul de gradient utilise la simulation de Barnes-Hut, laquelle s'exécute bien plus rapidement que la méthode Exact. La simulation de Barnes-Hut permet d'appliquer la technique t-SNE à des jeux de données réels volumineux. L'algorithme Exact permettra d'éviter plus efficacement les erreurs liées aux voisins les plus proches.

Init. Sélectionnez **Aléatoire** ou **ACP** pour l'initialisation de l'incorporation.

Champ cible. Sélectionnez l'affichage sous forme de table des couleurs pour la zone cible sur le graphique de sortie. Le graphique utilisera une seule couleur si aucune zone cible n'est indiquée à cet endroit.

Optimisation

Perplexité. La perplexité est liée au nombre de voisins les plus proches qui sont utilisés dans d'autres algorithmes d'apprentissage divers. Les jeux de données plus volumineux nécessitent généralement une plus grande perplexité. Sélectionnez de préférence une valeur comprise entre **5** et **50**. La valeur par défaut est **30**, et l'intervalle est **2 - 9999999**.

Exagération précoce. Ce paramètre détermine à quel point les clusters naturels de l'espace d'origine seront proches dans l'espace intégré, ainsi que l'espace qui les séparera. La valeur par défaut est **12**, et l'intervalle est **2 - 9999999**.

Taux d'apprentissage. Si le taux d'apprentissage est trop élevé, les données peuvent ressembler à une "balle" : tous les points sont à peu près équidistants de leurs voisins les plus proches. Si le taux d'apprentissage est trop faible, il se peut que tous les points soient compressés dans un nuage dense avec peu de valeurs éloignées. Si la fonction de coût se retrouve bloquée dans un mauvais minimum local, il est recommandé d'essayer d'augmenter le taux d'apprentissage. La valeur par défaut est **200**, et l'intervalle est **0 - 9999999**.

Itérations max Nombre maximal d'itérations pour l'optimisation. La valeur par défaut est **1000**, et l'intervalle est **250 - 9999999**.

Taille angulaire. Taille angulaire d'un noeud distant mesurée à partir d'un point. Entrez une valeur entre **0** et **1**. La valeur par défaut est **0,5**.

Valeur de départ aléatoire

Définir une valeur de départ aléatoire. Sélectionnez cette option et cliquez sur **Générer** pour générer la valeur de départ utilisée par le générateur de nombres aléatoires.

Condition d'arrêt de l'optimisation

Itérations max sans avancement. Le nombre maximum d'itérations sans avancement à effectuer avant d'arrêter l'optimisation, utilisé après 250 itérations initiales avec exagération précoce. Sachez que l'avancement n'est vérifié que toutes les 50 itérations. Cette valeur est donc arrondie au prochain multiple de 50. La valeur par défaut est **300**, et l'intervalle est **0 - 9999999**.

Norme de gradient min. Si la norme de gradient est inférieure à ce seuil minimum, l'optimisation s'arrêtera. La valeur par défaut est **1.0E-7**.

Mesure. La métrique à utiliser pour calculer la distance entre des instances dans un tableau de fonctions. Si la métrique est une chaîne, le paramètre de métrique doit correspondre à l'une des options autorisées par `scipy.spatial.distance.pdist`, ou une métrique répertoriée dans `pairwise.PAIRWISE_DISTANCE_FUNCTIONS`. Sélectionnez l'un des types de métrique disponibles. La valeur par défaut est **euclidean**.

Lorsque le nombre d'enregistrements est supérieur à. Spécifiez la méthode de représentation des jeux de données volumineux. Vous pouvez spécifier la taille maximale des jeux de données ou utiliser les 2 000 points par défaut. Lorsque vous sélectionnez les options **Intervalle** ou **Echantillon**, les performances des jeux de données volumineux sont optimisées. Vous pouvez également choisir de représenter tous les points de données en sélectionnant **Utiliser toutes les données**, mais sachez que vous risquez de réduire considérablement les performances du logiciel.

- **Casier.** Sélectionnez cette option pour permettre la création de casiers lorsque le jeu de données contient plus d'enregistrements que le nombre spécifié. La création d'intervalles applique une fine grille au graphique avant que soit effectué le traçage réel et compte le nombre de connexions apparaissant dans chacune des cellules de la grille. Dans le graphique final, une connexion est représentée dans chaque cellule, au niveau du centroïde de l'intervalle (moyenne de tous les emplacements de connexion de l'intervalle).
- **Echantillon.** Sélectionnez cette option pour échantillonner de façon aléatoire les données dans le nombre d'enregistrements spécifié.

Le tableau suivant montre la relation entre les paramètres de l'onglet Expert de la boîte de dialogue du noeud t-SNE SPSS Modeler et ceux de la bibliothèque Python t-SNE.

Tableau 36. Relation entre les propriétés du noeud et les paramètres de la bibliothèque Python

Paramètre SPSS Modeler	Nom du script (nom de la propriété)	Paramètre Python t-SNE
Mode	`mode_type`	
Type de visualisation	`n_components`	`n_components`
Méthode	`method`	`method`
Initialisation de l'incorporation	`init`	`init`
Cible	`target_field`	`target_field`
Perplexité	`perplexity`	`perplexity`

Tableau 36. Relation entre les propriétés du noeud et les paramètres de la bibliothèque Python (suite)

Paramètre SPSS Modeler	Nom du script (nom de la propriété)	Paramètre Python t-SNE
Exagération précoce	early_exaggeration	early_exaggeration
Taux d'apprentissage	learning_rate	learning_rate
Itérations max	n_iter	n_iter
Taille angulaire	angle	angle
Définir une valeur de départ aléatoire	enable_random_seed	
Valeur de départ aléatoire	random_seed	random_state
Itérations max sans avancement	n_iter_without_progress	n_iter_without_progress
Norme de gradient min	min_grad_norm	min_grad_norm
Exécuter t-SNE avec plusieurs perplexités	isGridSearch	

Options de sortie du noeud t-SNE

Indiquez des options pour la sortie du noeud t-SNE dans l'onglet **Sortie**.

Nom de la sortie - Indique le nom de la sortie générée lorsque le noeud est en cours d'exécution. Si vous sélectionnez **Automatique**, le nom de la sortie est défini automatiquement.

Sortie à l'écran. Sélectionnez cette option pour générer et afficher la sortie dans une nouvelle fenêtre. La sortie est également ajoutée au gestionnaire des sorties.

Sortie dans un fichier. Sélectionnez cette option pour enregistrer la sortie dans un fichier. Cette sélection active les zones **Fichier** et **Type de fichier**. Le noeud t-SNE doit pouvoir accéder à ce fichier de sortie si vous voulez créer des tracés à l'aide d'autres champs à des fins de comparaison, ou si vous voulez utiliser sa sortie en tant que prédicteurs dans des modèles de classification ou de régression. Le modèle t-SNE crée un fichier de résultats avec des champs de coordonnées x, y (et z) qui est facilement accessible à l'aide d'un noeud source Fichier fixe. Pour plus d'informations, voir @@@@@@.

Le tableau suivant montre la relation entre les paramètres de l'onglet Sortie de la boîte de dialogue du noeud t-SNE SPSS Modeler et ceux de la bibliothèque Python t-SNE.

Tableau 37. Relation entre les propriétés du noeud et les paramètres de la bibliothèque Python

Paramètre SPSS Modeler	Nom du script (nom de la propriété)	Paramètre Python t-SNE
Nom de la sortie	output_Rename	output_Rename
Mode de sortie	output_to	output_to
Fichier	full_filename	full_filename
Type de fichier	output_file_type	output_file_type
Cible	target_field	target_field

Accès et traçage de données t-SNE

Si vous utilisez l'option **Sortie dans le fichier** pour enregistrer une sortie t-SNE dans des fichiers, vous pouvez ensuite créer des tracés à l'aide d'autres champs à des fins de comparaison. Vous pouvez également utiliser la sortie en tant que prédicteurs dans des modèles de classification ou de régression. Le modèle t-SNE crée un fichier de résultats avec des champs de coordonnées x, y (et z) qui est facilement accessible à l'aide d'un noeud source Fichier fixe. Cette section fournit des informations d'exemple.

1. Dans la boîte de dialogue du noeud t-SNE, ouvrez l'onglet **Sortie**.

2. Sélectionnez **Sortie dans le fichier** et saisissez un nom de fichier. Utilisez le type de fichier HTML défini par défaut. L'exécution du modèle génèrera trois fichiers de sortie dans votre emplacement de sortie :
   - Un fichier texte (`result_xxxxxx.txt`)
   - Un fichier HTML (le nom de fichier que vous avez saisi)
   - Un fichier PNG (`tsne_chart_yyyyyy.png`)

   Le fichier texte contient les données dont vous avez besoin. Toutefois, pour des raisons techniques, il se peut qu'il s'affiche dans un format standard ou scientifique. S'il s'affiche dans un format scientifique (1.11111111e+01), vous devez créer un nouveau flux qui reconnaît le format :

Accès aux données de tracé t-SNE lorsque le fichier texte s'affiche dans un format numérique scientifique

1. Créez un nouveau flux (**Fichier** > **New Stream**).
2. Accédez à **Outils** > **Propriétés de flux** > **Options**, puis sélectionnez **Formats des nombres** et **Scientifique (#.###E+##)** pour le format d'affichage des nombres.
3. Ajoutez un noeud source Fichier fixe à votre espace de travail et utilisez les paramètres suivants dans l'onglet Fichier :
   - Ignorer les lignes des en-têtes : 1
   - Longueur de l'enregistrement : 54
   - Début tSNE_x : 3, Longueur : 16
   - Début tSNE_y : 20, Longueur : 16
   - Début tSNE_z : 36, Longueur: 16
4. Dans l'onglet Type, les chiffres doivent être reconnus comme réels. Cliquez sur Lire les valeurs. Les valeurs des champs doivent ressembler à :

Tableau 38. Exemple de valeurs de champ

Champ	Mesure	Valeurs
tSNE_x	Continu	[-7.07176703,7.14338837]
tSNE_y	Continu	[-9.2188112,8.89647667]
tSNE_x	Continu	[-9.95892882,9.95742482]

5. Ajoutez un noeud Sélectionner au flux afin de pouvoir supprimer les deux lignes de texte du bas dans le fichier qui sont lues comme valeurs nulles :

   ```
   *****************
   Exécutez t-SNE (durée totale de 9,5 s)
   ```

   Dans l'onglet Paramètres du noeud Sélectionner, choisissez **Supprimer** pour le mode et utilisez la condition `@NULL(tSNE_x)` pour supprimer les lignes.
6. Ajoutez un noeud Type et un noeud d'exportation Fichier plat au flux pour créer un intermédiaire d'offres à valeur ajoutée. Noeud source de fichier qui sera copié et collé dans votre flux original.

Accès aux données de tracé t-SNE lorsque le fichier texte s'affiche dans un format numérique standard

1. Créez un nouveau flux (**Fichier** > **New Stream**).
2. Ajoutez un noeud source Fichier fixe à votre espace de travail. Les trois noeuds suivants sont les seuls éléments nécessaires pour accéder aux données t-SNE.

Figure 48. Stream for accessing t-SNE plot data in standard numeric format

3. Utilisez les paramètres suivants dans l'onglet Fichier du noeud source Fichier fixe :
   - Ignorer les lignes des en-têtes : 1
   - Longueur de l'enregistrement : 29
   - Début tSNE_x : 3, Longueur : 12
   - Début tSNE_y : 16, Longueur : 12
4. Dans l'onglet Filtre, vous pouvez renommer `champ1` et `champ2` en `tsneX` and `tsneY`.
5. Ajoutez un noeud Fusionner pour le connecter à votre flux à l'aide de la méthode de fusion **Ordre**.
6. Vous pouvez désormais utiliser un noeud Tracépour tracer `tsneX` versus `tsneY` et le colorer avec votre champ à l'étude.

Nuggets de modèle t-SNE

Les nuggets de modèle t-SNE contiennent toutes les informations rassemblées par le modèle t-SNE. Les onglets suivants sont disponibles.

Graphique

L'onglet **Graphique** affiche la sortie de graphique du noeud t-SNE. Un graphique à nuage de points pyplot affiche le résultat à basses dimensions. Si vous n'avez pas sélectionné l'option **Exécuter t-SNE avec plusieurs perplexités** dans l'onglet Expert du noeud t-SNE, un seul graphique est inclus au lieu des six graphiques correspondant aux différentes perplexités.

Sortie texte

L'onglet **Sortie texte** affiche les résultats de l'algorithme t-SNE. Si vous avez sélectionné le type de visualisation **2D** dans l'onglet Expert du noeud t-SNE, le résultat affiché à cet endroit représente la valeur de point en deux dimensions. Si vous avez sélectionné **3D**, le résultat représente la valeur de point en trois dimensions.

Noeud E-Tracé (Bêta)

Les noeuds E-Tracé (Bêta) montrent les relations existant entre les champs numériques. Le noeud E-Tracé (Bêta) ressemble au noeud Tracé, mais il offre d'autres options et de nouvelles capacités de graphique. Utilisez ce noeud pour vous familiariser avec les nouvelles fonctions de graphique dans SPSS Modeler.

Le noeud E-Tracé (Bêta) fournit des nuages de points, des graphiques curvilignes et des graphiques à barres pour illustrer les relations entre les champs numériques. La nouvelle interface de graphique de ce noeud est intuitive, moderne et personnalisable. De plus, les graphiques de données sont interactifs. Pour plus d'informations, voir «Utilisation d'un graphique E-Tracé», à la page 287.

Noeud E-Tracé (Bêta), onglet Nuage

Les graphiques Nuage comparent les valeurs d'un champ *Y* à celles d'un champ *X*. En général, ces champs correspondent respectivement à une variable dépendante et à une variable indépendante.

Champ X. Dans la liste, sélectionnez le champ à afficher sur l'axe *x* horizontal.

Champ Y. Dans la liste, sélectionnez le champ à afficher sur l'axe *y* vertical.

Superposition. Il existe plusieurs méthodes pour mettre en évidence les catégories des valeurs de données. Par exemple, vous pouvez utiliser un champ *récolteprincipale* en tant que superposition de couleurs afin d'indiquer les valeurs *revenuest* et *valeurréclamation* de la récolte principale cultivée par les demandeurs. Sélectionnez les champs pour le mappage des couleurs, des tailles et des formes dans la sortie. Sélectionnez également tous les autres champs qui vous intéressent pour les inclure dans la sortie interactive. Pour plus d'informations, voir «Apparences, superpositions, panneaux et animation», à la page 198.

Une fois les options du e-tracé définies, vous pouvez exécuter le tracé directement à partir de la boîte de dialogue en cliquant sur **Exécuter**. Cependant, si vous le souhaitez, vous pouvez utiliser l'onglet Options pour ajouter des spécifications.

Noeud E-Tracé (Bêta), onglet Options

Nombre maximal d'enregistrements dans le tracé. Spécifiez la méthode de représentation des jeux de données volumineux. Vous pouvez spécifier le nombre de modalités maximales des jeux de données ou utiliser les 2 000 enregistrements par défaut. Lorsque vous sélectionnez l'option **Echantillon**, les performances des jeux de données volumineux sont optimisées. L'option d'échantillonnage échantillonne de façon aléatoire les données dans le nombre d'enregistrements saisi dans le champ de texte. Vous pouvez également choisir de représenter tous les points de données en sélectionnant **Utiliser toutes les données**, mais sachez que vous risquez de réduire considérablement les performances du logiciel.

Onglet Apparence, noeud E-Tracé (Bêta)

Si vous le souhaitez, vous pouvez indiquer un titre et un sous-titre avant la création du graphique. Ces options peuvent également être renseignées ou modifiées après la création du graphique.

Titre. Saisissez le texte à utiliser comme titre du graphique.

Sous-titre. Saisissez le texte à utiliser comme sous-titre du graphique.

Utilisation d'un graphique E-Tracé

Le noeud E-Tracé (Bêta) fournit des nuages de points, des graphiques curvilignes et des graphiques à barres pour illustrer les relations entre les champs numériques. La nouvelle interface de graphique introduite dans ce noeud bêta dispose d'un grand nombre de capacités exclusives.

Figure 49. E-Plot (Beta) scatterplot graph

Dans l'onglet Graphique, le coin supérieur gauche fournit une barre d'outils pour effectuer un zoom avant sur une section spécifique du graphique, un zoom arrière, revenir à la vue complète initiale et enregistrer le graphique pour un usage externe :

Figure 50. Toolbar

En bas de la fenêtre, vous pouvez utiliser le curseur pour effectuer un zoom avant sur une section spécifique du graphique. Déplacez les petits contrôles rectangulaires vers la droite et la gauche pour effectuer un zoom. Pour utiliser ce curseur, vous devez commencer par l'activer dans les options de la boîte d'outils.

Figure 51. Zoom slider

A gauche de la fenêtre, vous trouverez des contrôles pour modifier la plage de valeurs affichée. Pour utiliser ces contrôles, vous devez d'abord renseigner des options dans Data Map. Dans l'exemple ci-dessous, un champ appelé `PM25` est sélectionné pour la carte de couleur, un champ appelé `PM10` est sélectionné pour la carte de taille et un champ appelé `City` est sélectionné pour la carte de forme. Vous pouvez survoler les barres de couleur verticales pour surligner les zones correspondantes du graphique, ou faire glisser les triangles supérieur et inférieur.

Figure 52. Range controls

A droite de la fenêtre, un ensemble d'options extensibles est disponible. Vous pouvez les utiliser pour interagir avec les données et modifier l'apparence du graphique en temps réel :

Figure 53. Expandable options

Options basiques

Figure 54. Basic options

Sélectionnez le thème sombre ou clair, indiquez un titre et un sous-titre, sélectionnez un type de graphique (nuages de points, linéaire, à barres) et choisissez la série à afficher sur l'axe Y. Si vous sélectionnez le graphique **Linéaire**, seuls les champs de l'axe Y seront affichés, et seuls ces mêmes champs seront disponibles dans les options Data Map pour les cartes de couleur et de taille. Si vous sélectionnez le graphique **Barres**, seules les options de la carte de couleur seront disponibles dans les options Data Map. Pour les séries, tous les champs **Interested** que vous avez sélectionnés dans l'onglet Nuage du noeud E-Tracé seront disponibles ici.

Options Data Map

Figure 55. Data map options

Sélectionnez un champ continu ou un champ catégoriel pour **Carte de couleur**. Si un champ continu est sélectionné, les couleurs de vert à rouge s'affichent. Plus la valeur est faible, plus la couleur est proche du rouge. Et plus la valeur est élevée, plus la valeur est proche du vert. Si un champ catégoriel est sélectionné, la couleur du champ s'affiche conformément à la palette de couleurs définie.
Size Map ne prend en charge que les champs continus. Plus la valeur sur le graphique est faible, plus la taille du nuage sera petite.
Shape Map ne prend en charge que les champs catégoriels. La forme affichée sur la carte est définie par un champ catégoriel qui scinde la visualisation en éléments de formes différentes, un élément pour chaque catégorie.

Options de la palette

Figure 56. Palette options

Utilisez la palette si vous voulez personnaliser les couleurs du titre et des séries. Sélectionnez le titre ou les séries dans le menu déroulant, cliquez sur **Edit Predefined Colors**, puis sur **More** pour choisir une couleur. Vous pouvez également utiliser les champs RGB ou Hex pour indiquer une couleur exacte.

Options de la boîte à outils

Figure 57. Toolbox options

Utilisez les options de la boîte à outils pour activer ou désactiver le curseur du zoom, définir les propriétés du quadrillage et activer ou désactiver le suivi de la souris. Le suivi de la souris affiche les coordonnées exactes de votre position lorsque vous survolez le graphique.

Exploration de graphiques

Tandis que le mode d'édition vous permet de modifier la mise en forme et l'aspect du graphique, le mode d'interaction vous permet d'explorer de manière analytique les données et les valeurs représentées par le graphique. Le principal objectif de l'exploration est d'analyser les données puis d'identifier les valeurs à l'aide de bandes, de zones et de marquages pour créer des noeuds Sélectionner, Calculer ou Equilibrer. Pour sélectionner ce mode, sélectionnez **Affichage > Mode d'interaction** dans les menus (ou cliquez sur l'icône de la barre d'outils).

Bien que certains graphiques puissent utiliser tous les outils d'exploration, d'autres n'en acceptent qu'un seul. Le mode d'interaction comprend :

- La définition et la modification de bandes, qui permettent de fractionner les valeurs le long d'un axe d'échelle *x*. Pour plus d'informations, voir «Utilisation de bandes».
- La définition et l'édition de zones, qui permettent d'identifier un groupe de valeurs dans une zone rectangulaire. Pour plus d'informations, voir «Présentation des zones», à la page 295.
- Le marquage ou l'annulation du marquage d'éléments pour choisir vous-même les valeurs à utiliser pour créer un noeud Sélectionner ou Calculer. Pour plus d'informations, voir «Présentation des éléments marqués», à la page 297.
- La création de noeuds à l'aide des valeurs identifiées par les bandes, les zones, les éléments marqués et les liens de relations à utiliser dans votre flux. Pour plus d'informations, voir «Génération de noeuds à partir de graphiques», à la page 298.

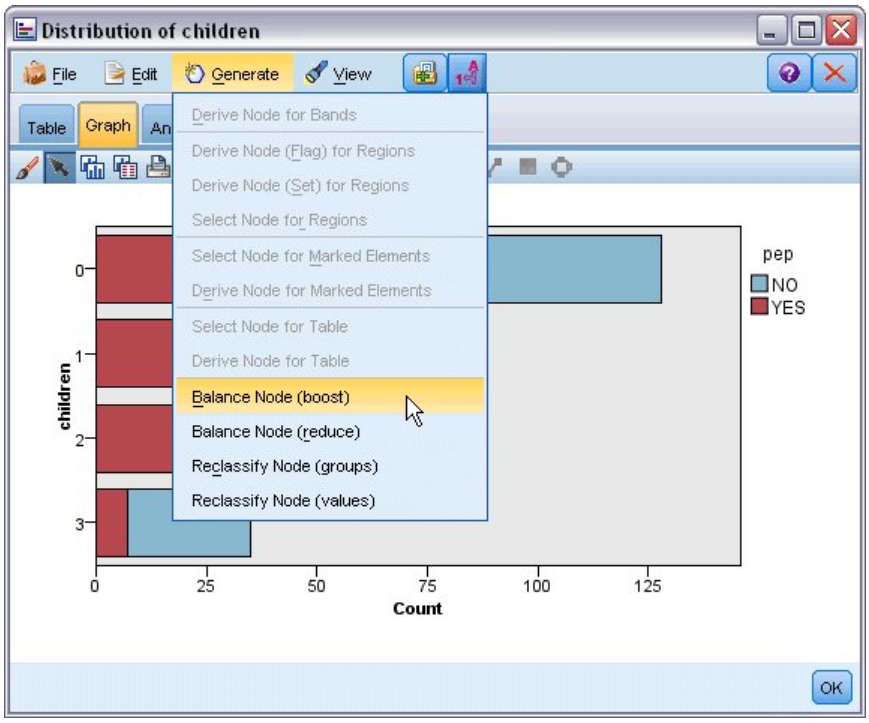


Figure 58. Graphique avec le menu Générer affichant

Utilisation de bandes

Dans tout graphique doté d'un champ d'échelle sur l'axe *x*, vous pouvez dessiner des lignes de bande verticales pour fractionner l'intervalle de valeurs sur l'axe *x*. Si un graphique contient plusieurs panneaux, une ligne de bande dessinée sur un panneau est également représentée sur les autres panneaux.

Certains graphiques n'acceptent pas les bandes. Voici certains des graphiques pouvant contenir des bandes : les histogrammes, les graphiques à barres et proportion, les graphiques tracés (linéaires, de dispersion, horaires, etc.), résumés, et les graphiques d'évaluation. Dans les graphiques divisés en panneaux, les bandes apparaissent sur tous les panneaux. Et dans certains cas d'une matrice SPLOM, vous verrez s'afficher une ligne de bande horizontale du fait que l'axe sur lequel a été dessinée la bande de champ/variable a été inversé.

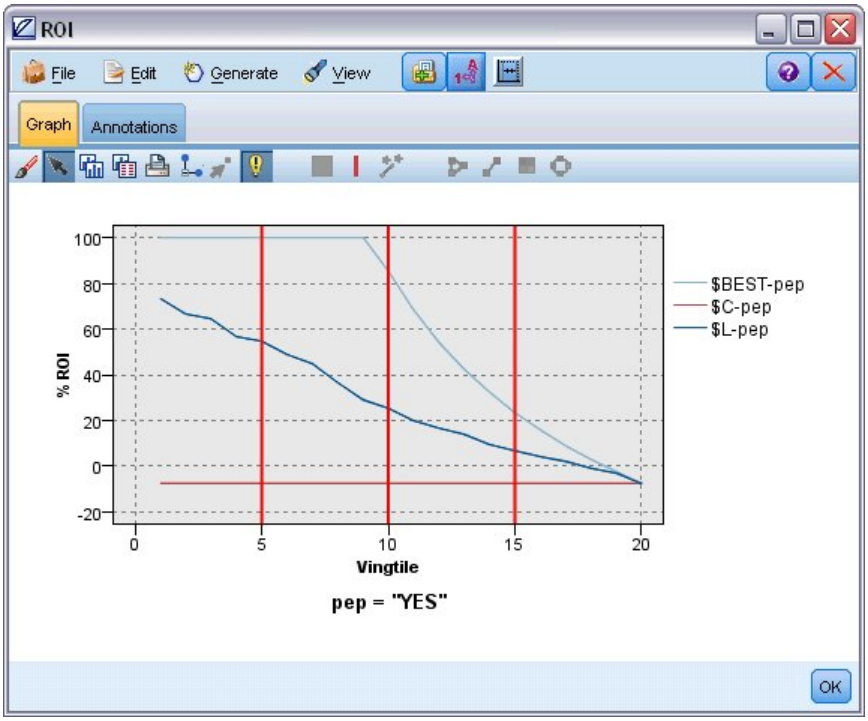


Figure 59. Graphique à trois bandes

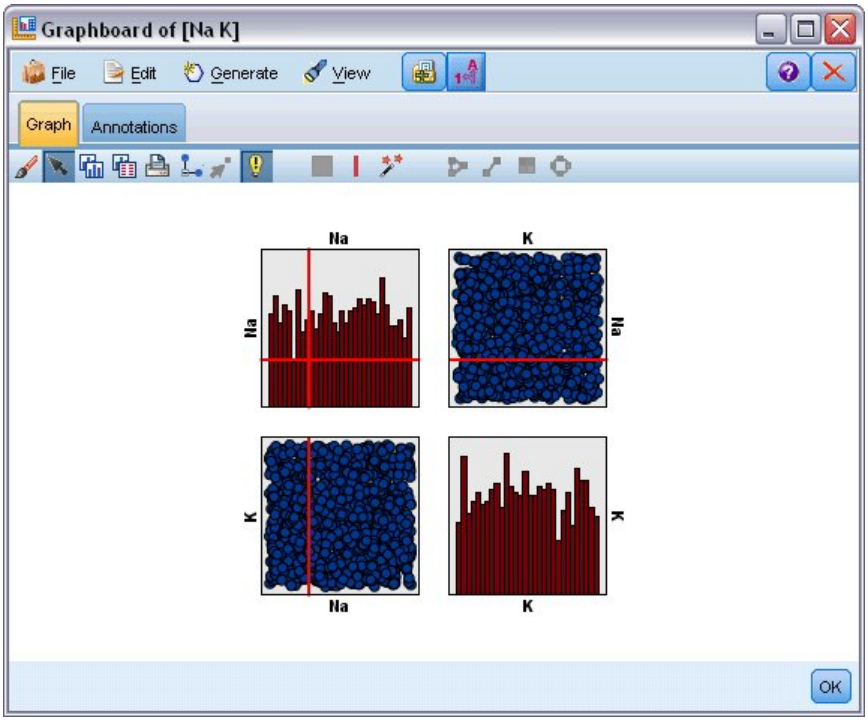


Figure 60. SPLOM avec bandes

Définition des bandes

Dans un graphique sans bande, l'ajout d'une ligne de bande fractionne le graphique en deux bandes. La valeur de la ligne de bande représente le point de départ, également appelé limite inférieure, de la

deuxième bande lors de la lecture du graphique de gauche à droite. De la même manière, dans un graphique à deux bandes, l'ajout d'une ligne de bande fractionne l'une de ces bandes en deux, créant ainsi une troisième bande. Par défaut, les bandes sont nommées *bandN*, où *N* correspond au nombre de bandes, de gauche à droite, sur l'axe *x*.

Une fois que vous avez défini une bande, vous pouvez utiliser la fonction glisser-déposer pour la repositionner sur l'axe *x*. Vous pouvez accéder à d'autres raccourcis en cliquant avec le bouton droit à l'intérieur de la bande pour des tâches telles que renommer, supprimer ou créer des noeuds pour cette bande en particulier.

Pour définir des bandes :

1. Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez **Affichage > Mode d'interaction**.
2. Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Dessiner une bande.

Figure 61. Bouton de barre d'outils Dessiner des bandes

3. Dans un graphique qui accepte les bandes, cliquez sur le point de valeur de l'axe *x* au niveau duquel vous voulez définir une ligne de bande.

Remarque : Vous pouvez également cliquer sur l'icône de la barre d'outils **Diviser le graphique en bandes** et saisir le nombre de bandes égales souhaitées, puis cliquer sur **Fractionner**.

Figure 62. Icône de fractionnement utilisée pour développer la barre d'outils et afficher les options permettant de fractionner l'axe en bandes

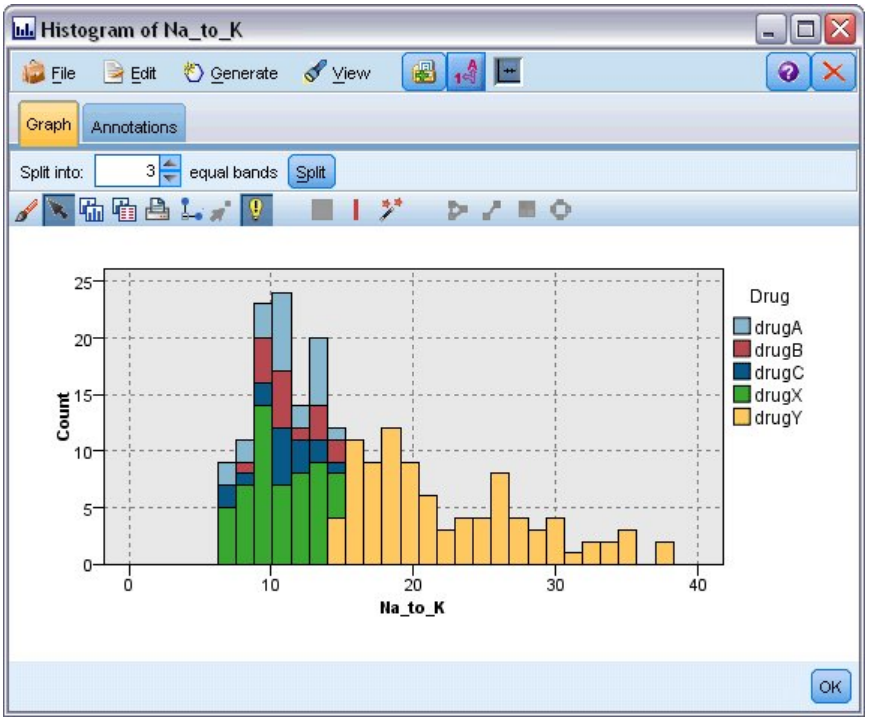


Figure 63. Barre d'outils Création de bandes égales avec bandes activées

Modification, changement de nom, et suppression de bandes

Vous pouvez modifier les propriétés des bandes existantes dans la boîte de dialogue Modifier les bandes du graphique ou via les menus contextuels dans le graphique lui-même.

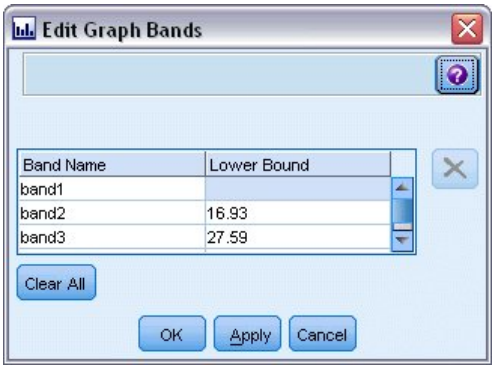


Figure 64. Boîte de dialogue Modifier les bandes du graphique

Pour modifier des bandes :

1. Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez **Affichage > Mode d'interaction**.
2. Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Dessiner une bande.
3. Dans les menus, choisissez **Edition > Bandes du graphique**. La boîte de dialogue Modifier les bandes du graphique s'ouvre.
4. Si vous avez plusieurs champs dans votre graphique (graphiques SPLOM par exemple), vous pouvez sélectionner le champ souhaité dans la liste déroulante.
5. Ajoutez une nouvelle bande en saisissant un nom et une limite inférieure. Appuyez sur la touche Entrée pour commencer une nouvelle ligne.

6. Modifiez la frontière d'une bande en ajustant la valeur de la **Limite inférieure**.
7. Renommez une bande en saisissant un nouveau nom de bande.
8. Supprimez une bande en sélectionnant la ligne dans le tableau, puis en cliquant sur le bouton de suppression.
9. Cliquez sur **OK** pour appliquer vos changements et fermer la boîte de dialogue.

Remarque : Vous pouvez également supprimer et renommer les bandes directement dans le graphique en cliquant avec le bouton droit de la souris sur la ligne de la bande et en choisissant l'option souhaitée dans les menus contextuels.

Présentation des zones

Dans tout graphique contenant deux axes d'échelle (ou d'intervalle), vous pouvez dessiner des zones pour regrouper des valeurs dans un rectangle, appelé zone. Une **zone** est une partie du graphique définie par ses valeurs *X* et *Y* minimales et maximales. Si un graphique contient plusieurs panneaux, une zone dessinée sur un panneau est également représentée sur les autres panneaux.

Certains graphiques n'acceptent pas les zones. Voici certains des graphiques qui acceptent les zones : les graphiques tracés (linéaires, de dispersion, en bulles, horaires, etc.), les matrices SPLOM et les résumés. Ces zones sont dessinées dans un espace X,Y, et il est par conséquent impossible de les définir dans les graphiques en 1D, 3D ou animés. Dans les graphiques divisés en panneaux, les zones apparaissent sur tous les panneaux. Dans le cas d'une matrice de diagramme de dispersion (SPLOM), une zone correspondante apparaît dans les graphiques supérieurs correspondants, mais pas dans les graphiques en diagonale car ils n'affichent qu'un seul champ d'échelle.

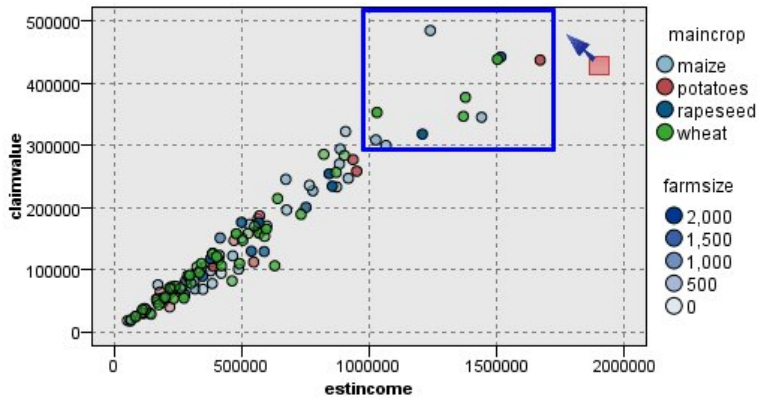


Figure 65. Définition d'une zone dont les valeurs des demandes sont élevées

Définition des zones

Quelque soit l'endroit où vous définissez une zone, vous créez un groupement de valeurs. Par défaut, chaque nouvelle zone est appelée *Zone<N>*, où *N* correspond au nombre de zones déjà créées.

Une fois que vous avez défini une zone, vous pouvez cliquer avec le bouton droit de la souris sur la ligne de la zone pour accéder à certains raccourcis de base. Vous pouvez toutefois accéder à de nombreux autres raccourcis en cliquant avec le bouton droit à l'intérieur de la zone (et non sur la ligne) pour des tâches telles que le changement de nom, la suppression ou la création de noeuds Sélectionner et Calculer pour cette zone en particulier.

Vous pouvez sélectionner des sous-ensembles d'enregistrements en fonction de leur appartenance à une zone particulière ou à une zone parmi d'autres. Vous pouvez également incorporer à l'enregistrement des informations sur la zone en générant un noeud Calculer de sorte à ajouter un indicateur aux

enregistrements en fonction de leur appartenance à une zone. Pour plus d'informations, voir «Génération de noeuds à partir de graphiques», à la page 298.

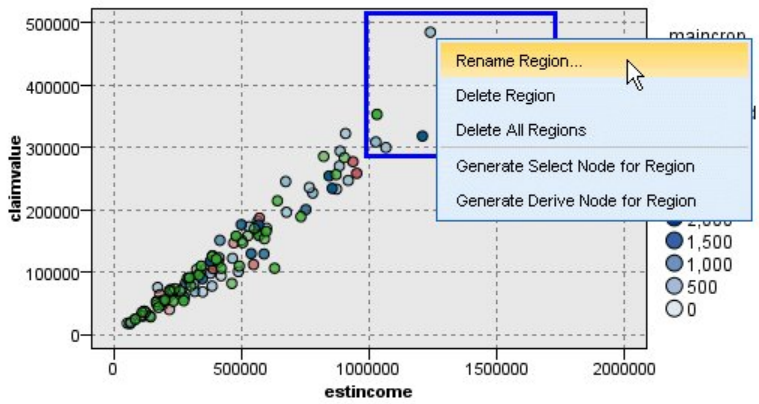


Figure 66. Exploration de la zone dont les valeurs des demandes sont élevées

Pour définir des zones :

1. Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez **Affichage > Mode d'interaction**.
2. Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Dessiner une zone.

Figure 67. Bouton de barre d'outils Dessiner une zone

3. Dans un graphique qui accepte les zones, cliquez et faites glisser votre souris pour dessiner la zone rectangulaire.

Modification, changement de nom, et suppression de zones

Vous pouvez modifier les propriétés des zones existantes dans la boîte de dialogue Modifier les zones du graphique ou via les menus contextuels dans le graphique lui-même.

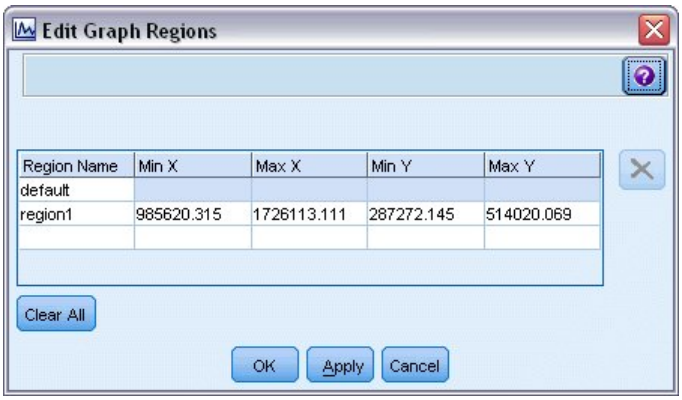


Figure 68. Spécification des propriétés des zones définies

Pour modifier des zones :

1. Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez **Affichage > Mode d'interaction**.

2. Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Dessiner une zone.
3. Dans les menus, choisissez **Edition > Zones du graphique**. La boîte de dialogue Modifier les zones du graphique s'ouvre.
4. Si vous avez plusieurs champs dans votre graphique (graphiques SPLOM par exemple), vous devez définir le champ de la zone dans les colonnes *Champ A* et*Champ B*.
5. Pour ajouter une nouvelle zone sur une nouvelle ligne, saisissez un nom, sélectionnez les noms des champs (le cas échéant) et définissez les limites minimum et maximum pour chaque champ. Appuyez sur la touche Entrée pour commencer une nouvelle ligne.
6. Modifiez les limites existantes de la zone en rectifiant les valeurs **Minimum** et **Maximum** de *A* et de *B*.
7. Pour renommer une zone, modifiez son nom dans le tableau.
8. Pour supprimer une zone, sélectionnez la ligne dans le tableau puis cliquez sur le bouton de suppression.
9. Cliquez sur **OK** pour appliquer vos changements et fermer la boîte de dialogue.

Remarque : Vous pouvez également supprimer et renommer les zones directement dans le graphique en cliquant avec le bouton droit de la souris sur la ligne de la zone et en choisissant l'option souhaitée dans les menus contextuels.

Présentation des éléments marqués

Vous pouvez marquer des éléments, comme des barres, des secteurs et des points de n'importe quel graphique. Il n'est pas possible de marquer les lignes, les zones et les aires dans les graphiques autres que Tracé horaire, Courbes et Evaluation, car dans ces cas, les lignes renvoient à des champs. Chaque fois que vous marquez un élément, vous mettez avant tout en évidence toutes les données représentées par cet élément. Dans tout graphique où la même observation est représentée en plusieurs endroits (matrice SPLOM par exemple), le marquage est synonyme de brossage. Vous pouvez marquer des éléments figurant dans des graphiques, y compris dans des bandes et des zones. Chaque fois que vous marquez un élément, puis retournez au mode d'édition, le marquage reste visible.

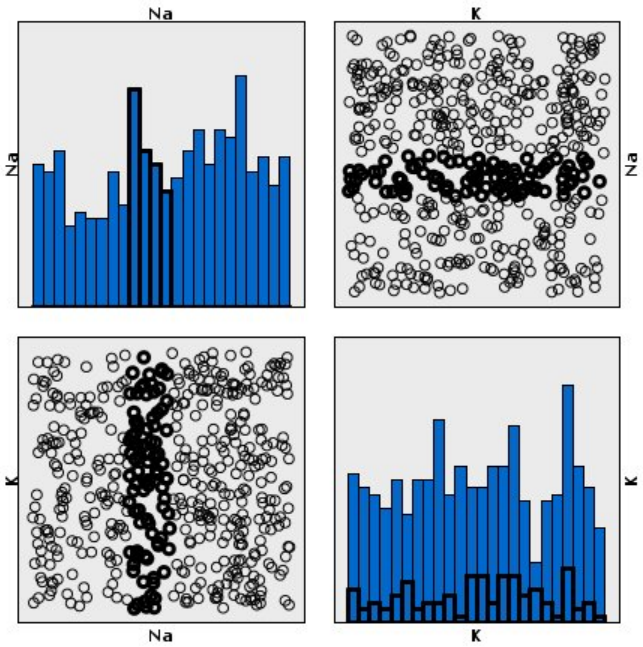


Figure 69. Marquage d'éléments dans une matrice SPLOM

Vous pouvez marquer et annuler le marquage d'éléments en cliquant sur les éléments dans le graphique. Lorsque vous cliquez pour la première fois sur un élément pour le marquer, l'élément apparaît avec une couleur de bordure épaisse pour indiquer qu'il a été marqué. Si vous cliquez à nouveau sur l'élément, la bordure disparaît et l'élément n'est plus marqué. Pour marquer plusieurs éléments, vous pouvez maintenir enfoncée la touche Ctrl tout en cliquant sur les éléments, ou vous pouvez faire glisser la souris autour de chacun des éléments que vous souhaitez marquer à l'aide de la "baguette magique". Souvenez-vous que si vous cliquez sur une autre zone ou sur un autre élément tout en maintenant enfoncée la touche Ctrl, tous les éléments déjà marqués sont désélectionnés.

Vous pouvez générer des noeuds Sélectionner et Calculer à partir des éléments marqués dans votre graphique. Pour plus d'informations, voir «Génération de noeuds à partir de graphiques».

Pour marquer des éléments :

1. Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez **Affichage > Mode d'interaction**.
2. Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Marquer des éléments.
3. Cliquez sur l'élément dont vous avez besoin ou cliquez et faites glisser votre souris pour dessiner une ligne autour de la région contenant plusieurs éléments.

Génération de noeuds à partir de graphiques

L'une des options les plus puissantes offertes par les graphiques IBM SPSS Modeler est la possibilité de générer des noeuds à partir d'un graphique ou d'une sélection dans le graphique. Vous pouvez ainsi, dans un graphique de tracé horaire, générer des noeuds Calculer et Sélectionner en fonction d'une sélection ou d'une zone de données, ce qui entraîne la définition de sous-jeux de données. Par exemple, vous pouvez utiliser la puissance de cette fonction pour identifier et exclure les valeurs éloignées.

Chaque fois que vous dessinez une bande, vous pouvez également générer un noeud Calculer. Dans les graphiques à deux axes d'échelle, vous pouvez générer des noeuds Calculer ou Sélectionner à partir des zones dessinées dans votre graphique. Dans les graphiques contenant des éléments marqués, vous pouvez générer des noeuds Calculer, des noeuds Sélectionner, et dans certains cas des noeuds Filtrer à partir de ces éléments. La génération de noeuds Equilibrer est activée pour tout graphique représentant une distribution de nombres.

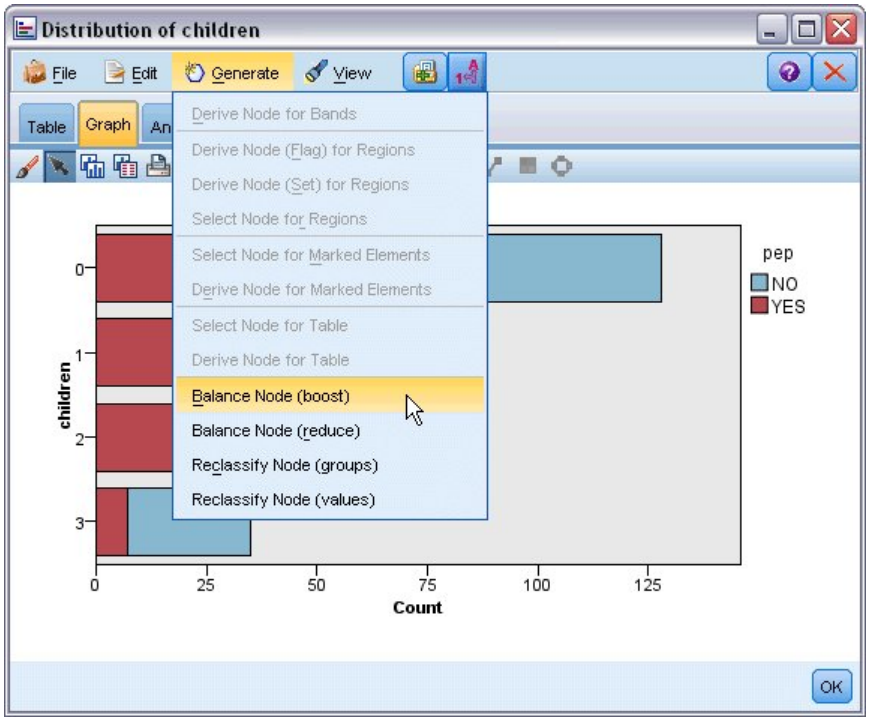


Figure 70. Graphique avec le menu Générer affichant

Chaque fois que vous générez un noeud, il est placé directement sur l'espace de travail de flux afin que vous puissiez le connecter à un flux existant. Les noeuds suivants peuvent être générés à partir de graphiques : Sélectionner, Calculer, Equilibrer, Filtrer et Recoder.

Noeuds Sélectionner

Les noeuds Sélectionner peuvent être générés pour tester l'inclusion d'enregistrements dans une zone et l'exclusion de tous les enregistrements non compris dans la zone, ou l'inverse dans le cas d'un traitement en aval.

- **Pour les bandes**. Vous pouvez générer un noeud Sélectionner qui inclut ou exclut les enregistrements compris dans cette bande. Le **noeud Sélectionner pour les bandes** est disponible uniquement via les menus contextuels car vous devez sélectionner la bande à utiliser dans le noeud Sélectionner.
- **Pour les zones**. Vous pouvez générer un noeud Sélectionner qui inclut ou exclut les enregistrements compris dans une zone.
- **Pour les éléments marqués**. Vous pouvez générer des noeuds Sélectionner pour capturer les enregistrements correspondant aux éléments marqués ou aux liens du graphique Relations.

Noeuds Calculer

Les noeuds Calculer peuvent être générés à partir de zones, de bandes et d'éléments marqués. Tous les graphiques peuvent produire des noeuds Calculer. Dans le cas des graphiques d'évaluation, une boîte de dialogue de sélection du modèle s'affiche. Dans le cas des graphiques Relations, le **Noeud dériver (“Et”)** et le **Noeud dériver (“Ou”)** sont possibles.

- **Pour les bandes**. Vous pouvez générer un noeud Calculer qui produit une catégorie pour chaque intervalle marqué sur l'axe, à l'aide des noms de bande répertoriés en tant que noms de catégorie dans la boîte de dialogue Modifier les bandes.
- **Pour les zones**. Vous pouvez générer un noeud Calculer (**Calculer en tant que booléen**) qui crée un champ booléen appelé *Dans_zone*, les booléens étant définis sur *T* pour les enregistrements inclus dans une zone et sur *F* pour les autres enregistrements. Vous pouvez également générer un noeud Calculer

(**Calculer en tant qu'ensemble**) qui produit un ensemble avec une valeur pour chaque zone et un nouveau champ appelé *zone* pour chaque enregistrement, qui prend comme valeur le nom de la zone dans laquelle est compris l'enregistrement. Les enregistrements qui ne sont compris dans aucune zone reçoivent le nom de la zone par défaut. Les noms des valeur deviennent les noms des zone répertoriés dans la boîte de dialogue Modifier les zones.

- **Pour les éléments marqués**. Vous pouvez générer un noeud Calculer qui calcule un booléen dont la valeur est *True (vrai)* pour tous les éléments marqués et *False (faux)* pour tous les autres enregistrements.

Noeuds Equilibrer

Les noeuds Equilibrer peuvent être générés pour corriger des déséquilibres dans les données. Il est par exemple possible de réduire la fréquence des valeurs courantes (utilisez l'option de menu **Noeud Equilibrer (réduire)**) ou d'augmenter l'occurrence des valeurs sous-représentées (utilisez l'option de menu **Noeud Equilibrer (augmenter)**). La génération de noeuds Equilibrer est activée pour tout graphique représentant une distribution de nombres, tel que Histogramme, Points, Résumé, Barre de nombres, Secteurs de nombres, et Courbes .

Noeuds Filtrer

Les noeuds Filtrer peuvent être générés pour renommer ou filtrer les champs en fonction de lignes ou de noeuds marqués dans le graphique. Dans le cas de graphiques d'évaluation, la ligne la plus appropriée ne génère pas de noeud Filtrer.

Noeuds Recoder

Les noeuds Recoder peuvent être générés pour recoder des valeurs. Cette option est utilisée pour les graphiques de distribution. Vous pouvez générer un noeud Recoder pour des **groupes**, afin de recoder des valeurs spécifiques d'un champ affiché en fonction de leur appartenance à un groupe (sélectionnez les groupes à l'aide de la combinaison Ctrl+clic dans l'onglet **Tables**). Vous pouvez également générer un noeud Recoder pour des **valeurs**, afin de recoder les données d'un ensemble de valeurs existant. Il peut s'agir par exemple de recoder les données d'un ensemble de valeurs standard afin de fusionner les données financières issues de différentes sociétés en vue de leur analyse.

Remarque : Si ces valeurs sont prédéfinies, vous pouvez les lire dans IBM SPSS Modeler en tant que fichier plat et utiliser une distribution pour toutes les afficher. Ensuite, créez directement à partir du graphique un noeud (de valeurs) Recoder pour le champ. Cette opération place l'ensemble des valeurs cible dans la colonne (liste déroulante) *Nouvelles valeurs* du noeud Recoder.

Lors de la définition des options pour le noeud Recoder, la table active un mappage clair des anciennes valeurs définies vers les nouvelles valeurs que vous spécifiez :

- **Valeur d'origine.** Cette colonne répertorie les valeurs existantes des champs sélectionnés.
- **Nouvelle valeur.** Utilisez cette colonne pour saisir les nouvelles valeurs de catégorie ou sélectionnez-en une dans la liste déroulante. Lorsque vous générez automatiquement un noeud Recoder avec les valeurs provenant d'un graphique Proportion, ces valeurs sont incluses dans la liste déroulante. Cela vous permet de mapper les valeurs existantes rapidement avec un ensemble de valeurs connu. Par exemple, les organisations de santé regroupent parfois les diagnostics différemment selon le réseau et les paramètres régionaux. Après une fusion ou un rachat, toutes les parties doivent recoder les données nouvelles ou même existantes de manière homogène. Au lieu d'attribuer un type manuellement à chaque cible à partir d'une longue liste, vous pouvez lire la principale liste des valeurs dans IBM SPSS Modeler, exécuter un graphique Proportion pour le champ *Diagnostic*, et générer un noeud Recoder (valeurs) pour ce champ directement à partir du graphique. Ce processus rend toutes les valeurs Diagnostic cible disponibles à partir de la liste déroulante Nouvelles valeurs.

Pour plus d'informations sur le noeud Recoder, voir «Paramétrage des options du noeud Recoder», à la page 171.

Génération de noeuds à partir de graphiques

Vous pouvez utiliser le menu Générer de la fenêtre de sortie du graphique pour générer des noeuds. Le noeud généré est placé dans l'espace de travail de flux. Pour utiliser ce noeud, connectez-le à un flux existant.

Pour générer un noeud à partir d'un graphique :

1. Vérifiez que vous êtes en mode d'interaction. Dans les menus, choisissez **Affichage > Mode d'interaction**.
2. Dans la barre d'outils du mode d'interaction, cliquez sur le bouton Zone.
3. Définissez les bandes, les zones ou tous les éléments marqués nécessaires pour générer votre noeud.
4. Dans le menu Générer, choisissez le type de noeud que vous souhaitez produire. Seuls les noeuds possibles sont activés.

Remarque : Vous pouvez également générer des noeuds directement à partir du graphique en cliquant avec le bouton droit de la souris et en choisissant l'option souhaitée dans les menus contextuels.

Modification des visualisations

Tandis que le mode d'interaction vous permet d'explorer de manière analytique les données et les valeurs représentées par la visualisation, le mode d'édition vous permet de modifier la mise en forme et l'aspect de la visualisation. Par exemple, vous pouvez modifier la police et les couleurs en fonction du guide de style de votre entreprise. Pour sélectionner ce mode, choisissez **Affichage > Mode d'édition** dans les menus (ou cliquez sur l'icône de la barre d'outils).

En mode d'édition, plusieurs barres d'outils ont un impact sur différents aspects de la présentation de la visualisation. Si certaines ne vous sont pas utiles, vous pouvez les masquer pour augmenter l'espace de la boîte de dialogue dans laquelle apparaît le graphique. Pour sélectionner ou désélectionner des barres d'outils, cliquez sur le nom de la barre d'outils souhaitée dans le menu Vue.

Remarque : Pour ajouter des détails supplémentaires à vos visualisations, vous pouvez appliquer un titre, des notes de bas de page et des libellés d'axes. Pour plus d'informations, voir «Ajout de titres et de notes de bas de page», à la page 312.

Plusieurs options sont disponibles pour modifier une visualisation en **mode d'édition**. Vous pouvez :

- Modifier le texte et le mettre en forme.
- Modifier la couleur de remplissage, la transparence et le motif des cadres et des éléments graphiques.
- Modifier la couleur et le tracé des bordures et des lignes.
- Faire pivoter et modifier la forme et le rapport d'aspect des points.
- Modifier la taille des éléments graphiques (par exemple, les barres et les points).
- Ajuster l'espace entourant les éléments à l'aide des marges et via l'extension
- Spécifier le format des numéros.
- Modifier les paramètres des axes et de l'échelle.
- Trier, exclure et réduire les catégories sur un axe catégoriel.
- Définir l'orientation des panneaux.
- Appliquer des transformations à un système de coordonnées.
- Modifier les statistiques, les types d'élément graphique et les modificateurs de collision.
- Modifier la position de la légende.

- Appliquer des feuilles de style de visualisation.

Les rubriques suivantes expliquent comment effectuer ces tâches. Nous vous recommandons également de lire les règles générales sur la modification de graphiques.

Comment basculer vers le mode d'édition

A partir des menus, sélectionnez :

Vue > **Mode Edition**

Règles générales d'édition de visualisations

Mode d'édition

Toutes les modifications s'effectuent dans le mode d'édition. Pour activer le mode d'édition, sélectionnez dans les menus :

Vue > **Mode Edition**

Sélection

Les options d'édition disponibles dépendent de la sélection effectuée. Différentes options de barre d'outils et de palette de propriétés sont activées selon les éléments sélectionnés. Seules les options activées s'appliquent à la sélection en cours. Par exemple, si un axe est sélectionné, les onglets Echelle, Graduations principales et Graduations secondaires sont disponibles dans la palette des propriétés.

Voici quelques conseils pour sélectionner des éléments dans une visualisation :

- Cliquez sur un élément pour le sélectionner.
- Sélectionnez un élément graphique (par exemple, des points dans un diagramme de dispersion ou des barres dans un diagramme à barres), en cliquant une seule fois. Après une première sélection, cliquez de nouveau pour réduire la sélection à des groupes d'éléments graphiques ou à un seul élément graphique.
- Appuyez sur la touche Echap pour désélectionner tous les éléments.

Palettes

Lorsqu'un élément est sélectionné dans la visualisation, les différentes palettes sont mises à jour pour refléter la sélection. Les palettes contiennent des commandes pour apporter des modifications à la sélection. Les palettes peuvent être des barres d'outils ou un panneau avec plusieurs commandes et onglets. Les palettes peuvent être masquées, pour garantir que la palette nécessaire est affichée pour effectuer des modifications. Consultez le menu Affichage pour savoir quelles palettes sont actuellement affichées.

Vous pouvez repositionner les palettes en cliquant et déplaçant l'espace vide dans une palette barre d'outils ou sur le côté gauche d'autres palettes. Le contrôle visuel vous indique où vous pouvez ancrer la palette. Pour les palettes qui ne sont pas des barres d'outils, vous pouvez également cliquer sur le bouton fermer pour masquer la palette et sur le bouton détacher pour afficher la palette dans une fenêtre séparée. Cliquez sur le bouton aide pour afficher de l'aide concernant une palette en particulier.

Paramètres automatiques

Certains paramètres proposent une option **-auto-**. Cette option indique que des valeurs automatiques sont appliquées. Les paramètres automatiques utilisés dépendent de la visualisation et des valeurs de données

spécifiques. Vous pouvez entrer une valeur pour remplacer le paramètre automatique. Pour restaurer le paramètre automatique, supprimez la valeur actuelle et appuyez sur Entrée. Le paramètre **-auto-** s'affiche de nouveau.

Suppression/Masquage des éléments

Vous pouvez supprimer/masquer divers éléments dans une visualisation. Par exemple, vous pouvez masquer la légende ou le libellé d'axe. Pour supprimer un élément, sélectionnez-le et appuyez sur Supprimer. Si la suppression de l'élément n'est pas autorisée, rien ne se produit. Si vous supprimez un élément par erreur, appuyez sur Ctrl+Z pour annuler la suppression.

Etat

Certaines barres d'outils reflètent l'état de la sélection actuelle, contrairement à d'autres. La palette des propriétés reflète toujours cet état. Si une barre d'outils ne reflète *pas* l'état de la sélection, la rubrique décrivant cette barre d'outils l'indique.

Edition et formatage de texte

Vous pouvez éditer le texte en place et changer le formatage d'un bloc de texte entier. Veuillez noter que vous ne pouvez pas modifier du texte directement lié à des valeurs de données. Par exemple, vous ne pouvez pas éditer un libellé de graduation car le contenu du libellé provient des données sous-jacentes. Mais vous pouvez formater n'importe quel texte de la visualisation.

Modifier le texte

1. Double-cliquez sur le bloc de texte. Cette action sélectionne tout le texte. Toutes les barres d'outils sont alors désactivées, car vous ne pouvez modifier aucune autre partie de la visualisation pendant la modification du texte.
2. Tapez le nouveau texte pour remplacer le texte existant. Vous pouvez également cliquer de nouveau sur le texte pour afficher un curseur. Placez le curseur à la position souhaitée et entrez le texte supplémentaire.

Formatage du texte

1. Sélectionnez le cadre qui contient le texte. Ne double-cliquez pas sur le texte.
2. Mettez le texte en forme à l'aide de la barre d'outils des polices. Si la barre d'outils n'est pas activée, vérifiez que seul le *cadre* contenant le texte est sélectionné. Si le texte lui-même est sélectionné, la barre d'outils est désactivée.

Vous pouvez modifier la police :

- Couleur
- Famille (par exemple, Arial ou Verdana)
- Taille (l'unité utilisée est le point, sauf si vous indiquez une unité différente telle que le pica, pc)
- Pondération
- Alignement par rapport au cadre du texte

Le formatage s'applique à tout le texte figurant dans le cadre. Vous ne pouvez pas changer le formatage de certaines lettres ou de certains mots dans un bloc de texte spécifique.

Modification des couleurs, des motifs, des pointillés et de la transparence

Dans une visualisation, de nombreux éléments différents comportent un remplissage et une bordure. L'exemple le plus évident est celui d'une barre dans un diagramme à barres. La couleur des barres est la couleur de remplissage. Les barres peuvent également être entourées d'une bordure unie noire.

D'autres éléments moins évidents d'une visualisation comportent également des couleurs de remplissage. Si la couleur de remplissage est transparente, le remplissage n'est pas nécessairement visible. Par exemple, supposons que le texte se trouve dans un libellé d'axe. Ce texte semble "flotter" mais figure en fait dans un cadre comportant une couleur de remplissage transparente. Vous pouvez voir le cadre en sélectionnant le libellé d'axe.

Tout cadre dans la visualisation peut avoir un style de remplissage et de bordure, y compris le cadre autour de la visualisation entière. De plus, n'importe quel remplissage possède un niveau d'opacité/transparence associé qui peut être ajusté.

Modifier les couleurs, les motifs, les pointillés et la transparence

1. Sélectionnez l'élément à formater. Par exemple, sélectionnez les barres d'un diagramme à barres ou un cadre contenant du texte. Si la visualisation est fractionnée par une variable ou un champ catégoriel, vous pouvez également sélectionner le groupe correspondant à une catégorie individuelle. Vous pouvez ainsi changer l'apparence par défaut attribuée à ce groupe. Par exemple, vous pouvez modifier la couleur d'un des groupes d'empilement dans un graphique à barres⌂ empilés.
2. Pour changer la couleur de remplissage, la couleur de bordure ou le motif de remplissage, utilisez la barre d'outils des couleurs.

   Remarque : Cette barre d'outils ne reflète pas l'état de la sélection en cours.

   Pour changer une couleur ou un remplissage, vous pouvez cliquer sur le bouton pour sélectionner l'option affichée ou sur la flèche de la liste déroulante pour sélectionner une autre option. Pour les couleurs, il existe une couleur paraissant blanche et traversée d'une ligne diagonale rouge. C'est la couleur transparente. Cette couleur peut être utilisée par exemple pour masquer les bordures des barres dans un histogramme.

   - Le premier bouton contrôle la couleur de remplissage. Si la couleur est associée à un champ continu ou ordinal, ce bouton change la couleur de remplissage par la couleur associée à la valeur la plus élevée dans les données. Vous pouvez utiliser l'onglet Couleur sur la palette des propriétés pour modifier la couleur associée à la valeur la plus faible et aux données manquantes. La couleur des éléments changera de manière incrémentale de la couleur Faible à la couleur Elevée, au fur et à mesure que les valeurs des données sous-jacentes augmentent.
   - Le deuxième bouton contrôle la couleur de bordure.
   - Le troisième bouton contrôle le motif de remplissage. Le motif de remplissage utilise la couleur des bordures. Par conséquent, le motif de remplissage n'est visible que s'il existe une couleur de bordure visible.
   - La quatrième commande est composée d'un curseur de défilement et d'une zone de texte qui contrôlent l'opacité de la couleur et du motif de remplissage. Un pourcentage peu élevé signifie moins d'opacité et plus de transparence. 100 % indique une couleur complètement opaque (aucune transparence).
3. Pour changer les tirets d'une bordure ou d'une ligne, utilisez la barre d'outils de ligne.

   Remarque : Cette barre d'outils ne reflète pas l'état de la sélection en cours.

   De même qu'avec l'autre barre d'outils, vous pouvez cliquer sur le bouton pour sélectionner l'option affichée ou cliquer sur la flèche du menu déroulant pour choisir une autre option.

Changement de la forme et du rapport d'aspect des points et rotation des points

Vous pouvez faire pivoter des points, attribuer une forme prédéfinie différente ou changer le rapport d'aspect (le rapport entre la largeur et la hauteur).

Modifier les points

1. Sélectionnez les points. Vous ne pouvez pas changer la forme et le rapport d'aspect de points individuels ni faire pivoter ces points.
2. Utilisez la barre d'outils des symboles pour modifier les points.

- Le premier bouton permet de changer la forme des points. Cliquez sur la flèche de la liste déroulante et sélectionnez une forme prédéfinie.
- Le deuxième bouton vous permet de faire pivoter les points sur une position de compas précise. Cliquez sur la flèche de la liste déroulante, puis faites glisser l'aiguille vers la position désirée.
- Le troisième bouton permet de changer le rapport d'aspect. Cliquez sur la flèche de la liste déroulante, puis cliquez sur le rectangle qui apparaît et faites-le glisser. La forme du rectangle représente le rapport d'aspect.

Changement de la taille des éléments graphiques

Vous pouvez modifier la taille des éléments graphiques dans la visualisation. Ces éléments sont notamment les barres, les lignes et les points. Si la taille de l'élément graphique est déterminée par une variable ou un champ, la taille spécifiée est la taille *minimale*.

Comment changer la taille des éléments graphiques

1. Sélectionnez les éléments graphiques à redimensionner.
2. Utilisez le curseur ou entrez une taille spécifique pour l'option disponible dans la barre d'outils des symboles. L'unité utilisée est le pixel, sauf si vous indiquez une unité différente (vous trouverez ci-dessous une liste complète des abréviations d'unité). Vous pouvez également spécifier un pourcentage (par exemple, 30 %), ce qui signifie qu'un élément graphique utilise le pourcentage d'espace disponible spécifié. L'espace disponible dépend du type de l'élément graphique et de la visualisation.

Tableau 39. Abréviations d'unités valides

Abréviation	**Unité**
cm	centimètre
dans	pouce
mm	millimètre
pc	pica
pt	point
px	pixel

Spécification des marges et du remplissage

S'il y a trop ou pas assez d'espace autour ou à l'intérieur d'un cadre dans la visualisation, vous pouvez modifier ses paramètres de marge et de remplissage. La **marge** est la quantité d'espace séparant le cadre des autres éléments situés autour de ce cadre. Le **remplissage** est la quantité d'espace située entre la bordure et le *contenu* du cadre.

Comment spécifier les marges et le remplissage

1. Sélectionnez le cadre pour lequel vous souhaitez spécifier des marges et un remplissage. Il peut s'agir d'un cadre de texte, d'un cadre entourant une légende ou même d'un cadre de données affichant des éléments graphiques (par exemple, des barres et des points).
2. Utilisez l'onglet Marges de la palette des propriétés pour spécifier les paramètres. Toutes les tailles sont exprimées en pixels, sauf si vous indiquez une unité différente (par exemple, le centimètre cm ou le pouce in).

Formatage des nombres

Vous pouvez spécifier le format des nombres figurant dans les libellés de graduation sur un axe continu ou dans les libellés de valeurs de données affichant un nombre. Par exemple, vous souhaitez spécifier que les nombres affichés sur les libellés de graduation soient en milliers.

Pour spécifier les formats de nombre

1. Sélectionnez les libellés de graduation de l'axe continu ou les libellés de valeur de données si elles comportent des nombres.
2. Cliquez sur l'onglet **Format** dans la palette des propriétés.
3. Sélectionnez les options de formatage des nombres souhaitées :

   Préfixe. Un caractère à afficher devant le nombre. Par exemple, saisissez le symbole ($) si les nombres correspondent à des salaires en dollars U.S.

   Suffixe. Caractère à afficher derrière le nombre. Par exemple, saisissez le symbole du pourcentage (%) si les nombres sont des pourcentages.

   Chiffres entiers min.. Nombre de chiffres minimum à afficher dans la partie entière d'une représentation décimale. Si la valeur réelle ne contient pas le nombre minimum de chiffres, la partie entière de la valeur sera remplie de zéros.

   Chiffres entiers min.. Nombre de chiffres maximum à afficher dans la partie entière d'une représentation décimale. Si la valeur réelle dépasse le nombre de chiffres maximal, la partie entière de cette valeur sera remplacée par des astérisques.

   Chiffres décimaux min. Nombre minimum de chiffres à afficher dans la partie décimale d'une représentation décimale ou scientifique. Si la valeur réelle ne contient pas le nombre de chiffres minimum, la partie décimale de cette valeur sera complétée par des zéros.

   Chiffres décimaux max. Nombre maximum de chiffres à afficher dans la partie décimale d'une représentation décimale ou scientifique. Si la valeur réelle dépasse le nombre de chiffres maximal, la décimale est arrondie au nombre de chiffres approprié.

   Scientifique. Afficher les chiffres sous forme de notation scientifique. Cette notation est utile pour des nombres très grands ou très petits. **-auto-** permet à l'application de choisir si la notation scientifique est appropriée.

   Mise à l'échelle. Facteur d'échelle, représentant un numéro par lequel la valeur originale est divisée. Utilisez un facteur d'échelle si vous souhaitez que le libellé ne s'étende pas trop pour s'ajuster aux nombres élevés. Si vous modifiez le format des nombres des libellés de graduation, veillez à modifier le titre de l'axe pour indiquer comment les nombres doivent être interprétés. Par exemple, si votre axe d'échelle affiche des salaires et que les libellés sont 30 000, 50 000 et 70 000, vous pouvez saisir un facteur d'échelle de `1 000` pour afficher 30, 50 et 70. Vous devez ensuite modifier l'axe d'échelle pour inclure le texte en milliers.

   Parenthèses pour -ve. Si des parenthèses doivent entourer les valeurs négatives.

   Regroupement. Afficher un caractère entre les groupes de chiffres. Les paramètres régionaux actuels de votre ordinateur déterminent le caractère utilisé pour le regroupement des chiffres.

Changement des paramètres d'axe et d'échelle

Plusieurs options permettent de changer les axes et les échelles.

Comment changer les paramètres relatifs aux et aux échelles

1. Sélectionnez une partie de l'axe (par exemple, le libellé d'axe ou les libellés de graduation).
2. Utilisez les onglets Echelle, Graduations principales et Graduations secondaires de la palette des propriétés pour changer les paramètres d'axe et d'échelle.

Onglet Echelle

Remarque : L'onglet Echelle n'apparaît pas pour les graphiques dans lesquels les données sont pré-agrégées (par exemple, les histogrammes).

Type. Indique si l'échelle est linéaire ou transformée. Les transformations d'échelle aident à comprendre les données ou à émettre les hypothèses nécessaires à la déduction statistique. Dans les nuages de points, vous pouvez utiliser une échelle transformée si la relation entre les variables ou les champs indépendants et dépendants n'est pas linéaire. Les transformations d'échelle permettent également de rendre un

histogramme asymétrique plus symétrique de sorte qu'il ressemble à une distribution normale. Vous ne transformez que l'échelle à laquelle les données sont affichées et non pas les données elles-mêmes.

- **linéaire.** Indique une échelle non transformée linéaire.
- **log.** Indique une échelle transformée log de base 10. Pour accommoder les valeurs nulles et négatives, cette transformation utilise une version modifiée de la fonction de log. La fonction "log valide" est définie comme `sign(x) * log(1 + abs(x))`. Par conséquent, `safeLog(-99)` est égal à :

  ```
  sign(-99) * log(1 + abs(-99)) = -1 * log(1 + 99) = -1 * 2 = -2
  ```
- **puissance.** Indique une échelle transformée de puissance, utilisant un exposant de 0,5. Pour accommoder les valeurs négatives, cette transformation utilise une version modifiée de la fonction de puissance. La fonction "puissance valide" est définie comme `sign(x) * pow(abs(x), 0,5)`. Par conséquent, `safePower(-100)` est égal à :

  ```
  sign(-100) * pow(abs(-100), 0.5) = -1* pow(100, 0.5) = -1 * 10 = -10
  ```

Minimum/Maximum/Echelle faible convenable/Echelle élevée convenable. Indique la plage de l'échelle. Le fait de sélectionner **Echelle faible convenable** et **Echelle élevée convenable** permet à l'application de sélectionner une échelle appropriée en fonction des données. Le minimum et le maximum sont convenables car ils constituent généralement des valeurs entières supérieures ou inférieures aux valeurs de données maximum et minimum. Par exemple, si la plage de données va de 4 à 92, la valeur Echelle faible convenable ou Echelle élevée convenable pour l'échelle peut être 0 ou 100 respectivement plutôt que les valeurs réelles de données maximum et minimum. Attention à ne pas définir une plage trop petit qui masquerait des éléments importants. Veuillez également noter qu'il est impossible de définir un minimum et un maximum explicites, si l'option **Inclure le zéro** est sélectionnée.

Marge faible/Marge élevée. Crée des marges en bas et/ou en haut de l'axe. La marge apparaît, perpendiculaire à l'axe sélectionné. L'unité est le pixel sauf si vous avez indiqué une autre unité (cm ou in par exemple). Par exemple, si vous définissez la **Marge haute** sur 5 pour l'axe vertical, une marge horizontale de 5 px est créée le long de la partie supérieure du cadre de données.

Inverser. Indique si l'échelle est inversée.

Inclure zéro. Indique que l'échelle doit inclure 0. Cette option est généralement utilisée pour les graphiques à barres afin de s'assurer que les barres commencent à 0 plutôt qu'à une valeur proche de la hauteur de la barre la plus petite. Si cette option est sélectionnée, les options **Minimum** et **Maximum** sont désactivées car vous ne pouvez pas définir de valeur minimum et maximum personnalisée pour l'intervalle de l'échelle.

Onglets Graduations principales/Graduations secondaires

Les **graduations** ou **marques de graduation** sont les lignes qui apparaissent sur un axe. Celles-ci indiquent des valeurs à des intervalles ou catégories spécifiques. Les **graduations principales** sont les graduations avec des libellés. Celles-ci sont également plus longues que les autres marques de graduation. Les **graduations secondaires** sont des graduations qui apparaissent entre les graduations principales. Certaines options sont spécifiques au type de graduation, mais la majorité des options est disponible pour les graduations principales et secondaires.

Afficher les graduations. Indique si les graduations principales ou secondaires apparaissent sur un graphique.

Afficher les quadrillages. Indique si le quadrillage apparaît au niveau des graduations principales ou secondaires. Les **quadrillages** sont les lignes qui traversent l'intégralité d'un graphique d'un axe à l'autre.

Position. Indique la position des marques de graduation par rapport à l'axe.

Longueur. Spécifie la longueur des marques de graduation. L'unité est le pixel sauf si vous avez indiqué une autre unité (cm ou in par exemple).

Base. *Ne s'applique qu'aux graduations principales.* Indique la valeur à laquelle la première graduation principale apparaît.

Delta. *Ne s'applique qu'aux graduations principales.* Indique la différence entre les graduations principales. C'est-à-dire que les graduations principales apparaîtront toutes les *n*èmes valeurs, où *n* est la valeur delta.

Divisions. *Ne s'applique qu'aux graduations secondaires.* Définit le nombre de divisions de graduations secondaires entre les graduations principales. Le nombre de graduations secondaires est inférieur de une unité au nombre de divisions. Par exemple, supposons qu'il existe des graduations principales à 0 et 100. Si vous entrez 2 comme nombre de divisions de graduation secondaire, il y aura *une* graduation secondaire à 50, divisant l'intervalle 0–100 et créant *deux* divisions.

Modification des modalités

Vous pouvez modifier les catégories sur un axe catégoriel de plusieurs façons :

- Changer l'ordre de tri de l'affichage des catégories.
- Excluez des catégories spécifiques.
- Ajoutez une catégorie qui n'apparaît pas dans le jeu de données.
- Fusionnez/combinez de petites modalités en une seule modalité.

Modifier l'ordre de tri des catégories

1. Sélectionnez un axe catégoriel. La palette Catégories affiche les catégories sur l'axe.

   Remarque : Si la palette n'est pas visible, vérifiez que vous l'avez bien activée. Dans IBM SPSS Modeler du menu Vue, choisissez **Catégories**.
2. Dans la palette Catégories, sélectionnez une option de tri dans la liste déroulante :

   Personnalisé. Trier les catégories en fonction de l'ordre dans lequel elles apparaissent dans la palette. Utilisez les flèches pour placer les catégories en haut ou en bas de la liste, ou les déplacer vers le haut ou vers le bas.

   Données. Trier les catégories en fonction de l'ordre dans lequel elles apparaissent dans le jeu de données.

   Nom. Trier les catégories dans l'ordre alphabétique, en utilisant les noms affichés dans la palette. Il peut s'agir de la valeur ou de libellé, selon si le bouton de la barre d'outils permettant d'afficher les valeurs ou les libellés est sélectionné ou non.

   Valeur. Triez les catégories en fonction de la valeur de données sous-jacente en utilisant les valeurs entre parenthèses de la palette. Seules les sources de données avec des métadonnées (fichier de données IBM SPSS Statistics par exemple) prennent en charge cette option.

   Statistique. Trier les catégories en fonction de la statistique calculée pour chaque catégorie. Il peut s'agir par exemple d'effectifs, de pourcentages et de moyennes. Cette option est uniquement disponible si une statistique est utilisée dans le graphique.

Ajout d'une catégorie

Par défaut, seules les catégories qui apparaissent dans le jeu de données sont disponibles. Si nécessaire, vous pouvez ajouter une modalité à la visualisation.

1. Sélectionnez un axe catégoriel. La palette Catégories affiche les catégories sur l'axe.

   Remarque : Si la palette n'est pas visible, vérifiez que vous l'avez bien activée. Dans IBM SPSS Modeler du menu Vue, choisissez **Catégories**.
2. Dans la palette Catégories, cliquez sur le bouton Ajouter une catégorie :

Figure 71. Bouton Ajouter une catégorie

3. Dans la boîte dialogue Ajouter une nouvelle modalité, saisissez le nom de la modalité.
4. Cliquez sur **OK**.

Exclure des catégories spécifiques

1. Sélectionnez un axe catégoriel. La palette Catégories affiche les catégories sur l'axe.

   Remarque : Si la palette n'est pas visible, vérifiez que vous l'avez bien activée. Dans IBM SPSS Modeler du menu Vue, choisissez **Catégories**.
2. Dans la palette Catégories, sélectionnez un nom de catégorie dans la liste Inclure, puis cliquez sur le bouton X. Pour inclure de nouveau la catégorie, sélectionnez son nom dans la liste Exclus, puis cliquez sur la flèche à droite de la liste.

Fusionner /combiner de petites catégories

Vous pouvez associer des catégories que vous n'avez pas besoin d'afficher séparément du fait de leur petite taille. Par exemple, si vous utilisez un diagramme circulaire comportant de nombreuses catégories, réduisez les catégories de 10 %. La réduction est disponible uniquement pour les statistiques basées sur une addition. Par exemple, il est impossible d'ajouter des moyennes ensemble car les moyennes ne sont pas additives. Par conséquent, la fusion/association de catégories en utilisant une moyenne n'est pas disponible.

1. Sélectionnez un axe catégoriel. La palette Catégories affiche les catégories sur l'axe.

   Remarque : Si la palette n'est pas visible, vérifiez que vous l'avez bien activée. Dans IBM SPSS Modeler du menu Vue, choisissez **Catégories**.
2. Dans la palette Catégories, sélectionnez **Réduire** et définissez un pourcentage. Toutes les catégories dont le pourcentage du total est inférieur au nombre spécifié sont réduites en une seule catégorie. Le pourcentage est basé sur la statistique présentée dans le tableau. La réduction est disponible uniquement pour les statistiques basées sur un dénombrement ou une addition (somme).

Modification de l'orientation des panels

Si vous utilisez des panneaux dans votre visualisation, vous pouvez modifier leur orientation.

Comment changer l'orientation des panneaux

1. Sélectionnez une partie de la visualisation.
2. Cliquez sur l'onglet **Panels** dans la palette des propriétés.
3. Sélectionnez une option dans **Présentation** :

Table. Dispose les panels sous forme de tableau, c'est-à-dire avec une ligne ou une colonne attribuée à chaque valeur individuelle.

Transposé. Dispose les panels sous forme de tableau, mais inverse également les lignes et les colonnes d'origine. Cette option n'est pas la même que transposer le graphique lui-même. Veuillez noter que l'axe *x* et l'axe *y* ne sont pas modifiés lorsque vous sélectionnez cette option.

Liste. Dispose les panels sous forme de liste, c'est-à-dire que chaque cellule représente une combinaison de valeurs. Les colonnes et les lignes ne sont plus assignées à des valeurs individuelles. Cette options réorganise les panels si nécessaire.

Transformation du système de coordonnées

De nombreuses visualisations sont affichées dans un système de coordonnées cartésiennes orthogonales. Vous pouvez transformer le système de coordonnées si nécessaire. Vous pouvez par exemple appliquer une transformation polaire au système de coordonnées, ajouter des effets d'ombrage oblique, et transposer les axes. Vous pouvez également annuler toutes ces transformations si elles ont déjà été appliquées à la visualisation actuelle. Par exemple, un diagramme en secteurs est placé dans un système de coordonnées polaires. Vous pouvez annuler la transformation polaire et afficher le graphique circulaire sous la forme d'un seul graphique à barres empilées dans un système de coordonnées rectangulaire.

Transformer le système de coordonnées

1. Sélectionnez le système de coordonnées à transformer. Vous sélectionnez le système de coordonnées en sélectionnant le cadre autour du graphique individuel.
2. Cliquez sur l'onglet **Coordonnées** dans la palette des propriétés.
3. Sélectionnez les transformations que vous voulez appliquer au système de coordonnées. Vous pouvez également désélectionner une transformation pour l'annuler.

   Transposé. L'opération consistant à changer l'orientation des axes est nommée **transposition**. Cette opération est comparable à la permutation des axes vertical et horizontal dans une visualisation en 2D.

   Polaire. Une transformation polaire dessine les éléments graphiques à un angle et une distance spécifiques du centre du diagramme. Un graphique circulaire est une visualisation en 1-D avec une transformation polaire qui dessine les barres à des angles spécifiques. Un graphique radar est une visualisation 2-D avec une transformation polaire qui dessine des éléments graphiques à un angle et une distance spécifiques du centre du graphique. Une visualisation 3-D inclurait également une dimension de profondeur supplémentaire.

   Oblique. Une transformation oblique ajoute un effet 3-D aux éléments graphiques. Cette transformation ajoute la profondeur aux éléments graphiques, mais la profondeur est purement décorative. Elle n'est influencée par aucune valeur de donnée particulière.

   Même rapport. Appliquer le même rapport indique que la même distance sur chaque échelle représente la même différence entre les valeurs de données. Par exemple, 2 cm sur les deux échelles représentent une différence de 1000.

   % de marge avant transformation. Si les axes sont tronqués après la transformation, il est conseillé d'ajouter des marges au graphique avant d'effectuer la transformation. Les insertions réduisent les dimensions d'un certain pourcentage avant que les transformations ne soient appliquées au système de coordonnées. Vous disposez d'un contrôle sur les dimensions *x* inférieure, *x* supérieure, *y* inférieure, et *y* supérieure, dans cet ordre.

   % de marge après transformation. Si vous souhaitez modifier le rapport hauteur/largeur du graphique, vous pouvez y ajouter des marges après avoir effectué la transformation. Les insertions réduisent les dimensions d'un certain pourcentage après l'application de transformations au système de coordonnées. Ces marges peuvent également être appliquées même si aucune transformation n'est effectuée sur le graphique. Vous disposez d'un contrôle sur les dimensions *x* inférieure, *x* supérieure, *y* inférieure, et *y* supérieure, dans cet ordre.

Changement des statistiques et des éléments graphiques

Vous pouvez convertir un élément graphique en un autre type, modifier la statistique utilisée pour dessiner l'élément graphique ou spécifier le modificateur de collision qui détermine ce qui se passe lorsque les éléments graphiques se chevauchent.

Convertir un élément graphique

1. Sélectionnez l'élément graphique à convertir.
2. Cliquez sur l'onglet **Elément** dans la palette des propriétés.
3. Sélectionnez un nouveau type d'élément graphique dans la liste Type.

Tableau 40. Types d'élément graphique

Type d'élément graphique	Description
Point	Un marqueur identifiant un point de données spécifique. Un point est utilisé dans des diagrammes de dispersion et d'autres visualisations associées.
95 %	Une forme rectangulaire tracée à une valeur de données spécifique et remplissant l'espace entre un valeur de donnée initiale et une autre. Un élément d'intervalle est utilisé dans les graphiques à barres et les histogrammes.
Courbes	Une ligne qui connecte les valeurs de données.
Chemin d'accès	Une ligne qui relie les valeurs de données dans leur ordre d'apparition dans le jeu de données.
Aire	Une ligne qui connecte les éléments de données avec la zone entre la ligne et une origine remplie.
Polygone	Une forme à plusieurs côtés comprenant une région de données. Un polygone peut être utilisé dans un diagramme de dispersion mis en intervalles ou une carte.
Schéma	Un élément constitué d'une case de moustaches et de marques indiquant les valeurs extrêmes. Un élément de schéma est utilisé pour les boîtes à moustaches.

Modifier la statistique

1. Sélectionnez l'élément graphique dont vous voulez modifier les statistiques.
2. Cliquez sur l'onglet **Elément** dans la palette des propriétés.

Comment spécifier le modificateur de collision

Le modificateur de collision détermine ce qui se passe lorsque des éléments graphiques se chevauchent.

1. Sélectionnez l'élément graphique dont vous souhaitez spécifier le modificateur de collision.
2. Cliquez sur l'onglet **Elément** dans la palette des propriétés.
3. Dans la liste déroulante Modificateur, sélectionnez un modificateur de collision. **-auto-** permet à l'application de déterminer quel modificateur de collision est adapté au type de l'élément graphique et à la statistique.

   Superposer. Représenter les éléments graphiques les uns sur les autres lorsqu'ils ont la même valeur.

   Pile. Empiler les éléments graphiques qui devraient normalement être superposés lorsqu'ils ont les mêmes valeurs de données.

   Dodge (regroupement). Déplace les éléments graphiques près d'autres éléments graphiques qui ont la même valeur, plutôt que de les superposer. Les éléments graphiques sont disposés symétriquement. C'est-à-dire que les éléments graphiques sont déplacés sur les côtés opposés d'une position centrale. Le dodging est très similaire au clustering (regroupement en clusters).

   Pile. Déplace les éléments graphiques près d'autres éléments graphiques qui ont la même valeur, plutôt que de les superposer. Les éléments graphiques sont arrangés de manière asymétrique. Autrement dit, les éléments graphiques sont empilés les uns sur les autres, l'élément graphique du bas étant positionné à une valeur spécifique sur l'échelle.

   Brouillage (normal). Repositionne de façon aléatoire les éléments graphiques à la même valeur de données au moyen d'une distribution normale.

   Brouillage (uniforme). Repositionne de manière aléatoire les éléments graphiques à la même valeur de date en utilisant une distribution uniforme.

Changement de la position de la légende

Si un graphique contient une légende, cette légende apparaît généralement à droite du graphique. Vous pouvez changer cette position si nécessaire.

Comment changer la position de la légende

1. Sélectionnez la légende.
2. Cliquez sur l'onglet **Légende** dans la palette des propriétés.
3. Sélectionnez une position.

Copie d'une visualisation et de données de visualisation

La palette Générale comprend des boutons pour copier la visualisation et ses données.

Figure 72. Bouton Copier la visualisation

Copie de la visualisation. Cette action copie la visualisation dans le presse-papiers sous forme d'image. Plusieurs formats d'image sont disponibles. Lorsque vous collez l'image dans une autre application, vous pouvez choisir une option "Collage spécial" pour sélectionner un des formats d'image disponibles pour le collage.

Figure 73. Bouton Copier les données de la visualisation

Copie des données de la visualisation. Cette action copie les données de la visualisation qui sont utilisées pour représenter la visualisation. Ces données sont copiées dans le presse-papiers comme texte normal⌂ ou comme texte au format HTML. Lorsque vous collez des données dans une autre application, vous pouvez choisir une option "Collage spécial" pour sélectionner un des formats d'image disponibles pour le collage.

Raccourcis clavier de l'éditeur de représentation graphique

Tableau 41. Raccourcis clavier

Touche de raccourci	Fonction
Ctrl+Barre d'espace	Basculer entre le mode d'exploration et le mode d'édition
Supprimer	Supprimer un élément de visualisation
Ctrl+Z	Annuler
Ctrl+Y	Rétablir
F2	Afficher la légende pour sélectionner des éléments dans le graphique

Ajout de titres et de notes de bas de page

Pour tous les types de graphique, vous pouvez ajouter un titre unique, une note de bas de page ou des libellés d'axe afin d'identifier ce que représente le graphique.

Ajout de titre aux graphiques

1. Dans les menus, sélectionnez **Edition > Ajouter un titre de graphique**. Une zone de texte contenant **<TITRE>** apparaît au-dessus du graphique.
2. Vérifiez que vous êtes en mode d'édition. Dans les menus, sélectionnez **Vue > Mode d'édition**.
3. Double-cliquez sur le texte **<TITRE>**.
4. Entrez le titre souhaité et appuyez sur Entrée.

Ajout de notes de bas de page aux graphiques

1. Dans les menus, sélectionnez **Edition > Ajouter une note de bas de page de graphique**. Une zone de texte contenant **<NOTE DE BAS DE PAGE>** apparaît sous le graphique.
2. Vérifiez que vous êtes en mode d'édition. Dans les menus, sélectionnez **Vue > Mode d'édition**.
3. Double-cliquez sur le texte **<NOTE DE BAS DE PAGE>**.
4. Entrez le titre souhaité et appuyez sur Entrée.

Utilisation de feuilles de style de graphique

Les informations de base relatives à l'affichage du graphique telles que les couleurs, polices, symboles et épaisseurs de lignes sont contrôlées par une feuille de style. Une feuille de style par défaut est fournie avec IBM SPSS Modeler ; néanmoins, vous pouvez la modifier si nécessaire. Par exemple, vous pouvez disposer d'un modèle de couleurs d'entreprise pour les présentations à utiliser dans les graphiques. Pour plus d'informations, voir «Modification des visualisations», à la page 301.

Dans les noeuds Graphiques, vous pouvez utiliser le Mode d'édition pour modifier les styles d'apparence d'un graphique. Vous pouvez ensuite utiliser le menu **Edition > Styles** pour enregistrer les changements sous forme de feuille de style qui s'applique à tous les graphiques que vous générez ensuite à partir du noeud Graphique actuel ou sous forme de nouvelle feuille de style par défaut pour tous les graphiques que vous produisez à l'aide d'IBM SPSS Modeler.

Quatre options de feuille de style sont disponibles à partir de l'option **Styles** du menu Edition :

- **Changer la feuille de style.** Cette option affiche une liste des feuilles de styles différentes stockées parmi lesquelles vous pouvez choisir afin de modifier l'apparence de vos graphiques. Pour plus d'informations, voir «Application de feuilles de style».
- **Stocker les styles dans le noeud.** Stocke les modifications apportées aux styles du graphique sélectionné de façon à ce qu'elles soient appliquées aux graphiques futurs créés à partir du même noeud Graphique, dans le flux actuel.
- **Stocker les styles comme valeur par défaut.** Stocke les modifications apportées aux styles du graphique sélectionné de façon à ce qu'elles soient appliquées aux graphiques futurs créés à partir d'un noeud Graphique, dans n'importe quel flux. Une fois cette option sélectionnée, vous pouvez utiliser l'option **Appliquer les styles par défaut** pour modifier d'autres graphiques existants afin qu'ils utilisent les mêmes styles.
- **Appliquer les styles par défaut.** Remplace les styles du graphique sélectionné par les styles actuellement enregistrés comme styles par défaut.
- **Appliquer les styles d'origine.** Rétablit les styles d'un graphique sur les styles fournis comme styles d'origine par défaut.

Application de feuilles de style

Vous pouvez appliquer une feuille de style de visualisation spécifiant les propriétés stylistiques de la visualisation. Par exemple, la feuille de style peut définir les polices, les tirets et couleurs, entre autres. Dans une certaine mesure, les feuilles de style sont un raccourci des modifications que vous auriez à effectuer manuellement. Notez néanmoins qu'une feuille de style est limitée aux modifications de *style*. Les autres changements telles que la position de la légende ou l'intervalle de l'échelle ne sont pas stockés dans la feuille de style.

Appliquer une feuille de style

1. A partir des menus, sélectionnez :
   Edition > **Styles** > **Permuter la feuille de style**
2. Utilisez la boîte de dialogue Permuter la feuille de style pour sélectionner une feuille de style.
3. Cliquez sur **Appliquer** pour appliquer la feuille de style à la visualisation sans fermer la boîte de dialogue. Cliquez sur **OK** pour appliquer la feuille de style et fermer la boîte de dialogue.

Boîte de dialogue Permuter/Sélectionner une feuille de style

Le tableau en haut de la boîte de dialogue répertorie toutes les feuilles de style de visualisation qui sont actuellement disponibles. Quelques feuilles de style sont préinstallées, tandis que d'autres ont été créées dans IBM SPSS Visualization Designer (un produit distinct).

Le bas de la boîte de dialogue montre des exemples de visualisations ainsi que des données d'échantillon. Sélectionnez l'une de ces feuilles de style afin d'en appliquer les styles aux visualisations en exemple. Ces exemples peuvent vous aider à déterminer la manière dont la feuille de style affecte votre visualisation actuelle.

La boîte de dialogue propose aussi les options suivantes.

Styles existants. Par défaut, une feuille de style peut écraser tous les styles dans la visualisation. Vous pouvez modifier ce comportement.

- **Ecraser tous les styles**. Lors de l'application de la feuille de style, remplacez tous les styles de la visualisation, y compris ceux modifiés dans la visualisation au cours de la session actuelle de modification.
- **Conserver les styles modifiés**. Lors de l'application de la feuille de style, remplacez uniquement les styles qui n'ont *pas* été modifiés dans la visualisation au cours de la session actuelle de modification. Les styles qui ont été modifiés au cours de la session actuelle de modification sont conservés.

Gérer. Gérer les modèles de visualisation, les feuilles de style et les cartes sur votre ordinateur. Vous pouvez importer, exporter, renommer et supprimer les modèles de visualisation, les feuilles de style et les cartes depuis votre ordinateur local. Pour plus d'informations, voir «Gérer les modèles, les feuilles de style et les fichiers cartes», à la page 229.

Emplacement. Modifier l'emplacement dans lequel les modèles de visualisation, les feuilles de style et les cartes sont stockés. L'emplacement actuel est indiqué à droite du bouton. Pour plus d'informations, voir la rubrique «Définition de l'emplacement des modèles, des feuilles de style et des cartes.», à la page 228.

Impression, enregistrement, copie et exportation de graphiques

Chaque graphique présente un certain nombre d'options relatives à l'enregistrement, à l'impression ou bien encore à l'exportation vers un autre format. La plupart de ces options sont disponibles dans le menu Fichier. En outre, à partir du menu Edition, vous pouvez choisir de copier le graphique, les données qu'il contient ou l'objet dessin Microsoft Office à utiliser dans une autre application.

Impression

Pour imprimer le graphique, utilisez l'option de menu ou le bouton **Imprimer**. Avant l'impression, vous pouvez utiliser les options **Mise en page** et **Aperçu avant impression** pour définir les options d'impression et prévisualiser la sortie.

Enregistrement des graphiques

Pour enregistrer le graphique dans un fichier de sortie IBM SPSS Modeler (`*.cou`), choisissez l'option **Fichier > Enregistrer** ou **Fichier > Enregistrer sous** dans les menus.

ou

Pour enregistrer le graphique dans le référentiel, choisissez **Fichier > Stocker la sortie** dans les menus.

Copie de graphiques

Pour copier le graphique en vue d'une utilisation dans une autre application, telle que MS Word ou MS PowerPoint, sélectionnez **Edition > Copier le graphique** dans les menus.

Copie des données

Pour copier les données en vue d'une utilisation dans une autre application, telle que MS Excel ou MS Word, sélectionnez **Edition > Copier les données** dans les menus. Par défaut, les données sont formatées en HTML. Utilisez **Collage spécial** dans l'autre application pour voir d'autres options de formatage lors du collage.

Copie d'un objet image Microsoft Office

Vous pouvez copier un graphique sous forme d'objet image Microsoft Office et l'utiliser dans des applications Microsoft Office, telles qu'Excel ou PowerPoint. Pour copier un graphique, choisissez **Edition > Copier l'objet image Microsoft Office** dans les menus. Le contenu est copié dans votre presse-papiers, au format binaire par défaut. Utilisez **Collage spécial** dans l'application Microsoft Office pour spécifier d'autres options de formatage lors du collage.

Notez que cette fonctionnalité n'est pas prise en charge par certains contenus, auquel cas l'option de menu **Copier l'objet image Microsoft Office** est désactivée. Notez également que le graphique, une fois collé dans une application Office, peut changer d'aspect, mais que ces données ne changent pas.

Six types de sortie graphique peuvent être copiés et collés dans Excel : Barres simples, Barres empilées, Diagramme à surfaces simple, Diagramme à surfaces groupé, Nuage de points simple et Nuage de points groupé. Si vous utilisez les options Panneau et Animation pour l'un de ces types de graphique, l'option **Copier l'objet image Microsoft Office** est désactivée dans SPSS Modeler. Pour les autres paramètres, tels qu'Apparences en option ou Superposition, cette option est partiellement prise en charge. Pour plus d'informations, reportez-vous au tableau suivant :

Tableau 42. Prise en charge de l'option Copier l'objet image Microsoft Office

Modèle de sortie graphique	Noeud de graphique Modeler	Type de graphique Modeler	Paramètre de base	Apparences en option	Superposition	Prise en charge de la copie d'objet image Microsoft	Commentaires
Barres simples	Représentation graphique	Barres	Oui	Non	N/A	Oui	
		Barres d'effectifs	Oui	Non	N/A	Oui	
	Proportion	Barres	Oui	N/A	Non	Oui	

Tableau 42. Prise en charge de l'option Copier l'objet image Microsoft Office (suite)

Modèle de sortie graphique	Noeud de graphique Modeler	Type de graphique Modeler	Paramètre de base	Apparences en option	Superposition	Prise en charge de la copie d'objet image Microsoft	Commentaires
Barres empilées	Représentation graphique	Barres	Oui	Oui	N/A	Oui avec limitation	Oui, uniquement pour la variable catégorielle dans Apparences en option.
		Barres d'effectifs	Oui	Oui	N/A	Oui avec limitation	Oui, uniquement pour la variable catégorielle dans Apparences en option.
	Proportion	Barres	Oui	N/A	Oui	Oui	
Boîte à moustaches	Représentation graphique	Boîte à moustaches	Oui	Non	N/A	Oui avec limitation	Oui, uniquement sous Windows.
		Boîte à moustaches	Oui	Oui	N/A	Non	
Diagramme à surfaces groupé	Représentation graphique	Diagramme à surfaces groupé	Oui	Non	N/A	Oui avec limitation	Oui, uniquement sous Windows.
		Diagramme à surfaces groupé	Oui	Oui	N/A	Non	

Tableau 42. Prise en charge de l'option Copier l'objet image Microsoft Office (suite)

Modèle de sortie graphique	Noeud de graphique Modeler	Type de graphique Modeler	Paramètre de base	Apparences en option	Superposition	Prise en charge de la copie d'objet image Microsoft	Commentaires
Nuage de points simple	Représentation graphique	Graphique à bulles	Oui	Non	N/A	Oui avec limitation	Uniquement Oui pour les variables continues dans les zones X et Y et la variable catégorielle dans la zone Tailles.
		Nuage de points	Oui	Non	N/A	Oui avec limitation	Oui, uniquement pour les variables continues dans les champs X et Y.
	Tracé	Point	Oui	N/A	Non	Oui avec limitation	Oui, uniquement pour les variables continues dans les champs X et Y.
Nuage de points groupé	Représentation graphique	Graphique à bulles	Oui	Oui	N/A	Non	
		Nuage de points	Oui	Oui	N/A	Oui avec limitation	Oui, uniquement pour les variables continues dans les champs X et Y et la variable catégorielle dans Apparences en option.
	Tracé	Point	Oui	N/A	Oui	Oui avec limitation	Uniquement Oui pour les variables continues dans les zones X et Y et la variable catégorielle dans la zone Superposition.

Exportation de graphiques

L'option **Exporter le graphique** vous permet d'exporter le graphique dans l'un des formats suivants : Bitmap (.bmp), JPEG (.jpg), PNG (.png), HTML (.html), PDF (.pdf) ou document ViZml (.xml) afin de les utiliser dans d'autres applications.

Remarque : si l'option PDF est sélectionnée, les graphiques sont exportés en tant que fichiers PDF à résolution élevée coupés pour tenir dans le graphique.

Pour exporter des graphiques, sélectionnez **Fichier > Exporter le graphique** dans les menus puis sélectionnez le format.

Exportation de tables

L'option **Exporter graphique** vous permet d'exporter le graphique dans l'un des formats suivants : délimité par des tabulations (.tab), délimité par des virgules (.csv), ou HTML (.html).

Pour exporter des tables, sélectionnez **Fichier > Exporter la table** dans les menus puis sélectionnez le format.

Chapitre 6. Noeuds de sortie

Présentation des noeuds de sortie

Les noeuds de sortie permettent d'obtenir des informations sur vos données et vos modèles. Ils permettent également d'exporter les données dans divers formats, afin de pouvoir les utiliser avec d'autres logiciels.

Les noeuds de sortie disponibles sont les suivants :

Le noeud Table affiche les données au format tabulaire (ces données peuvent également être écrites dans un fichier). Ainsi, vous pouvez passer en revue les valeurs de données ou les exporter dans un format facilement lisible.

Le noeud Matrice permet de créer un tableau dans lequel les relations entre les champs sont indiquées. Il s'agit généralement de deux champs symboliques, mais il peut également s'agir de champs booléens ou numériques.

Le noeud Analyse évalue la capacité des modèles prédictifs à générer des prévisions précises. Les noeuds Analyse comparent les valeurs prédites et les valeurs réelles d'un ou de plusieurs nuggets de modèle. Ils peuvent également comparer entre eux les modèles prédictifs.

Le noeud Audit données fournit un premier aperçu complet des données, notamment des statistiques récapitulatives, des histogrammes et distributions pour chaque champ, ainsi que des informations sur les valeurs éloignées, les valeurs manquantes et les valeurs extrêmes. Les résultats sont affichés dans une matrice facile à lire pouvant être triée et utilisée pour générer les noeuds de préparation des données et des graphiques grandeur nature.

Le noeud Transformation vous permet de sélectionner et de prévisualiser les résultats des transformations avant de les appliquer aux champs sélectionnés.

Le noeud Statistiques fournit des informations récapitulatives de base sur les champs numériques. Ill calcule les statistiques récapitulatives des champs individuels et des corrélations entre les champs.

Le noeud Moyennes compare les moyennes de groupes indépendants ou de paires de champs associés, afin de détecter toute différence sensible. Par exemple, vous pouvez comparer les revenus moyens avant et après l'application d'une augmentation, ou comparer les revenus des personnes ayant obtenu une augmentation avec ceux des personnes qui n'en ont pas eu.

Ce noeud permet de créer des rapports formatés contenant du texte fixe et des données, ainsi que des expressions calculées à partir de ces dernières. Le format du rapport est déterminé par des modèles texte définissant la structure du texte fixe et de la sortie de données. Vous pouvez définir un formatage de texte personnalisé en utilisant des balises HTML dans le modèle et en définissant des options dans l'onglet Sortie. Vous pouvez inclure des valeurs de données et d'autres sorties conditionnelles à l'aide des expressions CLEM du modèle.

Le noeud Valeurs globales analyse les données et calcule des valeurs récapitulatives pouvant être utilisées dans des expressions CLEM. Par exemple, vous pouvez utiliser ce noeud pour calculer les statistiques d'un champ *âge* puis utiliser la moyenne globale de ce champ *âge* dans des expressions CLEM en insérant la fonction `@GLOBAL_MEAN(age)`.

Le noeud Ajustement de simulation examine la distribution statistique des données dans chaque champ et génère (ou met à jour) un noeud Génération de simulation en affectant à chaque champ la distribution la mieux adaptée. Le noeud Génération de simulation peut ensuite être utilisé pour générer des données simulées.

Le noeud Evaluation de simulation évalue un champ cible prévu spécifique et présente des informations de distribution et de corrélation sur le champ cible.

Gestion des sorties

Le gestionnaire des sorties affiche les graphiques, les graphiques et les tableaux générés lors d'une session IBM SPSS Modeler. Vous pouvez toujours rouvrir une sortie en double-cliquant dessus dans le gestionnaire ; il est inutile de réexécuter le flux ou le noeud correspondant.

Pour afficher le gestionnaire des sorties

Ouvrez le menu Affichage et choisissez **Gestionnaires**. Cliquez sur l'onglet **Sorties**.

Dans le gestionnaire des sorties, vous pouvez effectuer les opérations suivantes :

- Afficher des objets de sortie existants, tels que des histogrammes, des graphiques Evaluation et des tableaux.
- Renommer des objets de sortie.
- Enregistrer des objets de sortie sur disque ou dans le IBM SPSS Collaboration and Deployment Services Repository (s'il est disponible).
- Ajouter des fichiers de sortie au projet actuel.
- Supprimer des objets de sortie non enregistrés de la session actuelle.
- Ouvrir les objets de sortie enregistrés ou les récupérer dans le IBM SPSS Collaboration and Deployment Services Repository (s'il est disponible).

Pour accéder à ces options, cliquez avec le bouton droit de la souris sur l'onglet Sorties.

Affichage de la sortie

La sortie à l'écran est affichée dans une fenêtre du navigateur de sortie. La fenêtre du navigateur de sortie comporte ses propres menus, lesquels vous permettent d'imprimer ou d'enregistrer la sortie, ou de l'exporter dans un autre format. Notez que les options proposées peuvent varier en fonction du type de sortie.

Impression, enregistrement et exportation de données. Pour plus d'informations, procédez comme suit :

- Pour imprimer la sortie, utilisez le bouton ou l'option de menu **Imprimer**. Avant l'impression, vous pouvez utiliser les options **Mise en page** et **Aperçu avant impression** pour définir les options d'impression et prévisualiser la sortie.
- Pour enregistrer la sortie dans un fichier de sortie IBM SPSS Modeler (`.cou`), choisissez l'option **Enregistrer** ou **Enregistrer sous** dans le menu Fichier.
- Pour enregistrer la sortie dans un autre format (texte ou HTML, par exemple), choisissez **Exporter** dans le menu Fichier. Pour plus d'informations, voir «Exportation des sorties», à la page 323.

  Notez que vous ne pouvez sélectionner ces formats que si la sortie contient des données pouvant être exportées de cette façon. Par exemple, les contenus d'un arbre de décisions peuvent être exportés sous forme de texte, mais les contenus d'un modèle K Moyennes n'ont pas de sens sous forme de texte.
- Pour enregistrer la sortie dans un référentiel partagé afin que les autres utilisateurs puissent le consulter via le IBM SPSS Collaboration and Deployment Services Deployment Portal, choisissez **Publier sur le Web** dans le menu Fichier. Notez que cette option requiert une licence distincte pour IBM SPSS Collaboration and Deployment Services.

Sélection de cellules et de colonnes. Le menu Edition contient plusieurs options permettant de sélectionner, de désélectionner et de copier des cellules et des colonnes, pour le type de sortie actuel. Voir «Sélection de cellules et de colonnes», à la page 323 pour plus d'informations.

Création de noeuds. Le menu Générer permet de créer des noeuds sur la base du contenu du navigateur de sortie. Les options varient en fonction du type de sortie et des éléments de la sortie actuellement sélectionnés. Pour plus d'informations sur les options de création de noeud pour un type de sortie donné, reportez-vous à la documentation propre à la sortie.

Publier sur le Web

La fonctionnalité Publier sur le Web vous permet de publier certains types de flux de sortie dans un IBM SPSS Collaboration and Deployment Services Repository partagé central qui forme la base de IBM SPSS Collaboration and Deployment Services. Si vous utilisez cette option, d'autres utilisateurs qui ont besoin de visualiser cette sortie peuvent le faire en utilisant un accès Internet et un compte IBM SPSS Collaboration and Deployment Services : il est inutile d'installer IBM SPSS Modeler.

Le tableau suivant répertorie les noeuds IBM SPSS Modeler qui prennent en charge la fonctionnalité Publier sur le Web. La sortie à partir de ces noeuds est stockée dans le IBM SPSS Collaboration and Deployment Services Repository dans un format d'objet de sortie (*.cou*), et peut-être affichée directement dans le IBM SPSS Collaboration and Deployment Services Deployment Portal.

D'autres types de sorties peuvent être affichés uniquement si l'application appropriée (par exemple,. IBM SPSS Modeler, pour des objets de flux) est installée sur l'ordinateur de l'utilisateur.

Tableau 43. Noeuds qui prennent en charge la fonctionnalité Publier sur le Web.

Type de noeud	Noeud
Graphes	toutes
Sortie	Tableau
	Matrice

Tableau 43. Noeuds qui prennent en charge la fonctionnalité Publier sur le Web (suite).

Type de noeud	Noeud
	Audit données
	Transformer
	Moyennes
	Analyse
	Statistiques
	Rapport (HTML)
IBM SPSS Statistics	Sortie des statistiques

Publication d'un résultat sur le Web

Pour publier un résultat sur le Web :

1. Dans un flux IBM SPSS Modeler, exécutez l'un des noeuds répertoriés dans le tableau. Un objet de sortie est alors créé (par exemple, un objet tableau , matrice ou rapport) dans une nouvelle fenêtre.
2. Dans la fenêtre des objets de sortie, sélectionnez :

   Fichier > **Publier sur le Web**

   Remarque : pour exporter de simples fichiers HTML à utiliser avec un navigateur Web standard, choisissez **Exporter** dans le menu Fichier et sélectionnez **HTML**.
3. Connectez-vous au IBM SPSS Collaboration and Deployment Services Repository

   Lorsque vous vous êtes connecté avec succès, le référentiel : boîte de dialogue Stocker apparaît et vous propose plusieurs options de stockage.
4. Lorsque vous avez choisi l'option de stockage de votre choix, cliquez sur **Stocker**.

Affichage du résultat publié sur le Web

Vous devez disposer d'un compte IBM SPSS Collaboration and Deployment Services configuré pour utiliser cette fonctionnalité. Si l'application appropriée est installée pour le type d'objet que vous souhaitez afficher (par exemple, IBM SPSS Modeler ou IBM SPSS Statistics), la sortie est affichée dans l'application même plutôt que dans le navigateur.

Pour afficher un résultat publié sur le Web :

1. Saisissez l'adresse *http://<repos_host>:<repos_port>/peb*

   dans votre navigateuroù *repos_host* et *repos_port* sont le nom d'hôte et le numéro de port de l'hôte IBM SPSS Collaboration and Deployment Services.
2. Saisissez les détails de connexion de votre compte IBM SPSS Collaboration and Deployment Services.
3. Cliquez sur **Référentiel de contenu**.
4. Recherchez ou accédez à l'objet que vous souhaitez afficher.
5. Cliquez sur le nom de l'objet. Pour certains types d'objets, tels que des graphiques, il est possible qu'il y ait un délai car l'objet est rendu dans le navigateur.

Affichage de la sortie dans un navigateur HTML

A partir de l'onglet Options avancées des nuggets de modèle Linéaire, Logistique et ACP/Facteur, vous pouvez afficher les informations affichées dans un navigateur distinct tel qu'Internet Explorer. Le format de sortie des informations est HTML ; vous pouvez alors les enregistrer et les réutiliser ailleurs, par exemple sur un réseau Intranet d'entreprise ou un site Internet.

Pour afficher les informations dans un navigateur, cliquez sur le bouton de lancement, situé sous l'icône de modèle, en haut à gauche de la boîte de dialogue de l'onglet Options avancées du nugget de modèle.

Exportation des sorties

Dans la fenêtre du navigateur de sortie, vous pouvez choisir d'exporter la sortie dans un autre format (texte ou HTML, par exemple). Les formats d'exportation varient selon le type de sortie, mais sont en général semblables aux options de type de fichier disponibles si vous sélectionnez l'option d'**enregistrement dans un fichier** dans le noeud de génération de la sortie.

Remarque : Vous ne pouvez sélectionner ces formats que si la sortie contient des données pouvant être exportées de cette façon. Par exemple, les contenus d'un arbre de décisions peuvent être exportés sous forme de texte, mais les contenus d'un modèle K Moyennes n'ont pas de sens sous forme de texte.

Pour exporter la sortie

1. Dans le navigateur de sortie, ouvrez le menu Fichier et choisissez **Exporter**. Sélectionnez ensuite le type de fichier à créer :
   - **Délimité par des tabulations (*.tab).** Cette option crée un fichier texte formaté contenant les valeurs de données. Ce style est souvent utilisé pour créer une représentation en texte brut des informations susceptibles d'être importées dans d'autres applications. Cette option est disponible pour les noeuds Table, Matrice et Moyennes.
   - **Délimité par des virgules (*.dat).** Cette option crée un fichier texte séparé par des virgules contenant les valeurs de données. Ce style est souvent utilisé pour générer rapidement un fichier de données susceptible d'être importé dans un tableur ou un autre logiciel d'analyse de données. Cette option est disponible pour les noeuds Table, Matrice et Moyennes.
   - **Format délimité par des tabulations transposé (*.tab).** Cette option est identique à l'option Délimité par des tabulations, à une exception près toutefois : les données sont transposées de façon à ce que les lignes et les colonnes représentent respectivement les champs et les enregistrements.
   - **Format délimité par des virgules transposé (*.dat).** Cette option est identique à l'option Délimité par des virgules, à une exception près : les données sont transposées de façon à ce que les lignes et les colonnes représentent respectivement les champs et les enregistrements.
   - **HTML (*.html).** Cette option permet d'écrire une sortie au format HTML dans des fichiers.

Sélection de cellules et de colonnes

Un certain nombre de noeuds, notamment les noeuds Table, Matrice et Moyennes, génèrent une sortie tabulaire. Les tableaux de sortie peuvent tous être affichés et utilisés de la même manière. Ainsi, il est possible, entre autres, de sélectionner des cellules, de copier tout ou partie du tableau dans le Presse-papiers, de générer de nouveaux noeuds à partir de la sélection actuelle, et d'enregistrer et d'imprimer le tableau.

Sélection de cellules. Pour sélectionner une cellule, cliquez dessus. Pour sélectionner un groupe de cellules, cliquez sur un angle du groupe voulu, faites glisser le pointeur de la souris jusqu'à l'angle opposé, puis relâchez le bouton de la souris. Pour sélectionner une colonne tout entière, cliquez sur son en-tête. Pour sélectionner plusieurs colonnes, cliquez sur leur en-tête en maintenant la touche Maj ou Ctrl enfoncée.

Toute nouvelle sélection annule la précédente. Si vous maintenez la touche Ctrl enfoncée lorsque vous effectuez une sélection, celle-ci sera ajoutée aux sélections existantes (la sélection précédente n'est pas supprimée). Cette fonction permet de sélectionner plusieurs zones non contiguës du tableau. Le menu Edition contient également les options **Sélectionner tout** et **Effacer la sélection**.

Réorganisation des colonnes. Les navigateurs de sortie des noeuds Table et Moyennes vous permettent de déplacer des colonnes du tableau en cliquant sur l'en-tête correspondant, puis en le faisant glisser vers l'emplacement souhaité. Vous ne pouvez déplacer qu'une seule colonne à la fois.

Noeud Table

Le noeud Table crée un tableau qui répertorie les valeurs dans vos données. Tous les champs et toutes les valeurs du flux sont comprises, ce qui facilite l'inspection des valeurs de vos données ou leur exportation sous une forme facilement lisible. De façon facultative, vous pouvez mettre en évidence des enregistrements qui satisfont à une certaine condition.

Remarque : Sauf dans le cas où vous travaillez sur de petits ensembles de données, il est recommandé de sélectionner un sous-ensemble des données à transmettre au noeud Table. Le noeud Table ne s'affiche pas correctement lorsque le nombre d'enregistrements dépasse une taille pouvant être contenue dans la structure d'affichage (par exemple, 100 millions de lignes).

Noeud Table - Onglet Paramètres

Surligner l'enregistrement quand. Pour mettre en évidence certains enregistrements du tableau, entrez une expression CLEM vraie pour chacun de ces enregistrements. Cette option est activée uniquement lorsque l'option **Sortie à l'écran** est sélectionnée.

Noeud Table - Onglet Format

L'onglet Format inclut les options permettant d'indiquer le formatage de chaque champ. Cet onglet est partagé avec le noeud Typer. Pour plus d'informations, voir «Onglet Paramètres du champ», à la page 157.

Noeud de sortie - Onglet Sortie

Pour les noeuds générant une sortie de type tableau, l'onglet Sortie vous permet de définir le format et l'emplacement des résultats.

Nom de la sortie - Spécifie le nom de la sortie générée lorsque le noeud est exécuté. L'option **Automatique** sélectionne un nom en fonction du noeud qui génère la sortie. Si vous le souhaitez, vous pouvez choisir **Personnalisé** pour indiquer un autre nom.

Sortie à l'écran (option par défaut)**.** Crée un objet de sortie à afficher en ligne. L'objet de sortie apparaît dans l'onglet Sorties de la fenêtre du gestionnaire lors de l'exécution du noeud de sortie.

Sortie dans un fichier. Enregistre la sortie dans un fichier lors de l'exécution du noeud. Si vous choisissez cette option, entrez un nom de fichier (ou parcourez l'arborescence et indiquez un nom de fichier à l'aide du sélecteur de fichiers), puis sélectionnez un type de fichier. Il se peut que tous les types de fichier ne soient pas disponibles pour certains types de sortie.

Remarque :

Les données de sortie des noeuds de sortie sont codées conformément aux règles suivantes :

- Lors de l'exécution d'un noeud de sortie, la valeur de codage du flux (définie dans l'onglet des options de flux) est définie sur la sortie.
- Une fois que la sortie a été générée, son codage n'est pas modifié, même si celui du flux l'est.
- Lors de l'exportation de la sortie du noeud de sortie, le fichier de sortie est exporté avec le codage de flux défini. Une fois la sortie créée, même si vous modifiez le codage du flux, celui-ci n'aura aucune incidence sur la sortie générée.

Prenez connaissance des exceptions suivantes qui s'appliquent à ces règles :

- Toutes les exportations HTML sont codées au format UTF-8.
- La sortie du noeud de sortie d'extension est générée par un script utilisateur personnalisé. Ainsi, le codage est contrôlé par le script.

Les options suivantes sont disponibles pour l'enregistrement de la sortie dans un fichier :

- **Données (délimitées par des tabulations) (*.tab).** Cette option crée un fichier texte formaté contenant les valeurs de données. Ce style est souvent utilisé pour créer une représentation en texte brut des informations susceptibles d'être importées dans d'autres applications. Cette option est disponible pour les noeuds Table, Matrice et Moyennes.
- **Données (séparées par des virgules) (*.dat).** Cette option crée un fichier texte séparé par des virgules contenant les valeurs de données. Ce style est souvent utilisé pour générer rapidement un fichier de données susceptible d'être importé dans un tableur ou un autre logiciel d'analyse de données. Cette option est disponible pour les noeuds Table, Matrice et Moyennes.
- **HTML (*.html).** Cette option permet d'écrire une sortie au format HTML dans des fichiers. Dans les sorties tabulaires (noeud Table, Matrice ou Moyennes), les ensembles de fichiers HTML comportent un panneau de contenu répertoriant les noms des champs, ainsi que les données sous forme de tableau HTML. Le tableau peut être partagé entre plusieurs fichiers HTML s'il contient plus de lignes que la valeur définie pour le paramètre **Lignes par page**. Dans ce cas, le panneau de contenu contient des liens correspondant à toutes les pages du tableau et permet de naviguer dans ce dernier. Dans le cas d'une sortie non tabulaire, un seul fichier HTML contenant les résultats du noeud est créé.

  Remarque : Si la sortie HTML ne contient des données de formatage que pour la première page, sélectionnez **Paginer la sortie** et ajustez le paramètre **Lignes par page** afin de regrouper toutes les sorties sur une même page. Si le modèle de sortie des noeuds (noeud Rapport, par exemple) contient des balises HTML personnalisées, assurez-vous d'avoir choisi le type de format **Personnalisé**.
- **Fichier texte (*.txt).** Cette option crée un fichier texte contenant la sortie. Ce style est souvent utilisé pour générer une sortie susceptible d'être importée dans d'autres applications (par exemple, traitement de texte ou création de présentations). Vous ne pouvez pas utiliser cette option avec tous les noeuds.
- **Objet de sortie (*.cou).** Les objets de sortie enregistrés dans ce format peuvent être ouverts et affichés dans IBM SPSS Modeler, ajoutés à des projets, ainsi que publiés et suivis via le IBM SPSS Collaboration and Deployment Services Repository.

Vue de sortie. Pour le noeud Moyennes, vous pouvez indiquer si vous souhaitez afficher par défaut une sortie simple ou avancée. Vous pouvez également basculer entre ces vues lorsque vous parcourez la sortie générée. Pour plus d'informations, voir «Navigateur de sortie du noeud Moyennes», à la page 346.

Format. Pour le noeud Rapport, vous pouvez choisir de formater automatiquement la sortie ou de la formater à l'aide des paramètres HTML indiqués dans le modèle. Sélectionnez **Personnalisé** pour permettre le formatage HTML dans le modèle.

Titre. Dans le cas du noeud Rapport, vous pouvez indiquer un titre facultatif qui apparaîtra en haut de la sortie sous forme de rapport.

Surligner le texte inséré. Dans le cas du noeud Rapport, cette option permet de surligner le texte généré par les expressions CLEM dans le modèle de rapport. Pour plus d'informations, voir la rubrique «Noeud Rapport - Onglet Modèle», à la page 348. Cette option n'est pas recommandée si vous avez opté pour un formatage **personnalisé**.

Lignes par page. Pour le noeud Rapport, indiquez le nombre de lignes à inclure sur chaque page lors du formatage **automatique** du rapport de sortie.

Transposer des données. Cette option transpose les données avant l'exportation, de sorte que les lignes et les colonnes représentent respectivement les champs et les enregistrements.

Remarque : Dans le cas de tableaux volumineux, les options ci-dessus ne fonctionnent pas toujours bien, surtout si vous utilisez un serveur distant. Il est alors préférable d'utiliser un noeud de sortie Fichier. Pour plus d'informations, voir «Noeud d'exportation Fichier à plat», à la page 381.

Navigateur du noeud Table

Le navigateur du noeud Table affiche les données tabulaires et vous permet d'exécuter des opérations standard : sélection et copie de cellules, réorganisation des colonnes, enregistrement et impression du tableau. Pour plus d'informations, voir «Sélection de cellules et de colonnes», à la page 323. Il s'agit des mêmes opérations que vous pouvez effectuer lors de l'aperçu des données dans un noeud.

Exportation des données tabulaires. Vous pouvez exporter des données depuis le navigateur du noeud Table en choisissant :

Fichier > **Exporter**

Pour plus d'informations, voir «Exportation des sorties», à la page 323.

Les données sont exportées dans le format d'encodage par défaut du système qui est spécifié dans le Panneau de configuration de Windows, ou si le système est en mode réparti, sur l'ordinateur serveur.

Recherche dans le tableau. Le bouton de recherche (icône représentant des jumelles) de la barre d'outils principale active la barre d'outils de recherche, qui permet de trouver des valeurs précises dans le tableau. Vous pouvez effectuer une recherche vers le début ou vers la fin du tableau, indiquer si la recherche doit respecter la casse ou non (bouton **Aa**), et interrompre une recherche en cours à l'aide du bouton Interrompre la recherche.

Création de noeuds. Le menu Générer contient des options permettant de générer des noeuds.

- **Noeud Sélectionner (Enregistrements).** Crée un noeud Sélectionner permettant de sélectionner les enregistrements auxquels au moins une cellule sélectionnée dans le tableau est associée.
- **Sélectionner (Et).** Crée un noeud Sélectionner permettant de sélectionner les enregistrements contenant *toutes* les valeurs sélectionnées dans le tableau.
- **Sélectionner (Ou)** Crée un noeud Sélectionner permettant de sélectionner les enregistrements contenant *n'importe quelle* valeur sélectionnée dans le tableau.
- **Calculer (Enregistrements).** Crée un noeud Calculer permettant de créer un champ booléen. Ce dernier contient la valeur *T* pour les enregistrements pour lesquels au moins une cellule du tableau est sélectionnée, et *F* (false - faux) pour les autres.
- **Calculer (Et).** Crée un noeud Calculer permettant de créer un champ booléen. Ce dernier indique la valeur *T* (true - vrai) pour les enregistrements contenant *toutes* les valeurs sélectionnées dans le tableau, et *F* (false - faux) pour les autres.
- **Calculer (Ou).** Crée un noeud Calculer permettant de créer un champ booléen. Ce dernier indique la valeur *T* (true - vrai) pour les enregistrements contenant *n'importe quelle* valeur sélectionnée dans le tableau, et *F* (false - faux) pour les autres.

Noeud Matrice

Le noeud Matrice permet de créer un tableau dans lequel les relations entre les champs sont indiquées. Il s'agit généralement de la relation entre deux champs catégoriels (indicateurs, nominaux ou ordinaux), mais il peut également s'agir de la relation entre des champs continus (intervalle numérique).

Noeud Matrice - Onglet Paramètres

L'onglet Paramètres permet de définir des options pour la structure de la matrice.

Champs. Permet de choisir l'un des types de sélection de champ suivants :

- **Sélectionné.** Cette option permet de sélectionner un champ catégoriel pour les lignes de la matrice et un pour les colonnes. Les lignes et les colonnes de la matrice sont définies par la liste des valeurs du champ catégoriel sélectionné. Les cellules de la matrice contiennent les statistiques récapitulatives sélectionnées plus bas.

- **Tous les indicateurs (valeurs vraies).** Cette option crée une matrice contenant une ligne et une colonne pour chaque champ booléen présent dans les données. Les cellules de la matrice indiquent le nombre d'enregistrements pour lesquels une combinaison de deux champs indicateurs est vraie. En d'autres termes, pour une ligne correspondant à *pain acheté* et une colonne correspondant à *fromage acheté*, la cellule à l'intersection de cette ligne et de cette colonne contient le nombre d'enregistrements pour lesquels *pain acheté* et *fromage acheté* sont vrais.
- **Tous les numériques.** Cette option crée une matrice contenant une ligne et une colonne pour chaque champ numérique. Les cellules de la matrice indiquent la somme des produits croisés pour la paire de champs correspondante. En d'autres termes, pour chaque cellule de la matrice, les valeurs du champ ligne et du champ colonne sont multipliées pour chaque enregistrement, puis additionnées.

Inclure les valeurs manquantes. Inclut les valeurs manquantes utilisateur (blancs) et les valeurs manquantes système ($null$) dans la sortie des lignes et des colonnes. Par exemple, si la valeur *Non applicable* est définie comme valeur manquante utilisateur pour le champ de colonne sélectionné, une autre colonne *Non applicable* est ajoutée, comme toute autre catégorie, au tableau (en supposant que cette valeur figure réellement dans les données). Si cette option est désélectionnée, la colonne *Non applicable* est exclue, quelle que soit sa fréquence.

Remarque : L'option d'ajout de valeurs manquantes ne s'applique que lorsque les champs sélectionnés sont affichés sous forme de tableau croisé. Les valeurs vides sont mappées avec les valeurs nulles ($null$) et sont exclues de l'agrégation pour le champ de fonction lorsque vous vous trouvez en mode **Sélectionné(e)(s)** et que le contenu est paramétré sur **Fonction**, et pour tous les champs numériques lorsque le mode est paramétré sur **Numériques**.

Contenus des cellules. Si vous avez choisi **Sélectionné(e)(s)** dans la zone Champs, vous pouvez indiquer le type de statistique à utiliser dans les cellules de la matrice. Sélectionnez des statistiques basées sur un comptage, ou un champ de superposition récapitulant les valeurs d'un champ numérique en fonction des valeurs des champs ligne et colonne.

- **Tableaux croisés.** Les valeurs des cellules indiquent le nombre et/ou le pourcentage d'enregistrements auxquels la combinaison de valeurs correspondante est associée. Vous pouvez indiquer les récapitulatifs de tableaux croisés de votre choix à l'aide des options de l'onglet Apparence. La valeur Chi-deux globale et la signification sont également affichées. Pour plus d'informations, voir «Navigateur de sortie du noeud Matrice», à la page 328.
- **Fonction.** Si vous sélectionnez une fonction récapitulative, les valeurs des cellules sont une fonction des valeurs de champ de superposition sélectionnées pour les observations où les valeurs de ligne et de colonne appropriées existent. Par exemple, si le champ ligne est *Région*, le champ colonne *Produit* et le champ de superposition *Revenu*, la cellule située à l'intersection de la ligne *Nord-est* et de la colonne *Widget* contiendra la somme (ou la moyenne, ou la valeur minimale ou maximale) des revenus provenant de la vente de widgets dans la région nord-est. La fonction récapitulative par défaut est **Moyenne**. Vous pouvez sélectionner une autre fonction pour la récapitulation du champ Fonction. Les options possibles sont les suivantes : **Moyenne**, **Somme**, **Ecart-type**, **Maximum** et **Minimum**.

Noeud Matrice - Onglet Apparence

L'onglet Apparence permet de définir des options de tri et de surlignage pour la matrice, ainsi que les statistiques présentées pour les matrices de tableau croisé.

Lignes et colonnes. Permet de contrôler le tri des en-têtes de ligne et de colonne de la matrice. La valeur par défaut est **Non trié**. Sélectionnez **Croissant** ou **Décroissant** en fonction de l'ordre de tri des en-têtes de ligne et de colonne voulu.

Superposition. Vous permet de surligner les valeurs extrêmes de la matrice. Les valeurs sont surlignées sur la base du nombre de cellules (pour les matrices de tableau croisé) ou des valeurs calculées (pour les matrices de fonction).

- **Mettre en évidence les plus grandes.** Cette option permet de mettre en évidence (en rouge) les valeurs les plus élevées de la matrice. Vous devez indiquer le nombre de valeurs à mettre en évidence.
- **Mettre en évidence les plus petites.** Cette option permet de mettre en évidence (en vert) les valeurs les moins élevées de la matrice. Vous devez indiquer le nombre de valeurs à mettre en évidence.

Remarque : s'il existe des valeurs ex-aequo, il est possible que le nombre de valeurs surlignées soit supérieur au nombre indiqué. Par exemple, dans le cas d'une matrice dont six cellules contiennent des zéros, si vous sélectionnez **Surligner le bas 5**, les six zéros seront surlignés.

Contenu de cellule de tableau croisé. Pour les tableaux croisés, vous pouvez indiquer les statistiques récapitulatives contenues dans la matrice dédiée aux matrices de tableau croisé. Ces options ne sont pas disponibles lorsque la fonction **Numériques** ou **Fonction** est sélectionnée dans l'onglet Paramètres.

- **Comptages.** Les cellules indiquent le nombre d'enregistrements dont la valeur de ligne a la valeur de colonne correspondante. Il s'agit uniquement du contenu de cellule par défaut.
- **Valeurs théoriques.** Valeur théorique du nombre d'enregistrements dans la cellule, en supposant qu'il n'existe aucune relation entre les lignes et les colonnes. Les valeurs théoriques sont basées sur la formule suivante :

  p(*row value*) * p(*column value*) * *total number of records*
- **Résiduels.** Différence entre les valeurs observées et les valeurs théoriques.
- **Pourcentage de lignes.** Pourcentage de tous les enregistrements dont la valeur de ligne a la valeur de colonne correspondante. Le pourcentage maximal pour une ligne est égal à 100.
- **Pourcentage de colonnes.** Pourcentage de tous les enregistrements dont la valeur de colonne a la valeur de ligne correspondante. Le pourcentage maximal pour une colonne est égal à 100.
- **Pourcentage du total.** Pourcentage de tous les enregistrements présentant la combinaison valeur de colonne/valeur de ligne. Le pourcentage maximal pour la matrice est égal à 100.
- **Inclure les totaux des lignes et colonnes.** Ajoute une ligne et une colonne à la matrice pour les totaux.
- **Appliquer les paramètres.** (Navigateur de sortie seulement) Vous permet de modifier l'apparence de la sortie du noeud Matrice sans avoir besoin de fermer et rouvrir le navigateur de sortie. Effectuez les changements dans cet onglet du navigateur de sortie, cliquez sur ce bouton et sélectionnez l'onglet Matrice pour visualiser l'impact des changements.

Navigateur de sortie du noeud Matrice

Le navigateur du noeud Matrice affiche les données sous la forme d'un tableau croisé dans lequel vous pouvez effectuer un certain nombre d'opérations : sélection de cellules, copie totale ou partielle de la matrice dans le Presse-papiers, création de noeuds en fonction d'une sélection, enregistrement et impression de la matrice. Le navigateur du noeud Matrice permet également d'afficher les sorties de certains modèles, comme les modèles Naive Bayes d'Oracle.

Les menus Fichier et Edition offrent les fonctions habituelles d'impression, d'enregistrement, d'exportation de sortie, ainsi que de sélection et de copie des données. Pour plus d'informations, voir «Affichage de la sortie», à la page 321.

Khi-deux. Pour le tableau croisé de deux champs catégoriels, le Pearson du Chi-deux global apparaît également sous le tableau. Ce test indique la probabilité que les deux champs ne soient pas liés, sur la base de la différence entre les valeurs observées et les valeurs théoriques en l'absence de relation. Par exemple, s'il n'existe aucune relation entre la satisfaction client et l'emplacement des magasins, vous vous attendez à des taux de satisfaction semblables pour tous les magasins. En revanche, si certains magasins enregistrent de forts taux de satisfaction par rapport aux autres, tout laisse à penser qu'il ne s'agit pas d'une simple coïncidence. Plus la différence est importante, plus la probabilité que cela soit dû uniquement à une erreur d'échantillonnage aléatoire est faible.

- Le test du Chi-deux indique la probabilité que les deux champs ne soient pas liés, auquel cas les différences éventuelles entre les effectifs observées et les prévisions d'effectif ne relèvent que du hasard. Si cette probabilité est infime (en général, inférieure à 5 %), la relation entre les deux champs est considérée comme significative.
- Si une seule colonne ou une seule ligne est utilisée (test du Chi-deux unilatéral), les degrés de liberté correspondent au nombre de cellules moins un. Pour un test du khi-deux bilatéral, les degrés de liberté correspondent au nombre de lignes moins le nombre de colonnes moins un.
- Soyez vigilant en interprétant les statistiques Chi-deux lorsque l'une des prévisions d'effectif de cellule est inférieure à cinq.
- Le test du Chi-deux n'est disponible que pour le tableau croisé de deux champs. (Lorsque l'option **Tous les booléens** ou **Numériques** est sélectionnée dans l'onglet Paramètres, ce test n'apparaît pas.)

Menu Générer. Le menu Générer contient des options permettant de générer des noeuds. Ces options sont disponibles uniquement pour les matrices à tableau croisé et vous devez avoir sélectionné au moins une cellule dans la matrice.

- **Noeud Sélectionner.** Crée un noeud Sélectionner permettant de sélectionner les enregistrements correspondant à au moins une cellule sélectionnée dans la matrice.
- **Noeud dériver (indicateur).** Crée un noeud Calculer permettant de créer un champ booléen. Ce dernier contient la valeur *T* (true - vrai) pour les enregistrements correspondant à au moins une cellule sélectionnée dans la matrice, et *F* (false - faux) pour les autres.
- **Noeud Calculer (Ensemble).** Crée un noeud Calculer permettant de créer un champ nominal. Le champ nominal contient une catégorie pour chaque ensemble contigu de cellules sélectionnées dans la matrice.

Noeud Analyse

Les noeuds Analyse vous permettent d'évaluer la capacité d'un modèle à générer des prévisions précises. Les noeuds Analyse comparent les valeurs prédites et les valeurs réelles (votre champ cible) d'un ou de plusieurs nuggets de modèle. Ils peuvent également être utilisés pour comparer des modèles prédictifs entre eux.

Lorsque que vous exécutez un noeud Analyse, un récapitulatif des résultats de l'analyse est automatiquement ajouté à la section Analyse de l'onglet Récapitulatif pour chaque nugget de modèle du flux exécuté. Les résultats détaillés de l'analyse apparaissent dans l'onglet Sorties de la fenêtre de gestionnaire ; ils peuvent également être écrits directement dans un fichier.

Remarque : Etant donné que les noeuds Analyse comparent les valeurs prédites aux valeurs réelles, ils ne sont utiles qu'avec les modèles supervisés (ceux qui requièrent un champ cible). Pour les modèles non supervisés, comme les algorithmes de classification non supervisée, aucun résultat réel n'est disponible pour servir de base à la comparaison.

Noeud Analyse - Onglet Analyse

L'onglet Analyse permet d'indiquer les détails de l'analyse.

Matrices de coïncidence (pour les cibles symboliques ou catégorielles). Affiche le motif des correspondances entre chaque champ généré (prédit) et le champ cible associé pour les cibles catégorielles (indicateur, nominal ou ordinal). Un tableau apparaît, dans lequel les lignes sont définies par des valeurs réelles et les colonnes par des valeurs prédites, chaque cellule indiquant le nombre d'enregistrements auxquels ce motif correspond. Cette fonction permet notamment d'identifier les erreurs systématiques dans les prévisions. Si plusieurs champs générés sont reliés au même champ de sortie alors qu'ils sont issus de modèles différents, le nombre de fois où ces champs sont en accord ou en désaccord est calculé et affiché. Lorsqu'ils sont en accord, d'autres statistiques correctes/incorrectes apparaissent.

Evaluation des performances. Affiche les statistiques d'évaluation des performances des modèles produisant des sorties catégorielles. Ces statistiques, affichées pour chaque catégorie des champs de sortie, indiquent la taille moyenne (en bits) des informations générées par le modèle utilisé pour la prévision des enregistrements appartenant à la catégorie en question. Elles tiennent compte des difficultés liées à la classification ; par conséquent, l'index d'évaluation de performances des prévisions précises portant sur des catégories rares sera supérieur à celui des prévisions précises portant sur des catégories courantes. Si le modèle ne permet pas d'obtenir des résultats pertinents pour une catégorie, l'index d'évaluation de performances de cette dernière sera de zéro.

Métriques d'évaluation (AUC & Gini, discriminants binaires uniquement). Pour les discriminants binaires, cette option présente les métriques d'évaluation de coefficient AUC (aire sous la courbe) et Gini. Ces deux métriques d'évaluation sont calculées ensemble pour chaque modèle binaire. Les valeurs des métriques sont indiquées dans une table dans le navigateur de sortie du noeud Analyse.

La métrique d'évaluation AUC est calculée en tant qu'aire sous une courbe ROC (Receiver Operator Characteristic) et constitue une représentation scalaire des performances attendues d'un discriminant. La métrique AUC est toujours comprise entre 0 et 1, un nombre élevé représentant un discriminant de meilleure qualité. Une courbe ROC diagonale entre les coordonnées (0,0) et (1,1) représente un discriminant aléatoire et comporte une métrique AUC de 0,5. Par conséquent, un discriminant réaliste n'aura pas de métrique AUC inférieure à 0,5.

La métrique d'évaluation de coefficient Gini est parfois utilisée comme alternative à la métrique d'évaluation AUC et les deux mesures sont étroitement liées. Le coefficient Gini est calculé en tant que double de l'aire entre la courbe ROC et la diagonale, ou sous la forme Gini = 2AUC - 1. Le coefficient Gini est toujours compris entre 0 et 1, un nombre élevé représentant un discriminant de meilleure qualité. Le coefficient Gini est négatif dans le cas peu probable où la courbe ROC se situe en dessous de la diagonale.

Niveau de confiance (si disponible) Pour les modèles qui génèrent un champ de fiabilité, cette option affiche des statistiques sur les valeurs de confiance et leurs relations avec les prévisions. Deux paramètres peuvent être définis pour cette option :

- **Seuil de.** Indique le niveau de confiance au-delà duquel la précision sera égale au pourcentage spécifié.
- **Améliorer l'exactitude.** Indique le niveau de confiance au-delà duquel la précision sera améliorée par le facteur spécifié. Par exemple, si la précision globale est de 90 % et que cette option est paramétrée sur 2, la valeur affichée correspondra au niveau de confiance requis pour une précision de 95 %.

Rechercher les champs prédits/de prédicteur avec. Détermine la façon dont les champs prédits sont en correspondance avec le champ cible d'origine.

- **Modéliser les métadonnées de champ de sortie.** Fait correspondre les champs prédits à la cible en fonction des informations du champ de modèle, ce qui autorise une correspondance même si un champ prédit a été renommé. Les informations du champ de modèle peuvent aussi être accédées pour tout champ prédit à partir de la boîte de dialogue Valeurs grâce à un noeud Typer. Pour plus d'informations, voir «Utilisation de la boîte de dialogue Valeurs», à la page 151.
- **Format de nom de champ.** Fait correspondre des champs en fonction de la convention de dénomination. Par exemple, des valeurs prédites générées par un nugget de modèle C5.0 pour une cible nommée *réponse* doivent se trouver dans un champ nommé *$C-réponse*.

Séparer par partition. Si un champ de partition est utilisé pour diviser des enregistrements en échantillons d'apprentissage, de test et de validation, sélectionnez cette option pour afficher les résultats séparément pour chaque partition. Pour plus d'informations, voir «Noeud Partitionner», à la page 185.

Remarque : lorsque vous séparez des enregistrements par partition, ceux dont le champ de partition contient des valeurs nulles sont exclus de l'analyse. Ce problème ne se pose jamais si un noeud Partitionner est utilisé, car ce type de noeud ne génère aucune valeur nulle.

Analyse définie par l'utilisateur. Permet d'indiquer le calcul d'analyse à utiliser pour l'évaluation des modèles. Vous pouvez utiliser des expressions CLEM pour spécifier les éléments calculés pour chaque enregistrement et pour combiner les scores de niveau enregistrement en un score global. Utilisez les fonctions **@TARGET** et **@PREDICTED** pour faire référence respectivement à la valeur cible (sortie réelle) et à la valeur prédite.

- **Si.** Indiquez une expression conditionnelle pour utiliser des calculs différents en fonction de certaines conditions.
- **Donc.** Indiquez le calcul à utiliser si la condition Si a la valeur true (vrai).
- **Sinon.** Indiquez le calcul à utiliser si la condition Si a la valeur false (faux).
- **Utiliser.** Sélectionnez les statistiques à utiliser pour calculer un score global à partir des scores individuels.

Décomposition de l'analyse par champ. Affiche les champs catégoriels disponibles pour la décomposition de l'analyse. Outre l'analyse globale, une analyse distincte sera effectuée pour chaque catégorie de chaque champ de décomposition.

Navigateur de sortie du noeud Analyse

Le navigateur de sortie du noeud Analyse affiche les résultats de l'exécution du noeud Analyse. Les options standard d'enregistrement, d'exportation et d'impression sont disponibles dans le menu Fichier. Pour plus d'informations, voir «Affichage de la sortie», à la page 321.

Lorsque vous accédez pour la première fois à la sortie du noeud Analyse, les résultats sont développés. Pour masquer les résultats après les avoir consultés, utilisez la commande de développement située à gauche des résultats à masquer ou cliquez sur le bouton **Réduire tout** pour réduire tous les résultats. Pour afficher de nouveau les résultats, utilisez la commande de développement située à gauche des résultats à afficher ou cliquez sur le bouton **Développer tout** pour développer tous les résultats.

Résultats du champ de sortie. La sortie du noeud Analyse contient une section pour chaque champ de sortie pour lequel il existe un champ de prévision créé par un modèle généré.

Comparaison. La section du champ de sortie contient une sous-section pour chaque champ de prévision associé au champ de sortie. Pour les champs de sortie catégoriels, la partie supérieure de cette section contient un tableau indiquant le nombre et le pourcentage de prévisions correctes et incorrectes, et le nombre total d'enregistrements dans le flux. Pour les champs de sortie numériques, cette section contient les informations suivantes :

- **Nombre minimal d'erreurs.** Affiche le nombre minimal d'erreurs (différence entre les valeurs observées et les valeurs prédites).
- **Nombre maximal d'erreurs.** Affiche le nombre maximal d'erreurs.
- **Nombre moyen d'erreurs.** Affiche le nombre moyen d'erreurs sur l'ensemble des enregistrements. Indique s'il existe un **biais** systématique (tendance à surestimer au lieu de sous-estimer ou inversement) dans le modèle.
- **Erreur absolue moyenne.** Affiche la moyenne des valeurs absolues des erreurs sur l'ensemble des enregistrements. Indique la grandeur moyenne des erreurs, indépendamment de la direction.
- **Ecart type.** Indique l'écart-type des erreurs.
- **Corrélation linéaire.** Indique la corrélation linéaire entre les valeurs prédites et réelles. Ces statistiques varient entre –1.0 et 1.0. Les valeurs proches de +1,0 indiquent une association positive forte, de sorte que les valeurs prédites élevées sont associées à des valeurs réelles élevées et les valeurs prédites faibles à des valeurs réelles faibles. Les valeurs proches de -1.0 indiquent une association négative forte, de sorte que les valeurs prédites élevées sont associées à des valeurs réelles faibles, et inversement. Les valeurs proches de 0.0 indiquent une association faible, de sorte que les valeurs prédites sont plus ou moins indépendantes des valeurs réelles. *Remarque* : Une entrée vide présente ici indique que la corrélation linéaire ne peut pas être calculée dans ce cas, car les valeurs réelles ou prédites sont constantes.

- **Occurrences.** Indique le nombre d'enregistrements utilisés dans l'analyse.

Matrice de coïncidences. Pour les champs de sortie catégoriels, si vous avez indiqué une matrice de coïncidences dans les options d'analyse, une sous-section contenant cette matrice apparaît. Les lignes représentent les valeurs réelles observées tandis que les colonnes représentent les valeurs prédites. La cellule du tableau indique le nombre d'enregistrements pour chaque combinaison valeurs prédites/valeurs réelles.

Evaluation des performances. Pour les champs de sortie catégoriels, si vous avez spécifié des statistiques d'évaluation des performances dans les options d'analyse, les résultats d'évaluation des performances apparaissent ici. Chaque catégorie de sortie est répertoriée avec les statistiques d'évaluation des performances correspondantes.

Rapport de valeurs de confiance. Pour les champs de sortie catégoriels, si vous avez spécifié des valeurs de fiabilité dans les options d'analyse, ces valeurs apparaissent ici. Les statistiques suivantes sont indiquées pour les valeurs de fiabilité du modèle :

- **Intervalle.** Indique l'intervalle (valeurs les plus faibles et les plus élevées) des valeurs de fiabilité pour les enregistrements des données du flux.
- **Moyenne correcte.** Indique la valeur de fiabilité moyenne des enregistrements correctement classés.
- **Moyenne incorrecte.** Indique la valeur de fiabilité moyenne des enregistrements non correctement classés.
- **Toujours correct au-dessus de.** Indique le seuil de fiabilité au-dessus duquel les prévisions sont toujours correctes et le pourcentage d'observations qui répondent à ce critère.
- **Toujours incorrect au-dessous de.** Indique le seuil de fiabilité en dessous duquel les prévisions sont toujours fausses et le pourcentage d'observations qui répondent à ce critère.
- **Degré d'exactitude au-dessus de = X %.** Indique le niveau de confiance correspondant à une exactitude de *X* %. *X* est approximativement la valeur spécifiée pour **Seuil de** dans les options d'analyse. Pour certains modèles et jeux de données, il n'est pas possible de choisir une valeur de fiabilité qui indique le seuil exact spécifié dans les options (généralement en raison de clusters d'observations similaires ayant la même valeur de fiabilité à proximité du seuil). La valeur de seuil indiquée est la valeur la plus proche du critère de précision spécifié pouvant être obtenue avec un même seuil de valeur de fiabilité.
- **Réduction correcte au-dessus = X.** Indique la valeur de confiance au niveau de laquelle l'exactitude est *X* fois meilleure qu'elle ne l'est pour le jeu de données global. *X* est la valeur spécifiée dans le champ **Améliorer la précision** des options d'analyse.

Accord entre. Si plusieurs modèles générés prédisant le même champ de sortie sont inclus dans le flux, vous pouvez également consulter des statistiques sur l'**accord** entre les prévisions générées par les modèles. Il peut s'agir du nombre et du pourcentage d'enregistrements pour lesquels les prévisions sont concordantes (pour les champs de sortie catégoriels), ou de statistiques récapitulant les erreurs (pour les champs de sortie continus). Pour les champs catégoriels, une analyse comparant les prévisions aux valeurs réelles est incluse pour le sous-ensemble d'enregistrements sur lesquels les modèles sont concordants (c'est-à-dire, génèrent la même valeur prédite).

Métriques d'évaluation. Pour les discriminants binaires, si vous avez demandé des métriques d'évaluation dans les options d'analyse, les valeurs des métriques d'évaluation de coefficient AUC et Gini s'affichent dans une table de cette section. La table contient une ligne par modèle de discriminant binaire. La table des métriques d'évaluation est affichée pour chaque zone de sortie plutôt que pour chaque modèle.

Noeud Audit données

Le noeud Audit données offre un premier aperçu complet des données que vous entrez dans IBM SPSS Modeler. Celles-ci sont présentées sous la forme d'une matrice très lisible que vous pouvez trier et à partir de laquelle vous pouvez générer des graphiques grandeur nature et divers noeuds de préparation des données.

- L'onglet Audit affiche un rapport qui fournit des statistiques récapitulatives, des histogrammes et des graphiques Proportion qui peuvent contribuer à une première compréhension des données. Le rapport affiche aussi l'icône de stockage devant le nom de champ.
- L'onglet Qualité du rapport d'audit affiche des informations sur les valeurs éloignées, les extrêmes et les valeurs manquantes, et propose des outils de gestion de ces valeurs.

Utilisation du noeud Audit données

Le noeud Audit données peut être connecté directement à un noeud source ou en aval d'un noeud Typer instancié. Vous pouvez également générer des noeuds de préparation des données sur la base des résultats. Par exemple, vous pouvez générer un noeud Filtrer qui exclut les champs contenant trop de valeurs manquantes pour être utiles à la modélisation et générer un super noeud qui attribue les valeurs manquantes à l'un des champs ou à tous les champs restants. Voilà où la puissance réelle de l'audit intervient, vous permettant non seulement d'évaluer l'état actuel de vos données, mais également d'agir sur la base de cette évaluation.

Filtrage ou échantillonnage des données. Etant donné que les audits initiaux sont particulièrement efficaces pour traiter les "données volumineuses", vous pouvez utiliser un noeud Echantillonner pour réduire le temps de traitement lors de l'exploration initiale en sélectionnant uniquement un sous-ensemble d'enregistrements. Vous pouvez également utiliser le noeud Audit données avec d'autres noeuds, tels que Sélection de fonction et Détection des anomalies, lors des phases exploratoires de l'analyse.

Noeud Audit données - Onglet Paramètres

L'onglet Paramètres vous permet de définir les paramètres de base de l'audit.

Par défaut. Vous pouvez tout simplement connecter le noeud au flux et cliquer sur **Exécuter** pour générer un rapport d'audit pour tous les champs, sur la base des paramètres par défaut, comme suit :

- Si aucun paramètre n'a été défini pour le noeud Typer, tous les champs sont inclus dans le rapport.
- Si des paramètres ont été définis pour le noeud Typer (qu'ils soient instanciés ou non), tous les champs *Entrée*,*Cible* et *Les deux* sont inclus dans l'affichage. S'il existe un seul champ *Cible*, utilisez-le en tant que champ de superposition. Si plusieurs champs *Cible* ont été définis, aucune superposition par défaut n'est spécifiée.

Utiliser les champs personnalisés. Choisissez cette option pour sélectionner les champs manuellement. Utilisez le sélecteur de champs à droite pour sélectionner les champs un par un ou par type.

Champ de superposition. Le champ de superposition permet de tracer les graphiques en miniature affichés dans le rapport d'audit. Pour un champ continu (intervalle numérique), les statistiques à deux dimensions (covariance et corrélation) sont également calculées. Si un seul champ *Cible* est présent sur la base des paramètres de noeud type, ce champ est utilisé comme champ de superposition par défaut, conformément à la description précédente. Vous pouvez également sélectionner **Utiliser les champs personnalisés** pour définir une superposition.

Affichage. Permet d'indiquer si des graphiques sont disponibles dans la sortie et de choisir les statistiques affichées par défaut.

- **Graphiques.** Affiche un graphique pour chaque champ sélectionné : un graphique de distribution (en barres), un histogramme ou un nuage de points, selon le type de graphique adapté aux données. Les

graphiques sont affichés sous forme de miniatures dans le rapport initial, mais vous pouvez également générer des graphiques en grandeur nature et des noeuds Graphiques. Pour plus d'informations, voir la rubrique «Navigateur de sortie du noeud Audit données», à la page 335.

- **Statistiques avancées/de base.** Indique le niveau de statistiques affiché par défaut dans la sortie. Bien que ce paramètre détermine l'affichage initial, toutes les statistiques sont disponibles dans la sortie, quel que soit ce paramètre. Pour plus d'informations, voir la rubrique «Afficher les statistiques», à la page 336.

Médiane et mode. Calcule la médiane et le mode de tous les champs figurant dans le rapport. Notez que, lorsque les jeux de données sont volumineux, ces statistiques peuvent augmenter le temps de traitement, car leur calcul dure plus longtemps. Pour le calcul de la médiane uniquement et dans certaines conditions, vous pouvez baser la valeur figurant dans le rapport sur un échantillon de 2000 enregistrements (plutôt que sur le jeu de données complet). Cet échantillonnage est effectué sur la base de chaque champ lorsque les limites de mémoire risquent d'être dépassées. Lorsque l'échantillonnage est actif, les résultats sont étiquetés comme tels dans la sortie (*Médiane de l'échantillon* plutôt que *Médiane*). Toutes les statistiques autres que la médiane sont systématiquement calculées sur la base du jeu de données complet.

Champs vides ou sans type. Lorsqu'ils sont utilisés avec des données instanciées, les champs sans type ne sont pas inclus dans le rapport d'audit. Pour inclure des champs sans type (y compris les champs vides), sélectionnez **Effacer toutes les valeurs** dans les noeuds Typer en amont. Cela garantit que les données ne sont pas instanciées, ce qui entraîne l'inclusion de tous les champs dans le rapport. Par exemple, cela peut se révéler utile si vous souhaitez obtenir la liste complète de tous les champs ou générer un noeud Filtrer qui exclut les champs vides. Pour plus d'informations, voir la rubrique «Filtrage de champs contenant des données manquantes», à la page 339.

Audit données - Onglet Qualité

L'onglet Qualité du noeud Audit données fournit des options de traitement des valeurs manquantes, des valeurs éloignées et des extrêmes.

Valeurs manquantes

- **Nombre d'enregistrements avec valeurs valides.** Sélectionnez cette option pour afficher le nombre d'enregistrements contenant des valeurs valides pour chaque champ évalué. Notez que les valeurs nulles (non définies), les valeurs non renseignées, les espaces blancs et les chaînes vides sont toujours traités comme des valeurs non valides.
- **Ventilation du nombre d'enregistrements avec valeurs non valides.** Sélectionnez cette option pour afficher le nombre d'enregistrements contenant chaque type de valeur non valide pour chaque champ.

Valeurs éloignées et extrêmes

Méthode de détection des valeurs éloignées et extrêmes. Deux méthodes sont prises en charge :

Ecart-type de la moyenne. Détecte les valeurs éloignées et extrêmes sur la base du nombre d'écarts-types par rapport à la moyenne. Par exemple, si un champ a une moyenne égale à 100 et un écart-type standard égal à 10, vous pouvez saisir 3,0 pour indiquer que toute valeur inférieure à 70 ou supérieure à 130 doit être traitée comme une valeur éloignée.

Intervalle interquartile. Détecte les valeurs éloignées et les extrêmes en fonction de l'intervalle interquartile (IQR), lequel représente l'intervalle dans lequel sont compris les deux quartiles centraux (entre les 25e et 75e centiles). Par exemple, si le paramètre par défaut est égal à 1,5, le seuil inférieur des valeurs éloignées est Q1 – 1,5 * IQR et le seuil supérieur, Q3 + 1,5*IQR. Cette option risque de ralentir les performances pour les jeux de données volumineux.

Navigateur de sortie du noeud Audit données

Le navigateur Audit données est un outil puissant permettant d'obtenir une présentation de vos données. L'onglet Audit affiche les graphiques en miniature, des icônes de stockage et les statistiques pour tous les champs, alors que l'onglet Qualité contient des informations sur les valeurs éloignées, les extrêmes et les valeurs manquantes. Sur la base des statistiques récapitulatives et des graphiques initiaux, vous pouvez décider de recoder un champ numérique, de calculer un nouveau champ ou de reclasser les valeurs d'un champ nominal. Si vous le souhaitez, vous pouvez également procéder à une exploration plus approfondie à l'aide d'outils de visualisation avancés. Vous pouvez le faire directement à partir du navigateur de rapport d'audit via le menu Générer pour créer plusieurs noeuds permettant de transformer ou de visualiser vos données.

- Triez les colonnes en cliquant sur l'en-tête de colonne ou réorganisez-les en les faisant glisser. La plupart des opérations de sortie standard sont également prises en charge. Pour plus d'informations, voir «Affichage de la sortie», à la page 321.
- Afficher les valeurs et les intervalles des champs en double-cliquant sur un champ de la colonne Mesure ou Unique.
- Utilisez la barre d'outils ou le menu Edition pour afficher ou masquer les libellés de valeur, ou pour sélectionner les statistiques à afficher. Pour plus d'informations, voir «Afficher les statistiques», à la page 336.
- Vérifiez les icônes de stockage à gauche des noms de champ. Le stockage des données décrit la façon dont les données sont stockées dans un champ. Par exemple, un champ comportant les valeurs 1 et 0 stocke des nombres entiers. Il est à différencier du niveau de mesure, qui décrit l'utilisation des données et n'a aucune incidence sur le stockage. Pour plus d'informations, voir «Définition du stockage et du formatage des champs», à la page 9.

Affichage et génération de graphiques

Si aucune superposition n'est sélectionnée, l'onglet Audit affiche des graphiques en barres (pour les champs nominaux ou indicateurs) ou des histogrammes (pour les champs de type Continu).

Dans le cas de la superposition d'un champ nominal ou indicateur, les valeurs de la superposition déterminent les couleurs des graphiques.

Dans le cas de la superposition d'un champ continu, des nuages de points en deux dimensions sont générés à la place des diagrammes en barres et des histogrammes unidimensionnels. Dans ce cas, l'axe *x* est mappé sur le champ de superposition, ce qui vous permet d'obtenir un tableau où tous les axes *x* sont à la même échelle.

- Pour les champs Indicateur ou Nominal, positionnez le curseur de la souris sur une barre pour afficher la valeur ou le libellé sous-jacent dans une info-bulle.
- Pour les champs de type Indicateur ou Nominal, utilisez la barre d'outils pour rendre verticale l'orientation horizontale des graphiques en miniature.
- Pour générer un graphique en grandeur nature à partir d'une miniature, double-cliquez sur cette dernière, puis sélectionnez **Sortie graphique** dans le menu Générer. *Remarque* : Lorsqu'un graphique en miniature repose sur des données échantillonnées, le graphique généré contient toutes les observations si le flux de données d'origine est resté ouvert.

  Vous ne pouvez générer un graphique que si le noeud Audit données qui a généré la sortie est connecté au flux.
- Pour générer un noeud Graphique correspondant, sélectionnez un ou plusieurs champs dans l'onglet Audit, puis choisissez **Noeud Graphique** dans le menu Générer. Le noeud obtenu est ajouté à l'espace de travail de flux ; il permet de recréer le graphique lorsque le flux est exécuté.
- si un champ d'ensemble de superposition contient plus de 100 valeurs, un avertissement est généré et la superposition n'est pas incluse.

Afficher les statistiques

La boîte de dialogue Afficher les statistiques vous permet de sélectionner les statistiques affichées dans l'onglet Audit. Les paramètres initiaux sont indiqués dans le noeud Audit données. Pour plus d'informations, voir la rubrique «Noeud Audit données - Onglet Paramètres», à la page 333.

minimum. La plus petite valeur d'une variable numérique.

maximum. La plus grande valeur d'une variable numérique.

somme. Somme ou total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.

Plage. Différence entre la valeur maximale et la valeur minimale d'une variable numérique (maximum – minimum).

Moyenne. Mesure de la tendance centrale. Moyenne arithmétique : somme divisée par le nombre d'observations.

Erreur standard de la moyenne. Mesure de la variation de la valeur de la moyenne d'un échantillon à l'autre issus de la même distribution. Cette mesure permet de comparer approximativement la moyenne observée avec une valeur hypothétique (autrement dit, vous pouvez conclure que ces deux valeurs sont différentes si le rapport de la différence avec l'erreur standard est inférieur à -2 ou supérieur à +2).

écart type. Mesure de la dispersion des valeurs autour de la moyenne, égale à la racine carrée de la variance. L'écart type est mesuré dans les mêmes unités que la variable d'origine.

Variance. Mesure de dispersion autour de la moyenne, égal à la somme des écarts au carré par rapport à la moyenne divisé par le nombre d'observations moins un. La variance se mesure en unités, qui sont égales au carré des unités de la variable.

Asymétrie. Mesure de l'asymétrie d'une distribution. La distribution normale est symétrique et présente une valeur de décalage de zéro. Une distribution avec un important décalage positif présente une longue queue vers la droite. Une distribution caractérisée par une importante asymétrie négative présente une extrémité gauche plus allongée. Pour simplifier, une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.

erreur standard d'asymétrie. Rapport de l'asymétrie avec son erreur standard pouvant être utilisé comme test de normalité (autrement dit, vous pouvez conclure à une anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur positive élevée indique une extrémité allongée vers la droite ; une valeur négative extrême produit une extrémité allongée vers la gauche.

Kurtosis. Mesure de l'importance des valeurs extrêmes. Dans le cas d'une distribution normale, la valeur de la statistique d'aplatissement est égale à zéro. Un kurtosis positif indique qu'on observe dans les données plus de valeurs extrêmes que dans une distribution normale. Une valeur négative indique que les données comportent moins de valeurs extrêmes qu'une distribution normale.

erreur standard de Kurtosis. Rapport de kurtosis avec son erreur standard pouvant être utilisé comme test de normalité (autrement dit, vous pouvez conclure à une anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur de kurtosis positive élevée indique que les extrémités de la distribution sont plus longues que celles d'une distribution normale ; une valeur de kurtosis négative présente des extrémités plus courtes (semblables à celles d'une distribution uniforme sous forme de boîtes).

unique. Evalue tous les effets simultanément, en ajustant chaque effet à tous les autres effets d'un type donné.

valide. Observations valides qui ne comportent pas la valeur système manquante ni une valeur manquante définie par l'utilisateur. Notez que les valeurs nulles (non définies), les valeurs non renseignées, les espaces blancs et les chaînes vides sont toujours traités comme des valeurs non valides.

Médiane. Valeur au-dessus et au-dessous de laquelle se trouvent la moitié des observations, le 50e percentile. Si le nombre de cellules est pair, la médiane correspond à la moyenne des deux cellules du milieu lorsqu'elles sont triées dans l'ordre croissant ou décroissant. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées.

Mode. Valeur qui revient le plus souvent. Si plusieurs valeurs partagent la plus grande fréquence d'occurrence, chacune d'elles constitue un mode.

La médiane et le mode sont supprimés par défaut pour améliorer les performances, mais peuvent être sélectionnés dans l'onglet Paramètres du noeud Audit données. Pour plus d'informations, voir la rubrique «Noeud Audit données - Onglet Paramètres», à la page 333.

Statistiques de superpositions

Si un champ de superposition continu (intervalle numérique) est utilisé, les statistiques suivantes sont également disponibles :

Covariance. Mesure d'association non normalisée entre deux variables, égale à la déviation des produits en croix divisée par N-1.

Navigateur Audit données - Onglet Qualité

L'onglet Qualité du navigateur Audit données affiche les résultats de l'analyse de la qualité des données. En outre, il vous permet de spécifier des traitements pour les valeurs éloignées, les extrêmes et les valeurs manquantes.

Attribution des valeurs manquantes : Le rapport d'audit répertorie le pourcentage d'enregistrements complets pour chaque champ, ainsi que le nombre de valeurs valides, de valeurs nulles et de valeurs non renseignées. Vous pouvez choisir d'attribuer les valeurs manquantes appropriées à des champs spécifiques, puis de générer un super noeud pour appliquer ces transformations.

1. Dans la colonne **Attribuer une entrée manquante**, spécifiez le type de valeur à attribuer, le cas échéant. Vous pouvez choisir d'attribuer des valeurs non renseignées ou nulles, ou les deux, ou d'indiquer une condition ou une expression personnalisée sélectionnant les valeurs à attribuer.

   Plusieurs types de valeur manquante sont reconnus par IBM SPSS Modeler :

   - **Valeurs système nulles ou manquantes.** Ces valeurs sont des valeurs "non-chaîne" qui ne sont pas renseignées dans la base de données ou dans le fichier source, et qui n'ont pas été spécifiquement définies comme "manquantes" dans un noeud source ou un noeud Typer. Les valeurs manquantes système sont affichées sous la forme **$null$**. Les chaînes vides ne sont pas considérées comme des valeurs nulles dans IBM SPSS Modeler, même si elles peuvent être traitées comme telles par certaines bases de données.
   - **Chaînes vides et espaces blancs.** Les chaînes vides et les espaces blancs (chaînes sans caractère visible) sont traités différemment des valeurs nulles. Dans la plupart des cas, les chaînes vides sont considérées comme des espaces blancs. Par exemple, si vous choisissez de traiter les espaces blancs comme blancs dans un noeud source ou un noeud Typer, ce paramètre s'applique également aux chaînes vides.
   - **Valeurs manquantes définies par l'utilisateur ou vides.** Ces valeurs sont des valeurs, telles que `inconnu`, `99` ou `-1`, qui sont explicitement définies comme manquantes dans un noeud source ou type. Vous pouvez également, si vous le souhaitez, préciser si les valeurs nulles et les espaces blancs doivent être traités comme des blancs ; un traitement spécial leur est alors appliqué et ils sont exclus de la plupart des calculs. Par exemple, vous pouvez utiliser la fonction `@BLANK` pour traiter comme des blancs ces valeurs, ainsi que d'autres types de valeur manquante.
2. Dans la colonne **Méthode**, spécifiez la méthode à utiliser.

Les méthodes suivantes sont disponibles pour attribuer des valeurs manquantes :

Colonne fixe. Remplacement par une valeur fixe (soit la moyenne du champ, soit la moitié de l'intervalle, soit une constante que vous indiquez).

Aléatoire. Remplacement par une valeur aléatoire fondée sur une loi normale ou uniforme.

Expression. Permet d'indiquer une expression personnalisée. Par exemple, vous pourriez remplacer les valeurs par une variable globale créée par le noeud Valeurs globales.

Algorithme. Remplacement par une valeur prévue par un modèle fondé sur l'algorithme C&RT. Chaque champ auquel une valeur est attribuée à l'aide de cette méthode est associé à un modèle C&RT distinct et à un noeud Remplacer qui remplace les valeurs non renseignées et les valeurs nulles par la valeur prédite par le modèle. Ensuite, un noeud Filtrer est utilisé pour supprimer les champs de prévision générés par le modèle.

3. Pour générer un super noeud Valeurs manquantes, choisissez les options suivantes :

   Générer > **Super noeud des valeurs manquantes**

   La boîte de dialogue Super noeud des valeurs manquantes s'affiche.
4. Sélectionnez **Tous les champs** ou **Champs sélectionnés uniquement**, puis indiquez une taille d'échantillon si vous le souhaitez. (L'échantillon spécifié est un pourcentage. Par défaut, 10 % des enregistrements sont échantillonnés.)
5. Cliquez sur **OK** pour ajouter le super noeud généré à l'espace de travail de flux.
6. Reliez le super noeud au flux pour appliquer les transformations.

Dans le super noeud, une combinaison de noeuds Remplacer, Filtrer et de nugget de modèle est utilisée. Pour comprendre le fonctionnement du super noeud, vous pouvez l'éditer et cliquer sur **Zoom avant**, puis ajouter, éditer ou supprimer des noeuds spécifiques dans le super noeud pour en affiner le comportement.

Gestion des valeurs éloignées et extrêmes : Le rapport d'audit répertorie le nombre de valeurs éloignées et d'extrêmes pour chaque champ en fonction des options de détection spécifiées dans le noeud Audit données. Pour plus d'informations, voir «Audit données - Onglet Qualité», à la page 334. Vous pouvez choisir de forcer, d'isoler ou de rendre nulles ces valeurs pour des champs spécifiques, selon vos besoins, puis de générer un super noeud pour appliquer les transformations.

1. Dans la colonne **Action**, spécifiez la gestion des valeurs éloignées et des extrêmes pour des champs spécifiques, si nécessaire.

   Les actions suivantes sont disponibles pour la gestion des valeurs éloignées et des extrêmes :

   - **Forcer.** Remplace les valeurs éloignées et extrêmes par la valeur la plus proche qui ne sera pas considérée comme extrême. Par exemple, si une valeur éloignée est définie comme étant supérieure ou inférieure à trois écarts-types, toutes les valeurs éloignées sont remplacées par la valeur supérieure ou inférieure comprise dans cet intervalle.
   - **Annuler.** Isole les enregistrements contenant des valeurs éloignées ou extrêmes pour le champ spécifié.
   - **Rendre nul.** Remplace les valeurs éloignées et les extrêmes par la valeur nulle ou manquante système.
   - **Forcer les valeurs éloignées/ignorer les extrêmes.** Ignore les valeurs extrêmes uniquement.
   - **Forcer les valeurs éloignées/rendre nulles les extrêmes.** Rend nulles les valeurs extrêmes uniquement.
2. Pour générer le super noeud, choisissez les options suivantes à partir des menus :

   Générer > **Super noeud de valeur éloignée et d'extrême**

   La boîte de dialogue Super noeud de valeur éloignée s'affiche.
3. Sélectionnez **Tous les champs** ou **Champs sélectionnés uniquement**, puis cliquez sur **OK** pour ajouter le super noeud généré à l'espace de travail de flux.
4. Reliez le super noeud au flux pour appliquer les transformations.

Si nécessaire, vous pouvez éditer le super noeud et effectuer un zoom avant à des fins de navigation ou de changement. Dans le super noeud, les valeurs sont supprimées, forcées ou rendues nulles par le biais des noeuds Sélectionner et/ou Remplacer appropriés.

Filtrage de champs contenant des données manquantes : A partir du navigateur Data Audit, vous pouvez créer un noeud Filtrer sur la base des résultats de l'analyse de la qualité à l'aide de la boîte de dialogue Générer un filtre à partir de Qualité.

Mode. Sélectionnez l'option souhaitée pour les champs indiqués : **Enlever** ou **Isoler**.

- **Champs sélectionnés.** Le noeud Filtrer inclut/exclut les champs sélectionnés dans l'onglet Qualité. Par exemple, vous pouvez trier le tableau en fonction de la colonne **% terminé(s)**, maintenir la touche Maj enfoncée tout en cliquant sur les champs les moins complets pour les sélectionner, puis générer un noeud Filtrer excluant ces champs.
- **Champs dont le pourcentage de qualité est supérieur à.** Le noeud Filtrer inclut/exclut les champs dont le pourcentage d'enregistrements complets est supérieur au seuil indiqué. La valeur de seuil par défaut est 50 %.

Filtrage de champs vides ou sans type

Une fois les valeurs de données instanciées, les champs vides ou sans type sont exclus des résultats d'audit et de la plupart des autres résultats dans IBM SPSS Modeler. Ces champs sont ignorés à des fins de modélisation, mais peuvent amplifier ou encombrer les données. Dans ce cas, vous pouvez utiliser le navigateur Audit données pour générer un noeud Filtrer supprimant ces champs du flux.

1. Pour vérifier que tous les champs figurent dans l'audit, y compris les champs vides ou sans type, cliquez sur **Effacer toutes les valeurs** dans le noeud Typer ou source en amont, ou définissez Valeurs sur <*Transférer*> pour tous les champs.
2. Dans le navigateur Audit données, triez le tableau en fonction de la colonne **% terminé(s)**, sélectionnez les champs ne contenant aucune valeur valide (ou un autre seuil) et utilisez le menu Générer pour générer un noeud Filtrer pouvant être ajouté au flux.

Sélection d'enregistrements contenant des valeurs manquantes : A partir du navigateur Audit données, vous pouvez créer un noeud Sélectionner sur la base des résultats de l'analyse de la qualité.

1. Dans le navigateur Audit données, sélectionnez l'onglet Qualité.
2. A partir du menu, sélectionnez :

   Générer > **Noeud Sélectionner les valeurs manquantes**

   La boîte de dialogue Générer le noeud Sélectionner s'affiche.

Sélectionnez cette option lorsque l'enregistrement est. Indiquez si les enregistrements doivent être conservés lorsque leur statut est **Valide** ou **Non valide**.

Rechercher les valeurs non valides dans. Indiquez où rechercher des valeurs non valides.

- **Tous les champs.** Le noeud Sélectionner recherche les valeurs non valides dans tous les champs.
- **Champs sélectionnés dans le tableau.** Le noeud Sélectionner ne vérifie que les champs sélectionnés dans le tableau de sortie Qualité.
- **Champs dont le pourcentage de qualité est supérieur à.** Le noeud Sélectionner ne vérifie que les champs dont le pourcentage d'enregistrements complets est supérieur au seuil indiqué. La valeur de seuil par défaut est 50 %.

Un enregistrement est considéré comme non valide lorsqu'il contient une valeur non valide. Indiquez la condition d'identification d'un enregistrement comme non valide.

- **N'importe quel champ ci-dessus.** Le noeud Sélectionner considère qu'un enregistrement n'est pas valide si l'*un* des champs spécifiés ci-dessus contient une valeur non valide pour cet enregistrement.

- **Tous les champs ci-dessus.** Le noeud Sélectionner considère qu'un enregistrement n'est pas valide si *tous* les champs spécifiés ci-dessus contiennent des valeurs non valides pour cet enregistrement.

Génération d'autres noeuds en vue d'une préparation de données

La plupart des noeuds servant à la préparation des données peuvent être générés directement à partir du navigateur Audit données, y compris les noeuds Recoder, Discrétiser et Calculer. Par exemple :

- Vous pouvez calculer un nouveau champ sur la base des valeurs *valeurréclamation* et *revenuferme* en les sélectionnant dans le rapport d'audit et en choisissant **Calculer** dans le menu Générer. Le nouveau noeud est ajouté à l'espace de travail de flux.
- De même, sur la base des résultats de l'audit, vous pouvez déterminer si le recodage de *revenuferme* en intervalles de type centile fournit une analyse plus précise. Pour générer un noeud Discrétiser, sélectionnez la ligne de champ dans l'affichage et choisissez **Discrétiser** dans le menu Générer.

Une fois le noeud généré et ajouté à l'espace de travail de flux, vous devez le relier au flux et ouvrir le noeud afin de spécifier les options des champs sélectionnés.

Noeud Transformation

La normalisation des champs d'entrée est une étape importante préalable à l'application de techniques d'évaluation traditionnelles, telles que la régression, la régression logistique et l'analyse discriminante. Ces techniques reposent sur des hypothèses relatives aux proportions normales des données qui peuvent ne pas s'appliquer à de nombreux fichiers de données brutes. Une méthode de traitement des données concrètes consiste à appliquer des transformations qui rapprochent un élément de données brutes d'une proportion plus normale. En outre, les champs normalisés sont facilement comparables entre eux. Par exemple, les revenus et l'âge se situent sur des échelles totalement différentes dans un fichier de données brutes. Une fois ces éléments normalisés, leur impact relatif est facile à interpréter.

Le noeud Transformation fournit un afficheur de résultats qui vous permet de procéder à une évaluation visuelle rapide de la meilleure transformation à utiliser. Vous pouvez voir en un coup d'oeil si les variables sont normalement réparties et, si nécessaire, choisir la transformation à appliquer. Vous pouvez choisir plusieurs champs et appliquer une transformation par champ.

Après avoir sélectionné les transformations préférées pour les champs, vous pouvez générer des noeuds Calculer ou Remplacer qui exécutent les transformations et connecter ces noeuds au flux. Le noeud Calculer crée des champs tandis que le noeud Remplacer transforme les champs existants. Pour plus d'informations, voir «Génération de graphiques», à la page 342.

Onglet Champs du noeud Transformation

Dans l'onglet Champs, indiquez les champs de données à utiliser pour afficher les transformations possibles et les appliquer. Seuls les champs numériques peuvent être transformés. Cliquez sur le bouton de sélection de champ, puis sélectionnez un ou plusieurs champs numériques dans la liste affichée.

Onglet Options du noeud Transformation

L'onglet Options permet d'indiquer les types de transformation à inclure. Vous pouvez décider d'inclure toutes les transformations disponibles ou sélectionner des transformations distinctes.

Dans ce dernier cas, vous pouvez également entrer un nombre pour décaler les données en vue des transformations inverse et logarithmique. Cela se révèle utile dans les cas où une large proportion de zéros dans les données pourrait biaiser les résultats de moyenne et d'écart-type.

Par exemple, supposez que vous avez un champ nommé *BALANCE* qui contient des valeurs nulles et que vous souhaitez lui appliquer une transformation inverse. Pour éviter tout biais indésirable, vous sélectionnez **Inverse (1/x)** et entrez 1 dans le champ **Utiliser un décalage de données**. (Notez que ce décalage n'est pas lié au décalage appliqué par la fonction séquentielle `@OFFSET` dans IBM SPSS Modeler.)

Toutes les formules. Indique que toutes les transformations disponibles doivent être calculées et figurer dans la sortie.

Sélectionner les formules. Permet de sélectionner les transformations à calculer et à afficher dans la sortie.

- **Inverse (1/x).** Indique que la transformation inverse doit être affichée dans la sortie.
- **Logarithme (log n).** Indique que la transformation $\log_n$ doit être affichée dans la sortie.
- **Logarithme (log 10).** Indique que la transformation $\log_{10}$ doit être affichée dans la sortie.
- **Exponentiel.** Indique que la transformation exponentielle (e^x) doit figurer dans la sortie.
- **Racine carrée.** Indique que la transformation racine carrée doit être affichée dans la sortie.

Onglet Sortie du noeud Transformation

L'onglet Sortie vous permet de préciser le format de sortie et l'emplacement de la sortie. Vous pouvez également choisir d'afficher les résultats à l'écran ou de les envoyer vers un des types de fichier standard. Pour plus d'informations, voir «Noeud de sortie - Onglet Sortie», à la page 324.

Afficheur de résultats du noeud Transformation

L'afficheur de résultats vous permet de consulter les résultats de l'exécution du noeud Transformation. L'afficheur est un outil puissant qui affiche plusieurs transformations par champ dans des vues miniatures de la transformation, vous permettant ainsi de comparer rapidement les champs. Utilisez les options du menu Fichier pour enregistrer, exporter ou imprimer la sortie. Pour plus d'informations, voir «Affichage de la sortie», à la page 321.

Sous chaque transformation (autre que Transformation sélectionnée), une légende est affichée sous le format :

```
Moyenne (écart-type)
```

Génération des noeuds pour les transformations

L'afficheur de résultats fournit un point de départ à la préparation des données. Par exemple, vous souhaitez normaliser le champ *AGE* pour pouvoir utiliser une technique d'évaluation (telle que la régression logistique ou l'analyse discriminante) qui suppose une proportion normale. D'après les graphiques initiaux et les statistiques récapitulatives, vous pouvez décider de transformer le champ *AGE* en fonction d'une distribution particulière (par exemple, log). Après avoir sélectionné la proportion préférée, vous pouvez générer un noeud de dérivation avec une transformation standardisée à utiliser pour l'évaluation.

Vous pouvez générer les noeuds d'opérations de champ suivants à partir de l'afficheur de résultats :

- Dériver
- Remplissage

Un noeud Calculer crée des champs avec les transformations souhaitées, tandis que le noeud Remplacer transforme les champs existants. Les noeuds sont placés dans l'espace de travail, sous la forme d'un super noeud.

Si vous sélectionnez la même transformation pour différents champs, un noeud Calculer ou Remplacer contient les formules de ce type de transformation pour tous les champs auxquels cette transformation s'applique. Par exemple, supposez que vous avez sélectionné les champs et les transformations sélectionnés pour générer un noeud Calculer .

Tableau 44. Exemple de génération de noeud Calculer.

Zone	Transformation
AGE	Proportion actuelle

Tableau 44. Exemple de génération de noeud Calculer (suite).

INCOME	Log
OPEN_BAL	Inverse
BALANCE	Inverse

Les noeuds suivants sont inclus dans le super noeud :

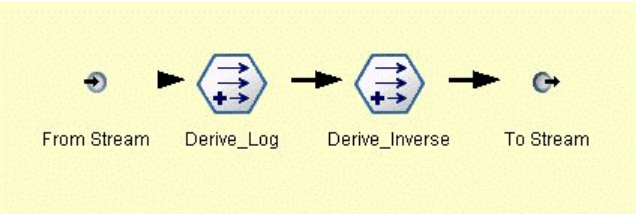


Figure 74. Super noeud dans l'espace de travail

Dans cet exemple, le noeud Derive_Log contient la formule logarithmique du champ *INCOME* et le noeud Derive_Inverse, les formules inverses des champs *OPEN_BAL* et *BALANCE*.

Pour générer un noeud

1. Pour chaque champ apparaissant dans l'afficheur de résultats, sélectionnez la transformation souhaitée.
2. Dans le menu Générer, choisissez **Noeud Calculer** ou **Noeud Remplacer**.

   Cela affiche la boîte de dialogue Générer le noeud Calculer ou Générer le noeud Remplacer, selon le cas.

   Choisissez **Transformation non standardisée** ou **Transformation standardisée (centrer-réduire)**, comme vous le souhaitez. La seconde option applique un score *z* à la transformation ; les scores *z* représentent les valeurs en tant que fonction de la distance par rapport à la moyenne de la variable dans les écarts-types. Par exemple, si vous appliquez la transformation logarithmique au champ *AGE* et que vous choisissez une transformation standardisée, l'équation finale du noeud généré est :

   ```
   (log(AGE)-Mean)/SD
   ```

Lorsqu'un noeud est généré et qu'il apparaît dans l'espace de travail de flux :

1. Connectez-le au flux.
2. Dans le cas d'un super noeud, vous pouvez double-cliquer sur le noeud pour consulter son contenu.
3. Vous pouvez double-cliquer sur un noeud Calculer ou Remplacer pour modifier les options des champs sélectionnés.

Génération de graphiques : Vous pouvez générer une sortie d'histogramme grandeur nature à partir d'un histogramme en miniature dans l'afficheur de résultats.

Pour générer un graphique

1. Double-cliquez sur un graphique en miniature dans l'afficheur de résultats.

   ou

   Sélectionnez un graphique en miniature dans l'afficheur de résultats.
2. Dans le menu Générer, sélectionnez **Sortie graphique**.

Vous affichez ainsi un histogramme avec une courbe de distribution normale en superposition. Cela vous permet de déterminer à quel point chaque transformation disponible correspond à la proportion normale.

Remarque : Vous ne pouvez générer un graphique que si le noeud Transformer qui a créé la sortie est connecté au flux.

Autres opérations : Dans l'afficheur des résultats, vous pouvez effectuer les opérations suivantes :

- Triez la grille de sortie sur la base de la colonne Champ.
- Exportez la sortie vers un fichier HTML. Pour plus d'informations, voir «Exportation des sorties», à la page 323.

Noeud Statistiques

Le noeud Statistiques fournit des informations récapitulatives de base sur les champs numériques. Ces statistiques peuvent porter sur des champs individuels et sur les corrélations entre les champs.

Noeud Statistiques - Onglet Paramètres

Examiner. Sélectionnez les champs sur lesquels obtenir des statistiques récapitulatives individuelles. Vous pouvez sélectionner plusieurs champs.

Statistiques. Sélectionnez les statistiques à créer. Les options disponibles sont les suivantes : **Comptage**, **Moyenne**, **Somme**, **Minimum**, **Maximum**, **Intervalle**, **Variance**, **Ecart-type**, **Erreur standard de la moyenne**, **Médiane** et **Mode**.

Corréler. Sélectionnez les champs à mettre en corrélation. Vous pouvez sélectionner plusieurs champs. Lorsque vous sélectionnez des champs de corrélation, la corrélation entre chaque champ Examiner et les champs de corrélation est indiquée dans la sortie.

Paramètres de corrélation. Vous pouvez définir les options d'affichage de la force des corrélations dans la sortie.

Paramètres de corrélation

IBM SPSS Modeler permet de définir les corrélations à l'aide de libellés descriptives afin de mettre en évidence des relations importantes. La **corrélation** mesure la force de la relation entre deux champs continus (intervalle numérique). Ses valeurs sont comprises entre –1.0 et 1.0. Les valeurs proches de +1,0 indiquent une association positive forte, de sorte que les valeurs élevées d'un champ sont associées aux valeurs élevées d'un autre champ, et les valeurs faibles du champ aux valeurs faibles de l'autre champ. Les valeurs proches de -1 indiquent une association négative forte, de sorte que les valeurs élevées d'un champ sont associées aux valeurs faibles de l'autre, et inversement. Les valeurs proches de 0,0 indiquent une association faible, de sorte que les valeurs des deux champs sont plus ou moins indépendantes.

A partir de la boîte de dialogue Paramètres de corrélation, vous pouvez contrôler l'affichage des libellés de corrélation, modifier les seuils définissant les catégories et modifier les libellés utilisées pour chaque intervalle. Etant donné que la manière dont vous définissez les valeurs de corrélation dépend essentiellement du type de problème, vous pouvez personnaliser les intervalles et les libellés en fonction de votre situation.

Afficher les libellés de force de corrélation en sortie. Par défaut, cette option est sélectionnée. Désélectionnez-la pour ne pas insérer les libellés descriptifs dans la sortie.

Force de corrélation. Deux options permettent de définir et d'étiqueter la force des corrélations :

- **Définir la force de corrélation par importance (1-p).** Applique un libellé aux corrélations en fonction de leur importance, cette dernière étant égale à 1 moins la signification, ou à 1 moins la probabilité que la différence de moyenne ne soit due qu'au hasard. Plus cette valeur est proche de 1, plus la probabilité que les deux champs ne soient *pas* indépendants (en d'autres termes, qu'une relation existe entre eux) est forte. En général, il est recommandé d'étiqueter les corrélations en fonction de leur importance plutôt qu'en fonction des valeurs absolues, car cela rend compte de la variabilité des données. Par exemple, un coefficient de 0,6 peut s'avérer très significatif dans un le jeu de données et

pas du tout dans un autre. Par défaut, les valeurs d'importance comprises entre 0 et 0,9 sont repérées par un libellé *Faible*, celles entre 0,9 et 0,95 par un libellé *Moyen*, et celles entre 0,95 et 1 par un libellé *Elevé*.

- **Définir la force de corrélation par valeur absolue.** Applique un libellé aux corrélations en fonction de la valeur absolue du coefficient de corrélation de Pearson, qui, comme indiqué précédemment, est comprise entre -1 et 1. Plus la valeur absolue de cette mesure est proche de 1, plus la corrélation est forte. Par défaut, les corrélations comprises entre 0 et 0,3333 (en valeur absolue) sont repérées par un libellé *Faible*, celles entre 0,3333 et 0,6666 par un libellé *Moyen*, et celles entre 0,6666 et 1 par un libellé *Elevé*. Toutefois, la signification d'une valeur donnée peut difficilement être généralisée d'un ensemble de donnée à un autre. C'est pourquoi, dans la plupart des cas, il est recommandé de définir des corrélations sur la base des probabilités et non des valeurs absolues.

Navigateur de sortie du noeud Statistiques

Le navigateur de sortie du noeud Statistiques affiche les résultats de l'analyse statistique et permet d'effectuer un certain nombre d'opérations : sélection de champs, création de noeuds en fonction d'une sélection, enregistrement et impression des résultats. Les options standard d'enregistrement, d'exportation et d'impression sont disponibles dans le menu Fichier, et celles d'édition dans le menu Edition. Pour plus d'informations, voir «Affichage de la sortie», à la page 321.

Lorsque vous accédez pour la première fois à la sortie du noeud Statistiques, les résultats sont développés. Pour masquer les résultats après les avoir consultés, utilisez la commande de développement située à gauche des résultats à masquer ou cliquez sur le bouton **Réduire tout** pour réduire tous les résultats. Pour afficher de nouveau les résultats, utilisez la commande de développement située à gauche des résultats à afficher ou cliquez sur le bouton **Développer tout** pour développer tous les résultats.

La sortie comporte une section pour chaque champ *Examiner*, contenant un tableau des statistiques demandées.

- **Comptage.** Indique le nombre d'enregistrements contenant des valeurs valides pour le champ.
- **Moyenne.** Indique la valeur moyenne du champ sur l'ensemble des enregistrements.
- **Somme.** Indique la somme des valeurs du champ sur l'ensemble des enregistrements.
- **Min.** Indique la valeur minimale du champ.
- **Max.** Indique la valeur maximale du champ.
- **Intervalle.** Indique la différence entre les valeurs minimale et maximale.
- **Variance.** Mesure de la variabilité des valeurs d'un champ. Pour la calculer, prenez la différence entre chaque valeur et la moyenne globale, élevez-la au carré, additionnez toutes les valeurs et divisez la somme par le nombre d'enregistrements.
- **Ecart type.** Autre mesure de la variabilité des valeurs d'un champ, correspondant à la racine carrée de la variance.
- **Erreur standard de la moyenne.** Mesure de la part d'incertitude dans l'estimation de la moyenne d'un champ si cette moyenne est censée s'appliquer à de nouvelles données.
- **Médiane.** Valeur médiane du champ, c'est-à-dire valeur qui sépare la première moitié des données de la seconde moitié (sur la base des valeurs du champ).
- **Mode.** Valeur unique la plus courante dans les données.

Corrélations. Si vous avez spécifié des champs de corrélation, la sortie contient également une section indiquant la corrélation de Pearson entre le champ Examiner et chaque champ de corrélation, ainsi que des libellés descriptifs facultatifs pour les valeurs de corrélation. Pour plus d'informations, voir «Paramètres de corrélation», à la page 343.

Menu Générer. Le menu Générer contient des options permettant de générer des noeuds.

- **Filtrer.** Permet de générer un noeud Filtrer, afin d'éliminer les champs non corrélés ou faiblement corrélés aux autres champs.

Génération d'un noeud Filtrer à partir de Statistics

Le noeud Filtrer, généré à partir du navigateur de sortie du noeud Statistiques, filtre les champs en fonction de leurs corrélations avec d'autres champs. Il trie les corrélations dans l'ordre de leur valeur absolue, prend les corrélations les plus grandes (selon le critère défini dans la boîte de dialogue Générer le filtre à partir des statistiques) et crée un filtre qui utilise tous les champs apparaissant dans l'une de ces corrélations.

Mode. Permet de définir le mode de sélection des corrélations. **Enlever** conserve les champs apparaissant dans les corrélations spécifiées. **Isoler** filtre les champs.

Inclure/exclure les champs apparaissant dans. Définissez le critère de sélection des corrélations.

- **Plus grand nombre de corrélations.** Sélectionne le nombre indiqué de corrélations et inclut/exclut les champs qui y apparaissent.
- **Plus grand pourcentage de corrélations (%).** Sélectionne le pourcentage spécifié (*n* %) de corrélations et inclut/exclut les champs qui apparaissent dans ces corrélations.
- **Corrélations supérieures à.** Sélectionne les corrélations supérieures, en valeur absolue, au seuil indiqué.

Noeud Moyennes

Le noeud Moyennes compare les moyennes de groupes indépendants ou de paires de champs associés, afin de détecter toute différence sensible. Par exemple, vous pouvez comparer les revenus moyens avant et après l'application d'une promotion, ou les revenus des clients qui ont et qui n'ont pas bénéficié de cette promotion.

Deux méthodes de comparaison des moyennes s'offrent à vous, selon vos données :

- **Entre groupes au sein d'un champ.** Pour comparer des groupes indépendants, sélectionnez un champ de test et un champ de regroupement. Par exemple, vous pouvez exclure un échantillon de clients "représentatifs" dans le cas d'une offre de promotion et comparer la moyenne des revenus de ce groupe avec celle de tous les autres clients. Dans ce cas, vous spécifiez un champ de test unique indiquant les revenus de chaque client, ainsi qu'un champ indicateur ou un champ nominal précisant si chacun a bénéficié de l'offre. Les échantillons sont indépendants dans le sens où chaque enregistrement est affecté à un groupe ou à un autre. Il est, en outre, impossible de lier un membre d'un groupe à un membre d'un autre groupe. Vous pouvez également définir un champ nominal comportant plus de deux valeurs pour comparer la moyenne de plusieurs groupes. Lorsqu'il est exécuté, le noeud effectue un test ANOVA unilatéral sur les champs sélectionnés. S'il n'existe que deux groupes de champs, les résultats du test ANOVA unilatéral sont globalement identiques à ceux d'un test *t* pour échantillons indépendants. Pour plus d'informations, voir «Comparaison des moyennes de groupes indépendants».
- **Entre paires de champs.** Lorsque vous comparez la moyenne de deux champs liés, vous devez réunir les groupes par paires pour que les résultats soient significatifs. Par exemple, vous pouvez comparer le revenu moyen d'un même groupe de clients avant et après l'application d'une promotion, ou bien encore les taux d'utilisation d'un service dans les paires époux-épouse pour voir s'il y a des différences. Chaque enregistrement contient deux mesures distinctes mais liées pouvant être comparées de manière significative. Lorsqu'il est exécuté, le noeud effectue un test *t* pour paires d'échantillons sur chaque paire de champs sélectionnée. Pour plus d'informations, voir «Comparaison de moyennes entre paires de champs», à la page 346.

Comparaison des moyennes de groupes indépendants

Sélectionnez **Entre groupes au sein d'un champ** dans le noeud Moyennes pour comparer la moyenne d'au moins deux groupes indépendants.

Champ de regroupement. Sélectionnez un champ indicateur ou un champ nominal comportant au moins deux valeurs distinctes et répartissant les enregistrements entre les différents groupes à comparer

(personnes qui ont bénéficié d'une offre et personnes qui n'en n'ont pas bénéficié, par exemple). Quel que soit le nombre de champs de test, vous ne pouvez sélectionner qu'un seul champ de regroupement.

Tester les champs. Sélectionnez un ou plusieurs champs numériques contenant les mesures à tester. Un test distinct est effectué pour chaque champ sélectionné. Par exemple, vous pouvez tester l'incidence que peut avoir une promotion sur l'utilisation, les revenus et l'attrition.

Comparaison de moyennes entre paires de champs

Sélectionnez **Entre paires de champs** dans le noeud Moyennes pour comparer la moyenne de différents champs. Ces champs doivent être liés d'une manière ou d'une autre pour que les résultats soient significatifs (revenus avant et après une promotion, par exemple). Vous pouvez également sélectionner plusieurs paires de champs.

Champ Un. Sélectionnez un champ numérique contenant la première des mesures à comparer. Dans une étude de type "avant-après", il s'agit du champ Avant.

Champ Deux. Sélectionnez le second champ à comparer.

Ajouter. Ajoute la paire sélectionnée à la liste Tester les paires de champs.

Si nécessaire, répétez les sélections de champ pour ajouter plusieurs paires à la liste.

Paramètres de corrélation. Permet de définir les options d'étiquetage de la force des corrélations. Pour plus d'informations, voir «Paramètres de corrélation», à la page 343.

Options du noeud Moyennes

L'onglet Options vous permet de définir les valeurs de seuil *p* utilisées pour étiqueter les résultats comme étant importants, marginaux ou non significatifs. Vous pouvez également éditer le libellé de chaque classement. L'importance est mesurée en termes de pourcentage et peut être définie globalement en soustrayant à 1 la probabilité d'obtention d'un résultat (la différence de moyenne entre deux champs, par exemple) aussi élevée ou plus élevée que le résultat généré de manière aléatoire. Par exemple, une valeur *p* supérieure à 0,95 indique une probabilité inférieure à 5% que le résultat soit dû exclusivement au hasard.

Libellés d'importance. Vous pouvez éditer les libellés servant à repérer chaque paire ou groupe de champs dans la sortie. Les libellés par défaut sont les suivantes : *Important*, *Marginal* et *Non significatif*.

Valeurs de césure. Indique le seuil de chaque rang. En général, les valeurs *p* supérieures à 0,95 sont considérées comme importantes et celles inférieures à 0,9 comme non significatives ; ces seuils peuvent être ajustés si nécessaire.

Remarque : Les mesures d'importance sont disponibles dans plusieurs noeuds. Les calculs possibles dépendent du noeud, ainsi que du type de la cible et des champs d'entrée utilisés, mais il est toujours possible de comparer les valeurs car toutes sont mesurées en pourcentage.

Navigateur de sortie du noeud Moyennes

Le navigateur de sortie du noeud Moyennes affiche les données sous forme de tableau croisé et vous permet d'exécuter des opérations standard : sélection et copie du tableau ligne par ligne, tri par colonne, et enregistrement et impression du tableau. Pour plus d'informations, voir «Affichage de la sortie», à la page 321.

Les informations particulières contenues dans le tableau dépendent du type de comparaison (groupes faisant partie d'un même champ ou de champs distincts).

Trier par. Permet de trier la sortie en fonction d'une colonne donnée. Cliquez sur la flèche vers le haut ou vers le bas pour modifier le sens du tri. Vous pouvez également cliquer sur l'en-tête de la colonne en fonction de laquelle réaliser le tri. (Pour modifier le sens du tri dans la colonne, cliquez à nouveau sur son en-tête.)

Afficher. Vous pouvez également choisir **Simple** ou **Options avancées** pour contrôler le niveau de détail de l'affichage. La vue avancée inclut toutes les informations de la vue simple, auxquelles elle ajoute des informations supplémentaires.

Sortie du noeud Moyennes par comparaison des groupes d'un champ

Lorsque vous comparez les groupes d'un champ, le nom du champ de regroupement apparaît au-dessus du tableau de sortie, et les moyennes et les statistiques liées sont indiquées séparément pour chaque groupe. Le tableau inclut une ligne distincte pour chaque champ de test.

Les colonnes suivantes apparaissent :

- **Champ.** Indique le nom des champs de test sélectionnés.
- **Moyennes par groupe.** Affiche la moyenne de chaque catégorie du champ de regroupement. Par exemple, vous pouvez comparer ceux qui ont bénéficié d'une offre spéciale (*Nouvelle promotion*) avec ceux qui n'en ont pas bénéficié (*Standard*). L'écart-type, l'erreur standard et le comptage sont également affichés dans la vue avancée.
- **Importance.** Affiche la valeur et le libellé d'importance. Pour plus d'informations, voir «Options du noeud Moyennes», à la page 346.

Sortie avancée

Dans la vue avancée, les colonnes supplémentaires suivantes apparaissent.

- **Test F.** Ce test est basé sur le rapport entre la variance entre les groupes et la variance au sein de chaque groupe. Si les moyennes sont identiques pour tous les groupes, le rapport *F* devrait être proche de 1, puisqu'il s'agit dans les deux cas de l'estimation de la même variance de population. Plus le rapport est élevé, plus la variation entre les groupes est importante et plus la probabilité d'une différence significative est forte.
- **df.** Affiche les degrés de liberté.

Sortie du noeud Moyennes par comparaison de paires de champs

Lorsque vous comparez des champs distincts, le tableau de sortie inclut une ligne pour chaque paire de champs sélectionnée.

- **Champ Un/Deux.** Affiche le nom des premier et second champs de chaque paire. L'écart-type, l'erreur standard et le comptage sont également affichés dans la vue avancée.
- **Moyenne Un/Deux.** Affiche la moyenne de chaque champ.
- **Corrélation.** Mesure la force de la relation entre deux champs continus (intervalle numérique). Les valeurs proches de +1.0 indiquent une association positive forte, et les valeurs proches de -1.0 indiquent une association négative forte. Pour plus d'informations, voir «Paramètres de corrélation», à la page 343.
- **Différence moyenne.** Affiche la différence entre les deux moyennes de champ.
- **Importance.** Affiche la valeur et le libellé d'importance. Pour plus d'informations, voir «Options du noeud Moyennes», à la page 346.

Sortie avancée

La sortie avancée ajoute les colonnes suivantes :

Intervalle de confiance de 95 %. Limites inférieure et supérieure de l'intervalle où la moyenne réelle est susceptible de figurer dans 95 % des échantillons possibles de cette taille au sein de cette population.

T-Test. La statistique *t* est obtenue en divisant la différence de moyenne par l'erreur standard correspondante. Plus la valeur absolue de cette statistique est élevée, plus la probabilité que les moyennes soient différentes est forte.

df. Affiche les degrés de liberté correspondant à la statistique.

Noeud Rapport

Ce noeud permet de créer des rapports formatés contenant du texte fixe et des données, ainsi que des expressions calculées à partir de ces données. Le format du rapport est déterminé par des modèles texte définissant la structure du texte fixe et de la sortie de données. Vous pouvez définir un formatage de texte personnalisé en utilisant des balises HTML dans le modèle et en définissant des options dans l'onglet Sortie. Les valeurs de données et autres sorties conditionnelles sont incluses dans le rapport à l'aide des expressions CLEM du modèle.

Alternatives au noeud Rapport

Le noeud Rapport est le plus souvent utilisé pour répertorier une sortie d'enregistrements ou d'observations d'un flux (par exemple, tous les enregistrements répondant à une certaine condition). A cet égard, il peut être considéré comme une alternative moins structurée au noeud Table.

- Si vous souhaitez un rapport répertoriant les informations de champ ou tout autre élément défini dans le flux plutôt que les données elles-mêmes (par exemple, des définitions de champ indiquées dans un noeud Typer), vous pouvez alors utiliser un script à la place.
- Pour générer un rapport incluant plusieurs objets de sortie (par exemple un ensemble de modèles, de tableaux et de graphiques générés par un ou plusieurs flux) et pouvant être créé sous plusieurs formats (texte, HTML et Microsoft Word/Office), vous pouvez utiliser un projet IBM SPSS Modeler.
- Pour produire une liste de noms de champ sans utiliser la génération de scripts, vous pouvez utiliser un noeud Table précédé d'un noeud Echantillonner qui supprime tous les enregistrements. Cette opération entraîne la création d'un tableau sans lignes, qui peut être transposé lors de l'exportation afin de produire une liste de noms de champ dans une seule colonne. (Pour ce faire, sélectionnez **Transposer les données** dans l'onglet Sortie du noeud Table.)

Noeud Rapport - Onglet Modèle

Création d'un modèle. Pour définir le contenu du rapport, vous devez créer un modèle dans l'onglet Modèle du noeud Rapport. Ce modèle se compose de lignes de texte définissant chacune un aspect du contenu du rapport, et de lignes de balises indiquant la portée de chaque ligne. Encadrées de crochets ([]), les expressions CLEM des lignes de contenu sont évaluées avant l'écriture de la ligne dans le rapport. Les portées suivantes sont disponibles pour les lignes du modèle :

Colonne fixe. Les lignes qui ne portent aucune indication sont considérées comme fixes. Les lignes fixes ne sont copiées qu'une fois dans le rapport, après l'évaluation des éventuelles expressions qu'elles contiennent. Par exemple, la ligne :

```
Voici mon rapport, imprimé le [@TODAY]
```

entraîne la copie dans le rapport d'une ligne contenant le texte indiqué et la date actuelle.

Global (Itérer TOUT). Les lignes figurant entre les balises spéciales `#ALL` et `#` sont copiées dans le rapport une fois pour chaque enregistrement contenu dans les données d'entrée. Les expressions CLEM (entre crochets) sont évaluées en fonction de l'enregistrement actuel de chaque ligne de sortie. Par exemple, les lignes :

```
#ALL
For record [@INDEX], the value of AGE is [AGE]
#
```

copient une ligne par enregistrement, indiquant le numéro de l'enregistrement et l'âge.

Pour générer la liste de tous les enregistrements :

```
#ALL
[Age] [Sexe] [Cholestérol]  [BP]
#
```

Conditionnel (Itérer SI). Les lignes figurant entre les balises spéciales `#WHERE <condition>>` et `#` sont copiées dans le rapport une fois pour chaque enregistrement pour lequel la condition spécifiée a la valeur true (vrai). La condition est une expression CLEM. (Dans la condition `WHERE`, les crochets sont facultatifs.) Par exemple, les lignes :

```
#WHERE [SEX = 'M']
Male at record no. [@INDEX] has age [AGE].
#
```

copient dans le fichier une ligne pour chaque enregistrement dans lequel le sexe a pour valeur *M*. Le rapport complet contient les lignes fixes, globales et conditionnelles définies après l'application du modèle aux données d'entrée.

Dans l'onglet Sortie, vous pouvez spécifier les options d'affichage ou d'enregistrement de résultats, qui seront communes à plusieurs types de noeud de sortie. Pour plus d'informations, voir «Noeud de sortie - Onglet Sortie», à la page 324.

Sortie des données au format HTML ou XML

Vous pouvez inclure des balises HTML ou XML directement dans le modèle pour générer des rapports dans l'un ou l'autre des formats. Par exemple, le modèle suivant génère un tableau HTML.

```
This report is written in HTML.
Seuls les enregistrements où Age est supérieur à 60 sont inclus.

<HTML>
  <TABLE border="2">
   <TR>
     <TD>Age</TD>
     <TD>BP</TD>
     <TD>Cholesterol</TD>
     <TD>Drug</TD>
   </TR>

  #WHERE Age > 60
    <TR>
       <TD>[Age]</TD>
       <TD>[BP]</TD>
       <TD>[Cholesterol]</TD>
       <TD>[Drug]</TD>
     </TR>
  #
  </TABLE>
</HTML>
```

Navigateur de sortie du noeud Rapport

Ce navigateur affiche le contenu du rapport généré. Les options standard d'enregistrement, d'exportation et d'impression sont disponibles dans le menu Fichier, et celles d'édition dans le menu Edition. Pour plus d'informations, voir «Affichage de la sortie», à la page 321.

Noeud Valeurs globales

Le noeud Valeurs globales analyse les données et calcule des valeurs récapitulatives pouvant être utilisées dans des expressions CLEM. Par exemple, vous pouvez utiliser un noeud Valeurs globales afin de calculer les statistiques pour un champ appelé *âge*, puis utiliser la moyenne globale du champ *âge* dans des expressions CLEM en insérant la fonction `@GLOBAL_MEAN(âge)`.

Noeud Valeurs globales - Onglet Paramètres

Valeurs globales à créer. Sélectionnez les champs pour lesquels vous souhaitez que des valeurs globales soient disponibles. Vous pouvez sélectionner plusieurs champs. Pour chaque champ, indiquez les statistiques à calculer en les sélectionnant dans les colonnes en regard du nom du champ.

- **Moyenne.** Indique la valeur moyenne du champ sur l'ensemble des enregistrements.
- **Somme.** Indique la somme des valeurs du champ sur l'ensemble des enregistrements.
- **Minimum.** Indique la valeur minimale du champ.
- **Maximum.** Indique la valeur maximale du champ.
- **Ecart-type.** L'écart-type est une mesure de variabilité des valeurs d'un champ ; il correspond à la racine carrée de la variance.

Opération(s) par défaut. Les options sélectionnées ici sont utilisées lorsque de nouveaux champs sont ajoutés à la liste des valeurs globales ci-dessus. Pour modifier l'ensemble de statistiques par défaut, sélectionnez ou désélectionnez les statistiques de votre choix. Vous pouvez également utiliser le bouton **Appliquer** pour appliquer les options par défaut à tous les champs de la liste.

Remarque : Certaines opérations ne sont pas applicables à des champs non numériques (par exemple Somme pour un champ date/heure). Les opérations qui ne peuvent pas être utilisées avec un champ sélectionné sont désactivées.

Effacer toutes les valeurs globales avant l'exécution. Cette option permet de supprimer toutes les valeurs globales avant d'en calculer de nouvelles. Si cette option n'est pas sélectionnée, les valeurs recalculées remplacent les anciennes, mais les valeurs globales non recalculées restent disponibles.

Afficher un aperçu des valeurs globales créées avant l'exécution. Si vous sélectionnez cette option, l'onglet Valeurs globales de la boîte de dialogue des propriétés du flux apparaît à la fin de l'exécution pour afficher les valeurs globales calculées.

Noeud Ajustement de simulation

Le noeud Ajustement de simulation ajuste un ensemble de distributions statistiques candidates à chaque champ figurant dans les données. L'ajustement de chaque distribution à un champ est évalué à l'aide d'un critère de qualité d'ajustement. Lors de l'exécution d'un noeud Ajustement de simulation, un noeud Génération de simulation est créé (ou un noeud existant est mis à jour). La distribution d'ajustement la mieux adaptée est affectée à chaque champ. Le noeud Génération de simulation peut ensuite être utilisé pour générer des données simulées pour chaque champ.

Bien que le noeud Ajustement de simulation soit un noeud terminal, il n'ajoute pas de modèle à la palette des modèles générés, n'ajoute pas de sortie ou de graphique à l'onglet Sorties et n'exporte pas de données.

Remarque : Si les données d'historique sont éparpillées (à savoir, s'il existe beaucoup de valeurs manquantes), il peut s'avérer difficile pour le composant d'ajustement de trouver suffisamment de valeurs valides pour ajouter les distributions aux données. Lorsque les données sont éparpillées, vous devez, soit les supprimer avant l'ajustement si elles ne sont pas requises, soit imputer les valeurs manquantes. L'utilisation des options figurant sur l'onglet **Qualité** du noeud Audit des données permet d'afficher le nombre d'enregistrements complets, d'identifier les champs éparpillés et de sélectionner une méthode

d'imputation. Si le nombre d'enregistrements est insuffisant pour l'ajustement de distribution, vous pouvez utiliser un noeud équilibrer pour augmenter le nombre d'enregistrements.

Utilisation d'un noeud Ajustement de simulation pour créer automatiquement un noeud Génération de simulation

Lors de la première exécution du noeud Ajustement de simulation, un noeud Génération de simulation est créé avec un lien de mise à jour vers le noeud Ajustement de simulation. Si le noeud Ajustement de simulation est réexécuté, un nouveau noeud Génération de simulation sera créé uniquement si le lien de mise à jour a été supprimé. Un noeud Ajustement de simulation peut également être utilisé pour mettre à jour un noeud Génération de simulation connecté. Le résultat varie suivant si les mêmes champs sont présents ou non dans les deux noeuds et si les champs sont déverrouillés ou non dans le noeud Génération de simulation. Pour plus d'informations, voir «Noeud Génération de simulation», à la page 54.

Un noeud Ajustement de simulation ne peut contenir qu'un lien de mise à jour vers un noeud Génération de simulation. Pour définir un lien de mise à jour vers un noeud Génération de simulation, procédez comme suit :

1. Cliquez avec le bouton droit de la souris sur le noeud Ajustement de simulation.
2. Dans le menu, sélectionnez **Définir le lien Mettre à jour**.
3. Cliquez sur le noeud Génération de simulation vers lequel vous souhaitez définir un lien de mise à jour.

Pour supprimer un lien de mise à jour entre un noeud Ajustement de simulation et un noeud Génération de simulation, cliquez avec le bouton droit de la souris sur le lien de mise à jour et sélectionnez **Supprimer un lien**.

Ajustement de distribution

Une distribution statistique représente la fréquence théorique de l'occurrence des valeurs qu'une variable peut prendre. Dans le noeud Ajustement de simulation, un ensemble de distributions statistiques théoriques est comparé à chaque champ de données. Les distributions disponibles pour l'ajustement sont décrites dans la rubrique «Distributions», à la page 65. Les paramètres de la distribution théorique sont ajustés pour s'adapter au mieux aux données en fonction de la mesure de la qualité d'ajustement ; il s'agit du critère Anderson-Darling ou Kolmogorov-Smirnov. Les résultats de l'ajustement de distribution par le noeud Ajustement de simulation indiquent quelles distributions ont été ajustées, les meilleures estimations de paramètres pour chaque distribution et la qualité d'ajustement de chaque distribution aux données. Lors de l'ajustement de distribution, les corrélations entre les champs dotés d'un type de stockage numérique et les contingences entre les champs dotés d'une distribution catégorielle sont également calculées. Les résultats de l'ajustement de distribution sont utilisés pour créer un noeud Génération de simulation.

Avant que les distributions soient ajustées à vos données, les 1 000 premiers enregistrements sont examinés à la recherche de valeurs manquantes. S'il manque trop de valeurs, l'ajustement de distribution n'est pas possible. Vous devez alors décider si l'une des options suivantes est adaptée :

- Utiliser un noeud en amont pour supprimer les enregistrements contenant des valeurs manquantes.
- Utiliser un noeud en amont pour imputer des valeurs pour les valeurs manquantes.

L'ajustement de distribution n'exclut pas les valeurs manquantes de l'utilisateur. Si vos données contiennent des valeurs manquantes de l'utilisateur et que vous souhaitez qu'elles soient exclues de l'ajustement de distribution, vous devez définir ces valeurs en tant que valeurs système manquantes.

Le rôle d'un champ n'est pas pris en compte lorsque les distributions sont ajustées. Par exemple, les champs dotés du rôle **Cible** sont traités de la même façon que ceux dotés des rôles **Entrée**, **Aucun**, **Les deux**, **Partition**, **Scission**, **Fréquence** et **ID**.

Les champs sont traités différemment au cours de l'ajustement de distribution, selon le type de stockage et le niveau de mesure. Le traitement des champs lors de l'ajustement de distribution est décrit dans le tableau ci-dessous.

Tableau 45. Ajustement de distribution en fonction du type de stockage et du niveau de mesure des champs

<table>
<tr><th rowspan="2">Type de stockage</th><th colspan="6">Niveau de mesure</th></tr>
<tr><td>Continu</td><td>Catégorielle</td><td>Indicateur</td><td>Nominal</td><td>Ordinal</td><td>Sans type</td></tr>
<tr><td>Chaîne</td><td>Irréalisable</td><td colspan="3">Les distributions catégorielles, dé à jouer et fixes sont ajustées</td><td rowspan="9">Le champ est ignoré et n'est pas transmis au noeud Génération de simulation.</td></tr>
<tr><td>Entier</td><td rowspan="6">Toutes les distributions sont ajustées. Les corrélations et contingences sont calculées.</td><td colspan="2" rowspan="6">La distribution catégorielle est ajustée. Les corrélations ne sont pas calculées.</td><td rowspan="6">Les distributions binomiales, binomiales négatives et de Poisson sont ajustées et les corrélations sont calculées.</td></tr>
<tr><td>Réel</td></tr>
<tr><td>Heure</td></tr>
<tr><td>Date</td></tr>
<tr><td>Horodatage</td></tr>
<tr><td>Inconnu</td><td colspan="4">Le type de stockage approprié est déterminé à partir des données.</td></tr>
</table>

Les champs dotés du niveau de mesure ordinal sont traités comme des champs continus et sont inclus dans la table de corrélations dans le noeud Génération de simulation. Si vous voulez qu'une distribution qui n'est ni binomiale, ni binomiale négative, ni de Poisson soit ajustée sur un champ ordinal, vous devez modifier le niveau de mesure du champ en continu. Si vous avez défini au préalable un libellé pour chaque valeur d'un champ ordinal et que vous modifiez ensuite le niveau de mesure en continu, les libellés seront perdus.

Les champs dotés de valeurs uniques ne sont pas traités différemment des champs contenant plusieurs valeurs lors de l'ajustement de distribution. Les champs dont le type de stockage est Heure, Date ou Horodatage sont traités comme des champs numériques.

Ajustement des distributions aux champs de scission

Si vos données contiennent un champ de scission et que vous souhaitez que l'ajustement de distribution soit réalisé séparément pour chaque scission, vous devez transformer les données à l'aide d'un noeud Restructurer en amont. A l'aide du noeud Restructurer, générez un nouveau champ pour chaque valeur du champ de scission. Ces données restructurées peuvent ensuite être utilisées pour l'ajustement de distribution dans le noeud Ajustement de simulation.

Noeud Ajustement de simulation - Onglet Paramètres

Nom de noeud source. Vous pouvez créer automatiquement le nom du noeud Génération de simulation généré (ou mis à jour) en sélectionnant **Automatique**. Le nom généré automatiquement est le nom spécifié dans le noeud Ajustement de simulation si un nom personnalisé a été spécifié (ou Génération de simulation si aucun nom personnalisé n'a été indiqué dans le noeud Ajustement de simulation). Sélectionnez **Personnalisé** pour spécifier un nom personnalisé dans le champ de texte adjacent. Si le champ de texte n'est pas modifié, le nom personnalisé par défaut est Génération de simulation.

Options d'ajustement : Ces options permettent d'indiquer le mode d'ajustement des distributions aux champs et l'évaluation de l'ajustement des distributions.

- **Nombre d'observations à échantillonner.** Il s'agit du nombre d'observations à utiliser lors de l'ajustement des distributions aux champs dans le jeu de données. Sélectionnez **Toutes les observations** pour ajuster les distributions à la totalité des enregistrements figurant dans les données. Si votre jeu de données est très volumineux, envisagez de limiter le nombre d'observations utilisées pour l'ajustement de distribution. Sélectionnez **Limiter aux N premières observations** pour n'utiliser que les N premières observations. Cliquez sur les flèches pour spécifier le nombre d'observations à utiliser. Vous pouvez également utiliser un noeud en amont pour prendre un échantillon aléatoire d'enregistrements pour l'ajustement de distribution.
- **Critère de la qualité d'ajustement (champs continus uniquement).** Pour les champs continus, sélectionnez le test Anderson-Darling ou Kolmogorov-Smirnoff pour tester la qualité d'ajustement aux distributions de rang lors de l'ajustement des distributions aux champs. Le test Anderson-Darling est sélectionné par défaut et est particulièrement recommandé si vous souhaitez garantir le meilleur ajustement possible dans les zones de fin. Les deux statistiques sont calculées pour chaque distribution candidate, mais seule la statistique sélectionnée est utilisée pour ordonner les distributions et déterminer la meilleure distribution d'ajustement.
- **Intervalles (distribution empirique uniquement).** Pour les champs continus, la distribution empirique est la fonction de distribution cumulée des données d'historique. Il s'agit de la probabilité de chaque valeur, ou plage de valeurs, qui est dérivée directement des données. Vous pouvez spécifier le nombre d'intervalles utilisés pour le calcul de la distribution empirique des champs continus en cliquant sur les flèches. La valeur par défaut est 100 et la valeur maximale, 1 000.
- **Champ de pondération (facultatif).** Si votre jeu de données contient un champ de pondération, cliquez sur l'icône de sélection des champs et sélectionnez le champ de pondération dans la liste. Le champ de pondération est ensuite exclu du processus d'ajustement de distribution. La liste affiche tous les champs du jeu de données dont le niveau de mesure est continu. Vous ne pouvez sélectionner qu'un seul champ de pondération.

Noeud Evaluation de simulation

Le noeud Evaluation de simulation est un noeud terminal qui évalue un champ spécifié, fournit une distribution du champ et crée des graphiques représentant les distributions et corrélations. Ce noeud est principalement utilisé pour évaluer les champs continus. Il vient donc compléter le graphique d'évaluation, qui est généré par un noeud Evaluation et est utile pour l'évaluation des champs discrets. Une autre différence réside dans le fait que le noeud Evaluation de simulation évalue une seule prévision sur plusieurs itérations, tandis que le noeud Evaluation évalue plusieurs prévisions, chacune avec une seule itération. Les itérations sont générées lorsque plusieurs valeurs sont indiquées pour un paramètre de distribution dans le noeud Génération de simulation. Pour plus d'informations, voir «Itérations», à la page 64.

Le noeud Evaluation de simulation est conçu pour une utilisation avec les données générées par les noeuds Ajustement de simulation et Génération de simulation. Toutefois, il peut être utilisé avec tout autre noeud. Vous pouvez insérer un nombre quelconque d'étapes de traitement entre le noeud Génération de simulation et le noeud Evaluation de simulation.

Important : Le noeud Evaluation de simulation requiert un minimum de 1 000 enregistrements comportant des valeurs valides pour le champ cible.

Noeud Evaluation de simulation - Onglet Paramètres

L'onglet Paramètres du noeud Evaluation de simulation vous permet de spécifier le rôle de chaque champ dans votre jeu de données et de personnaliser la sortie générée par la simulation.

Sélectionner un élément. Permet de basculer entre les trois vues du noeud Evaluation de simulation : Champs, Fonctions de densité et Sorties.

Vue Champs

Champ cible. Il s'agit d'un champ obligatoire. Cliquez sur la flèche pour sélectionner le champ cible de votre jeu de données dans la liste déroulante. Le champ sélectionné peut comporter un niveau de mesure continu, ordinal ou nominal, mais pas de date, ni de niveau de mesure non spécifié.

Champ d'itération (facultatif). Si vos données contiennent un champ d'itération indiquant à quelle itération appartient chaque enregistrement figurant dans vos données, vous devez le sélectionner ici. Cela signifie que chaque itération sera évaluée séparément. Seuls les champs dotés d'un niveau de mesure continu, ordinal ou nominal peuvent être sélectionnés.

Les données d'entrée sont déjà triées par itération. Cette option est activée uniquement si un champ d'itération est spécifié dans **Champ d'itération (facultatif)**. Sélectionnez cette option uniquement si vous êtes sûr que vos données d'entrée sont déjà triées sur le champ d'itération spécifié dans **Champ d'itération (facultatif)**.

Nombre maximal d'itérations à représenter graphiquement. Cette option est activée uniquement si un champ d'itération est spécifié dans **Champ d'itération (facultatif)**. Cliquez sur les flèches pour spécifier le nombre d'itérations à représenter graphiquement. La définition de ce nombre permet d'éviter de représenter un trop grand nombre d'itérations sur un seul graphique, ce qui le rendrait difficilement interprétable. Le niveau le plus bas pour le nombre maximal d'itérations est 2, tandis que le plus élevé est 50. Le nombre maximal d'itérations à représenter graphiquement est initialement défini sur 10.

Champs d'entrée pour la corrélation tornado. Le graphique de corrélation tornado est un graphique à barres qui affiche les coefficients de corrélation entre la cible spécifiée et chacune des entrées indiquées. Cliquez sur l'icône de sélection des champs pour sélectionner dans la liste des entrées simulées disponibles les champs d'entrée à inclure dans le graphique tornado. Seuls les champs d'entrée dotés de niveaux de mesure continus et ordinaux peuvent être sélectionnés. Les champs d'entrée nominaux, sans type et de date ne sont pas disponibles dans la liste et ne peuvent pas être sélectionnés.

Vue Fonctions de densité

Les options de cette vue permettent de personnaliser la sortie pour les fonctions de densité de probabilité et les fonctions de distribution cumulée pour les cibles continues, ainsi que les graphiques à barres des valeurs prédites pour les cibles catégorielles.

Fonctions de densité. Les fonctions de densité sont les principaux moyens permettant de sonder l'ensemble de résultats généré par votre simulation.

- **Fonction de densité de probabilité (PDF).** Sélectionnez cette option pour générer une fonction de densité de probabilité pour le champ cible. La fonction de densité de probabilité affiche la distribution des valeurs cible. Vous pouvez utiliser la fonction de densité de probabilité pour déterminer la probabilité que la cible se trouve dans une zone spécifique. Pour les cibles catégorielles (cibles dotées d'un niveau de mesure nominal ou ordinal), un graphique à barres affichant le pourcentage de cas compris dans chaque catégorie de la cible est généré.

- **Fonction de distribution cumulée (CDF).** Sélectionnez cette option pour générer une fonction de distribution cumulée pour le champ cible. La fonction de distribution cumulée affiche la probabilité que la valeur de la cible soit inférieure ou égale à une valeur donnée. Elle est disponible uniquement pour les variables continues.

Lignes de référence (continues). Ces options sont activées si les options **Fonction de densité de probabilité (PDF)** et/ou **Fonction de distribution cumulée (CDF)** sont sélectionnées. Elles permettent d'ajouter diverses lignes de référence verticale fixe aux fonctions de densité de probabilité et aux fonctions de distribution cumulée.

- **Moyenne.** Sélectionnez cette option pour ajouter une ligne de référence au niveau de la valeur moyenne du champ cible.
- **Médiane.** Sélectionnez cette option pour ajouter une ligne de référence au niveau de la valeur médiane du champ cible.
- **Ecarts types.** Sélectionnez cette option pour ajouter des lignes de référence de plus ou moins un nombre donné d'écarts types par rapport à la moyenne du champ cible. Cette option permet d'activer le champ adjacent **Nombre**. Cliquez sur les flèches pour spécifier le nombre d'écarts types. Le nombre minimal d'écarts types est 1 et le nombre maximal, 10. Le nombre d'écarts types est initialement défini sur 3.
- **Percentiles.** Sélectionnez cette option pour ajouter des lignes de référence au niveau de deux valeurs de centile de la distribution du champ cible. Cette option permet d'activer les champs de texte adjacents **Bas** et **Haut**. Par exemple, si vous saisissez la valeur 90 dans le champ de texte **Haut**, vous ajoutez une ligne de référence au niveau du 90ème percentile de la cible, qui représente la valeur au dessous de laquelle se trouvent 90 % des observations. De même, la valeur 10 dans le champ de texte **Haut** représente le dixième percentile de la cible, qui correspond à la valeur au dessous de laquelle se trouvent 10 % des observations.
- **Lignes de référence personnalisées.** Sélectionnez cette option pour ajouter des lignes de référence au niveau de certaines valeurs situées sur l'axe horizontal. Cette option permet d'activer la table adjacente **Valeurs**. Chaque fois que vous saisissez un nombre valide dans la table **Valeurs**, une nouvelle ligne vide est ajoutée au bas de la table. Un nombre *valide* est un nombre compris dans la plage de valeurs du champ cible.

Remarque : Lorsque plusieurs fonctions de densité ou fonctions de distribution (provenant de plusieurs itérations) s'affichent sur un seul graphique, les lignes de référence (autres que les lignes personnalisées) sont appliquées séparément à chaque fonction.

Cible catégorielle (PDF uniquement). Ces options sont activées uniquement si l'option **Fonction de densité de probabilité (PDF)** est sélectionnée.

- **Valeurs de catégorie à rapporter.** Pour les modèles avec des champs cible catégoriels, le résultat du modèle est un ensemble de probabilités prédites, une pour chaque catégorie, que la valeur cible se situe dans chaque catégorie. La catégorie avec la probabilité la plus élevée est considérée comme la modalité estimée et utilisée pour générer le graphique à barres pour la fonction de densité de probabilité. Sélectionnez **Modalité estimée** pour générer le graphique à barres. Sélectionnez **Probabilités prédites** pour générer des histogrammes de la distribution des probabilités prédites pour chaque catégorie du champ cible. Vous pouvez également sélectionner **Les deux** pour générer les deux types de graphique.
- **Regroupement pour l'analyse de sensibilité.** Les simulations incluant des itérations d'analyse de sensibilité génèrent un champ cible indépendant (ou un champ cible prédit à partir d'un modèle) pour chaque itération définie par l'analyse. Il existe une itération pour chaque valeur du paramètre de distribution en cours de variation. Lorsque des itérations sont présentes, le graphique à barres de la modalité estimée pour un champ cible catégoriel s'affiche sous forme de graphique à barres en cluster incluant les résultats de toutes les itérations. Sélectionnez **Regrouper les catégories** ou **Regrouper les itérations**.

Vue Sorties

Valeurs de centile des distributions cible. Ces options permettent de créer une table de valeurs de centile des distributions cible et de spécifier les percentiles à afficher.

Créer une table de valeurs de centile. Pour les champs cible continus, sélectionnez cette option pour générer une table des percentiles spécifiés des distributions cible. Choisissez l'une des options suivantes pour spécifier les percentiles :

- **Quartiles.** Les quartiles sont les 25ème, 50ème et 75ème percentiles de la distribution appliquée au champ cible. Les observations sont subdivisées en quatre groupes de même taille.
- **Intervalles.** Si vous voulez un nombre égal de groupes différent de quatre, sélectionnez **Intervalles**. Cette option permet d'activer le champ adjacent **Nombre**. Cliquez sur les flèches pour spécifier le nombre d'intervalles. Le nombre minimal d'intervalles est 2 et le nombre maximal, 100. Le nombre d'intervalles est initialement défini sur 10.
- **Percentiles personnalisés.** Sélectionnez cette option pour spécifier des percentiles individuels, par exemple, le 99ème percentile. Cette option permet d'activer la table adjacente **Valeurs**. Chaque fois que vous saisissez un nombre valide compris entre 1 et 100 dans la table **Valeurs**, une nouvelle ligne vide est ajoutée au bas de la table.

Sortie du noeud Evaluation de simulation

Lors de l'exécution du noeud Evaluation de simulation, la sortie est ajoutée au gestionnaire des sorties. Le navigateur de sortie du noeud Evaluation de simulation affiche les résultats de l'exécution du noeud Evaluation de simulation. Les options standard d'enregistrement, d'exportation et d'impression sont disponibles dans le menu **Fichier**, et celles d'édition dans le menu **Edition**. Pour plus d'informations, voir «Affichage de la sortie», à la page 321. Le menu **Vue** est uniquement activé si l'un des graphiques est sélectionné. Il n'est pas activé pour la table de distribution, ni pour les sorties d'information. Le menu **Vue** vous permet de sélectionner **Mode Edition** pour modifier la mise en forme et l'aspect du graphique, ou **Mode Exploration** pour explorer les données et valeurs représentées par le graphique. Le mode statique conserve la position actuelle des lignes de référence (et des curseurs) du graphique, empêchant ainsi de les déplacer. Il s'agit du seul mode qui vous permet de copier, d'imprimer ou d'exporter le graphique avec ses lignes de référence. Pour sélectionner ce mode, cliquez sur **Mode statique** dans le menu **Vue**.

Le navigateur de sortie Evaluation de simulation se compose de deux panneaux. A gauche de la fenêtre, un panneau de navigation affiche les représentations miniatures des graphiques qui ont été générés lors de l'exécution du noeud Evaluation de simulation. Lorsqu'une miniature est sélectionnée, la sortie du graphique s'affiche sur le panneau situé à droite de la fenêtre.

Panneau de navigation

Le panneau de navigation du navigateur de sortie contient des miniatures des graphiques qui sont générés par une simulation. Les miniatures affichées sur le panneau de navigation dépendent du niveau de mesure du champ cible et des options sélectionnées dans la boîte de dialogue du noeud Evaluation de simulation. Les miniatures sont décrites dans le tableau ci-dessous.

Tableau 46. Miniatures du panneau de navigation

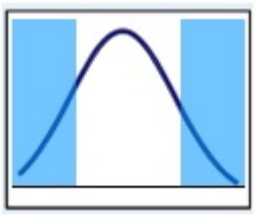
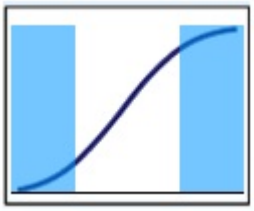
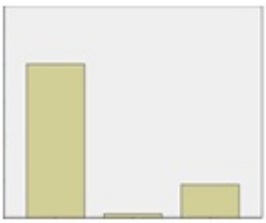
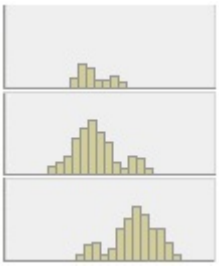
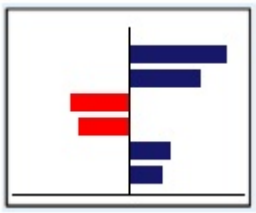
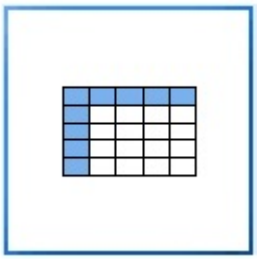
Miniature	Description	Commentaires
	Fonction de densité de probabilité	Cette miniature s'affiche uniquement si le niveau de mesure du champ cible est continu et si l'option **Fonction de densité de probabilité (PDF)** est sélectionnée dans la vue Fonctions de densité de la boîte de dialogue du noeud Evaluation de simulation. Si le niveau de mesure du champ cible est catégoriel, cette miniature ne s'affiche pas.
	Fonction de distribution cumulée	Cette miniature s'affiche uniquement si le niveau de mesure du champ cible est continu et si l'option **Fonction de distribution cumulée (CDF)** est sélectionnée dans la vue Fonctions de densité de la boîte de dialogue du noeud Evaluation de simulation. Si le niveau de mesure du champ cible est catégoriel, cette miniature ne s'affiche pas.
	Valeurs de modalité estimée	Cette miniature s'affiche uniquement si le niveau de mesure du champ cible est catégoriel, si l'option **Fonction de densité de probabilité (PDF)** est sélectionnée dans la vue Fonctions de densité de la boîte de dialogue du noeud Evaluation de simulation et si **Modalité estimée** ou **Les deux** est sélectionné dans la zone **Valeurs de catégorie à rapporter**. Si le niveau de mesure du champ cible est continu, cette miniature ne s'affiche pas.
	Probabilités de modalité estimée	Cette miniature s'affiche uniquement si le niveau de mesure du champ cible est catégoriel, si l'option **Fonction de densité de probabilité (PDF)** est sélectionnée dans la vue Fonctions de densité de la boîte de dialogue du noeud Evaluation de simulation et si **Probabilités prédites** ou **Les deux** est sélectionné dans la zone **Valeurs de catégorie à rapporter**. Si le niveau de mesure du champ cible est continu, cette miniature ne s'affiche pas.

Tableau 46. Miniatures du panneau de navigation (suite)

Miniature	Description	Commentaires
	Graphiques Tornado	Cette miniature s'affiche uniquement si un ou plusieurs champs d'entrée sont sélectionnés dans **Champs d'entrée pour la corrélation tornado** dans la vue Champs de la boîte de dialogue du noeud Evaluation de simulation.
	Table de distribution	Cette miniature s'affiche uniquement si le niveau de mesure du champ cible est continu et si l'option **Créer une table de valeurs de centile** est sélectionnée dans la vue Sorties de la boîte de dialogue du noeud Evaluation de simulation. Le menu **Vue** est désactivé pour ce graphique. Si le niveau de mesure du champ cible est catégoriel, cette miniature ne s'affiche pas.
	Informations	Cette miniature s'affiche en permanence. Le menu **Vue** est désactivé pour cette sortie.

Sortie de graphique

Les types de graphiques de sortie disponibles dépendent du niveau de mesure du champ cible, de l'utilisation ou non d'un champ d'itération et des options sélectionnées dans la boîte de dialogue du noeud Evaluation de simulation. Parmi les graphiques générés par une simulation, un certain nombre comporte des fonctions interactives que vous pouvez utiliser pour personnaliser l'affichage. Pour accéder aux fonctions interactives, cliquez sur **Options de graphique**. Tous les graphiques de simulation sont des visualisations graphiques.

Graphiques de fonction de densité de probabilité pour les cibles continues. Ce type de graphique affiche à la fois la probabilité et la fréquence, l'échelle de probabilité se trouvant sur l'axe vertical gauche et l'échelle de fréquence sur l'axe vertical droit. Le graphique contient également deux lignes de référence verticale déplaçables qui le divisent en zones distinctes. La table située au dessous du graphique affiche le pourcentage de distribution dans chacune des zones. Si plusieurs fonctions de densité sont affichées sur le même graphique (en raison d'itérations), la table contient une ligne distincte pour les probabilités associées à chaque fonction de densité et une colonne supplémentaire contenant le nom de l'itération et la couleur associée à chaque fonction de densité. Les itérations sont répertoriées dans la table par ordre alphabétique, en fonction du libellé d'itération. Si aucun libellé d'itération n'est disponible, la valeur d'itération est utilisée à la place. La table n'est pas modifiable.

Chaque ligne de référence comporte un curseur (triangle inversé) que vous pouvez utiliser pour la déplacer facilement. Chaque curseur est doté d'un libellé indiquant sa position actuelle. Par défaut, les curseurs se trouvent aux 5ème et 95ème percentiles de la distribution. S'il existe plusieurs itérations, les curseurs se trouvent aux 5ème et 95ème percentiles de la première itération répertoriée dans la table. Vous ne pouvez pas déplacer les lignes afin qu'elles se croisent.

Pour accéder à un certain nombre de fonctions supplémentaires, cliquez sur **Options de graphique**. Vous pouvez notamment définir explicitement les positions des curseurs, ajouter des lignes de référence fixe et modifier la vue du graphique afin qu'elle prenne la forme d'un histogramme plutôt que d'une courbe continue. Pour plus d'informations, voir «Options de graphique», à la page 360. Cliquez avec le bouton droit de la souris sur le graphique pour le copier ou l'exporter.

Graphiques de fonction de distribution cumulée pour les cibles continues. Ce type de graphique contient les deux mêmes lignes de référence verticale déplaçables et la même table associée que celles décrites pour le graphique de fonction de densité de probabilité. Les boutons du curseur et la table se comportent de la même façon que dans la fonction de densité de probabilité lorsqu'il existe plusieurs itérations. Les mêmes couleurs sont utilisées pour les fonctions de distribution et pour identifier la fonction de densité appartenant à chaque itération.

Ce graphique vous donne également accès à la boîte de dialogue Options de graphique, qui vous permet de définir explicitement les positions des curseurs, d'ajouter des lignes de référence fixe et d'indiquer si la fonction de distribution cumulée est affichée sous forme de fonction croissante (paramètre par défaut) ou décroissante. Pour plus d'informations, voir «Options de graphique», à la page 360. Cliquez avec le bouton droit de la souris sur le graphique pour le copier, l'exporter ou l'éditer. Sélectionnez **Editer** pour ouvrir le graphique dans une fenêtre de l'éditeur de représentation graphique flottante.

Graphique de valeurs de modalité estimée pour les cibles catégorielles. Pour les champs cible catégoriels, un graphique à barres présente les valeurs prédites. Ces dernières s'affichent sous forme de pourcentage du champ cible qui doit être inclus dans chaque catégorie. Pour les champs cible catégoriels avec des itérations d'analyse de sensibilité, les résultats pour la catégorie cible prévue s'affichent sous forme de graphique à barres en cluster incluant les résultats de toutes les itérations. Le graphique est divisé en cluster par catégorie ou par itération, suivant l'option sélectionnée dans la zone **Regroupement pour l'analyse de sensibilité** dans la vue Fonctions de densité de la boîte de dialogue du noeud Evaluation de simulation. Cliquez avec le bouton droit de la souris sur le graphique pour le copier, l'exporter ou l'éditer. Sélectionnez **Editer** pour ouvrir le graphique dans une fenêtre de l'éditeur de représentation graphique flottante.

Graphique de probabilités de modalité estimée pour les cibles catégorielles. Pour les champs cible catégoriels, un histogramme affiche la distribution des probabilités prédites pour chaque catégorie de la cible. Pour les champs cible catégoriels avec des itérations d'analyse de sensibilité, les histogrammes sont affichés par catégorie ou par itération, suivant l'option sélectionnée dans la zone **Regroupement pour l'analyse de sensibilité** dans la vue Fonctions de densité de la boîte de dialogue du noeud Evaluation de simulation. Si les histogrammes sont regroupés par catégorie, une liste déroulante contenant les libellés d'itération vous permet de choisir l'itération à afficher. Vous pouvez également sélectionner l'itération à afficher en cliquant à l'aide du bouton droit de la souris sur le graphique et en sélectionnant l'itération dans le sous-menu **Itération**. Si les histogrammes sont regroupés par itération, une liste déroulante contenant les noms de catégorie vous permet de choisir la catégorie à afficher. Vous pouvez également sélectionner la catégorie à afficher en cliquant à l'aide du bouton droit de la souris sur le graphique et en sélectionnant la catégorie dans le sous-menu **Catégorie**.

Ce graphique est uniquement disponible pour un sous-ensemble de modèles et l'option permettant de générer toutes les probabilités de groupe doit être sélectionnée sur le nugget de modèle. Par exemple, sur le nugget de modèle Logistique, vous devez sélectionner **Ajouter toutes les probabilités**. Les nuggets de modèle suivants prennent en charge cette option :

- Logistique, SVM, Bayes, Réseau de neurones et KNN
- Modèles d'exploration de base de données Db2/ISW pour la régression logistique, arbres de décisions et Naïve Bayes

Par défaut, l'option permettant de générer toutes les probabilités de groupe n'est pas sélectionnée sur ces nuggets de modèle.

Graphiques tornado. Le graphique tornado est un graphique à barres qui représente la sensibilité du champ cible par rapport à chacune des entrées indiquées. La sensibilité est mesurée par la corrélation de la cible avec chaque entrée. Le titre du graphique contient le nom du champ cible. Chaque barre du graphique représente la corrélation entre le champ cible et un champ d'entrée. Les entrées simulées incluses dans le graphique sont celles qui sont sélectionnées dans **Champs d'entrée pour la corrélation tornado** dans la vue Champs de la boîte de dialogue du noeud Evaluation de simulation. L'intitulé de chaque barre correspond à la valeur de corrélation. Les barres sont ordonnées par ordre décroissant, en fonction de la valeur absolue des corrélations. S'il existe des itérations, un graphique distinct est généré pour chacune d'elles. Chaque graphique comporte un sous-titre, qui contient le nom de l'itération.

Table de distribution. Cette table contient la valeur du champ cible, au dessous de laquelle se situe le pourcentage spécifié des observations. Elle comporte une ligne pour chaque valeur de centile indiquée dans la vue Sorties de la boîte de dialogue du noeud Evaluation de simulation. Les valeurs de centile peuvent être des quartiles, un nombre différent de percentiles espacés de manière uniforme, ou des percentiles spécifiés individuellement. La table de distribution contient une colonne par itération.

Informations. Cette section fournit un récapitulatif général des champs et enregistrements utilisés dans l'évaluation. Elle présente également les champs d'entrée et le nombre d'enregistrements, qui sont décomposés pour chaque itération.

Options de graphique

La boîte de dialogue Options de graphique permet de personnaliser l'affichage des graphiques activés pour les fonctions de densité de probabilité et les fonctions de distribution cumulée qui sont générés par une simulation.

Afficher. La liste déroulante **Vue** s'applique uniquement au graphique de la fonction de densité de probabilité. Vous pouvez l'utiliser pour basculer la vue du graphique d'une courbe continue à un histogramme. Cette fonction est désactivée lorsque plusieurs fonctions de densité (provenant de plusieurs itérations) s'affichent sur le même graphique. Lorsqu'il existe plusieurs fonctions de densité, elles ne peuvent être visualisées que sous forme de courbes continues.

Ordre. La liste déroulante **Ordre** s'applique uniquement au graphique de la fonction de distribution cumulée. Elle indique si la fonction de distribution cumulée s'affiche sous forme de fonction croissante (paramètre par défaut) ou décroissante. Lorsqu'elle s'affiche sous forme de fonction décroissante, la valeur de la fonction à un point donné de l'axe horizontal représente la probabilité que le champ cible se situe à la droite de ce point.

Positions du curseur. Le champ de texte **Supérieur** contient la position actuelle de la ligne de référence de droite. Le champ de texte **Inférieur** contient la position actuelle de la ligne de référence de gauche. Vous pouvez définir explicitement les positions des curseurs en saisissant des valeurs dans les champs de texte **Supérieur** et **Inférieur**. La valeur située dans le champ de texte **Inférieur** doit être strictement inférieure à celle figurant dans le champ de texte **Supérieur**. Vous pouvez supprimer la ligne de référence de gauche en sélectionnant **-Infini**, qui permet de définir la position sur l'infini négatif. Cette action désactive le champ de texte **Inférieur**. Vous pouvez supprimer la ligne de référence de droite en sélectionnant **Infini**, qui permet de définir sa position sur l'infini. Cette action désactive le champ de texte **Supérieur**. Vous ne pouvez pas supprimer les deux lignes de référence ; la sélection de l'option **-Infini** désactive la case à cocher **Infini** et inversement.

Lignes de référence. Vous pouvez ajouter diverses lignes de référence verticale fixe aux fonctions de densité de probabilité et aux fonctions de distribution cumulée.

- **Moyenne.** Vous pouvez ajouter une ligne de référence au niveau de la moyenne du champ cible.
- **Médiane.** Vous pouvez ajouter une ligne de référence au niveau de la médiane du champ cible.

- **Ecarts types.** Vous pouvez ajouter des lignes de référence de plus ou moins un nombre donné d'écarts types par rapport à la moyenne du champ cible. Vous pouvez saisir le nombre d'écarts types à utiliser dans le champ de texte adjacent. Le nombre minimal d'écarts types est 1 et le nombre maximal, 10. Le nombre d'écarts types est initialement défini sur 3.
- **Percentiles.** Vous pouvez ajouter des lignes de référence au niveau d'une ou deux valeurs de centile de la distribution pour le champ cible en saisissant des valeurs dans les champs de texte **Bas** et **Haut**. Par exemple, la valeur 95 dans le champ de texte **Haut** représente le 95ème percentile, qui est la valeur au dessous de laquelle se trouvent 95 % des observations. De même, la valeur 5 dans le champ de texte **Bas** représente le cinquième percentile, qui correspond à la valeur au dessous de laquelle se trouvent 5 % des observations. Pour le champ de texte **Bas**, la valeur de centile minimale est 0 et la valeur maximale, 49. Pour le champ de texte **Haut**, la valeur de centile minimale est 50 et la valeur maximale, 100.
- **Positions personnalisées.** Des lignes de référence peuvent être ajoutées à des valeurs spécifiques le long de l'axe horizontal. Vous pouvez supprimer des lignes de référence personnalisées en supprimant l'entrée de la grille.

Lorsque vous cliquez sur **OK**, les curseurs, les libellés situés au dessus des curseurs, les lignes de référence et la table située au dessous du graphique sont mis à jour pour refléter les options sélectionnées dans la boîte de dialogue Options de graphique. Cliquez sur **Annuler** pour fermer la boîte de dialogue sans apporter de modifications. Pour supprimer des lignes de référence, désélectionnez l'option associée dans la boîte de dialogue Options de graphique et cliquez sur **OK**.

Remarque : Lorsque plusieurs fonctions de densité ou fonctions de distribution s'affichent sur un seul graphique (en raison de résultats provenant d'itérations d'analyse de sensibilité), les lignes de référence (autres que les lignes personnalisées) sont appliquées séparément à chaque fonction. Seuls les lignes de référence pour la première itération s'affichent. Les libellés de ligne de référence incluent le libellé d'itération. Ce dernier est dérivé en amont, en général à partir d'un noeud Génération de simulation. Si aucun libellé d'itération n'est disponible, la valeur d'itération est utilisée à la place. Les options **Moyenne**, **Médiane**, **Ecarts types** et **Percentiles** sont désactivées pour les fonctions de distribution cumulée avec plusieurs itérations.

Noeud Sortie d'extension

Si **Sortie à l'écran** est sélectionné dans l'onglet **Sortie** de la boîte de dialogue de noeud Sortie d'extension, la sortie à l'écran s'affiche dans une fenêtre de navigateur de sortie. La sortie est également ajoutée au gestionnaire des sorties. La fenêtre du navigateur de sortie comporte ses propres menus, lesquels vous permettent d'imprimer ou d'enregistrer la sortie, ou de l'exporter dans un autre format. Le menu **Edition** contient uniquement l'option **Copier**. Le navigateur de sortie du noeud Sortie d'extension comporte deux onglets : l'onglet **Sortie de texte** qui affiche une sortie de type texte et l'onglet **Sortie graphique** qui affiche des graphiques et des diagrammes.

Si **Sortie dans le fichier** est sélectionné dans l'onglet **Sortie** de la boîte de dialogue de noeud Sortie d'extension, la fenêtre du navigateur de sortie ne s'affiche pas suite à l'exécution réussie du noeud Sortie d'extension.

Noeud Sortie d'extension - Onglet Syntaxe

Sélectionnez le type de syntaxe – **R** ou **Python for Spark**. Consultez les sections suivantes pour plus d'informations. Lorsque votre syntaxe est prête, vous pouvez cliquer sur **Exécuter** pour exécuter le noeud Sortie d'extension. Les objets de sortie sont ajoutés au gestionnaire de sorties ou, le cas échéant, au fichier indiqué dans le champ **Nom de fichier** de l'onglet **Sortie**.

Syntaxe R

Syntaxe R. Permet d'entrer ou de coller la syntaxe de script R personnalisé en vue de l'analyse des données dans ce champ.

Convertir les champs indicateurs. Indique comment sont traités les champs indicateurs. Deux options sont disponibles : **Chaînes en facteur, Entiers et Réels en double** et **Valeurs logiques (True, False)**. Si vous sélectionnez **Valeurs logiques (True, False)**, les valeurs originales des champs indicateurs sont perdues. Par exemple, si un champ a les valeurs Mâle et Femelle, elles sont remplacées par True et False.

Convertir les valeurs manquantes en valeur R 'non disponible' (NA). Lorsque cette option est sélectionnée, toute valeur manquante est convertie en valeur R `NA`. La valeur `NA` est utilisée par R pour identifier les valeurs manquantes. Des fonctions R que vous utilisez peuvent comporter un argument par le biais duquel il est possible de contrôler le comportement des fonctions lorsque les données contiennent `NA`. Par exemple, la fonction peut vous permettre de choisir d'exclure automatiquement les enregistrements qui contiennent `NA`. Si cette option n'est pas sélectionnée, les valeurs manquantes sont transmises à R en l'état et peuvent entraîner des erreurs lors de l'exécution du script R.

Convertir les champs date/heure en classes R avec contrôle spécial pour les fuseaux horaires. Lorsque cette option est sélectionnée, les variables de format de date et de date/heure sont converties en objets R date/heure. Vous devez sélectionner l'une des options suivantes :

- **R POSIXct.** Les variables de format de date ou de date/heure sont converties en objets R `POSIXct`.
- **R POSIXlt (liste).** Les variables de format de date ou de date/heure sont converties en objets R `POSIXlt`.

Remarque : Les formats POSIX sont des options avancées. Utilisez-les uniquement si le script R spécifie que les champs date/heure sont traités de telle manière que ces formats sont requis. Les formats POSIX ne s'appliquent pas aux variables de format horaire.

Syntaxe Python

Syntaxe Python. Permet d'entrer ou de coller la syntaxe de script Python personnalisé en vue de l'analyse des données dans ce champ. Pour plus d'informations relatives à Python for Spark, voir Python for Spark et Scriptage avec Python for Spark.

Noeud Sortie d'extension - Onglet Sortie de la console

L'onglet **Sortie de la console** contient les sorties reçues lorsque le script R ou le script Python for Spark de l'onglet Syntaxe est exécuté (par exemple, si un script R est utilisé, il affiche la sortie reçue de la console R lorsque le script R du champ **Syntaxe R** de l'onglet **Syntaxe** est exécuté). La sortie peut contenir des messages d'erreur ou d'avertissement R ou Python générés lors de l'exécution du script R ou Python. Cette sortie permet essentiellement de déboguer le script. L'onglet **Sortie de la console** contient également le script du champ **Syntaxe R** ou **Syntaxe Python**.

A chaque exécution du script Sortie d'extension, le contenu de l'onglet **Sortie de la console** est écrasé par la sortie reçue de la console R ou Python for Spark. La sortie ne peut pas être éditée.

Noeud Sortie d'extension - Onglet Sortie

Nom de la sortie - Spécifie le nom de la sortie générée lorsque le noeud est exécuté. Lorsque **Auto** est sélectionné, le nom "Sortie R" ou "Sortie Python" est attribué automatiquement à la sortie selon le type de script. Si vous le souhaitez, vous pouvez choisir **Personnalisé** pour indiquer un autre nom.

Sortie à l'écran. Sélectionnez cette option pour générer et afficher la sortie dans une nouvelle fenêtre. La sortie est également ajoutée au gestionnaire des sorties.

Sortie dans un fichier. Sélectionnez cette option pour enregistrer la sortie dans un fichier. Les boutons radio **Graphique de sortie** et **Fichier de sortie** sont alors activés.

Graphique de sortie. Activé uniquement si **Sortie dans le fichier** est sélectionné. Sélectionnez cette option pour enregistrer des graphiques générés à la suite de l'exécution du noeud Sortie d'extension dans

un fichier. Indiquez un nom de fichier à utiliser pour la sortie générée dans le champ **Nom de fichier**. Cliquez sur les points de suspension (**...**) pour choisir un fichier et un emplacement spécifiques. Indiquez le type du fichier dans la liste déroulante **Type de fichier**. Les types de fichier suivants sont disponibles :

- Objet de sortie (`.cou`)
- HTML (`.html`)

Texte de sortie. Activé uniquement si **Sortie dans le fichier** est sélectionné. Sélectionnez cette option pour enregistrer une sortie de texte générée suite à l'exécution du noeud Sortie d'extension dans un fichier. Indiquez un nom de fichier à utiliser pour la sortie générée dans le champ **Nom de fichier**. Cliquez sur les points de suspension (**...**) pour spécifier un fichier et un emplacement spécifiques. Indiquez le type du fichier dans la liste déroulante **Type de fichier**. Les types de fichier suivants sont disponibles :

- HTML (`.html`)
- Objet de sortie (`.cou`)
- Document texte (`.txt`)

Navigateur Sortie d'extension

Si **Sortie à l'écran** est sélectionné dans l'onglet **Sortie** de la boîte de dialogue de noeud Sortie d'extension, la sortie à l'écran s'affiche dans une fenêtre de navigateur de sortie. La sortie est également ajoutée au gestionnaire des sorties. La fenêtre du navigateur de sortie comporte ses propres menus, lesquels vous permettent d'imprimer ou d'enregistrer la sortie, ou de l'exporter dans un autre format. Le menu **Edition** contient uniquement l'option **Copier**. Le navigateur de sortie du noeud Sortie d'extension contient deux onglets :

- L'onglet **Sortie de texte** affiche la sortie de texte
- L'onglet **Sortie de graphique** affiche les graphiques

Si **Sortie dans le fichier** est sélectionné dans l'onglet **Sortie** de la boîte de dialogue de noeud Sortie d'extension au lieu de **Sortie à l'écran**, la fenêtre du navigateur de sortie ne s'affiche pas suite à l'exécution réussie du noeud Sortie d'eExtension.

Navigateur Sortie d'extension - Onglet Sortie de texte

L'onglet **Sortie de texte** affiche toutes les sorties de texte générées lors de l'exécution du script R ou Python for Spark sur l'onglet **Syntaxe** du noeud Sortie d'extension.

Remarque : Les messages d'erreur ou d'avertissement R ou Python for Spark renvoyés suite à l'exécution du script Sortie d'extension s'affichent toujours dans l'onglet **Sortie de la console** du noeud Sortie d'extension.

Navigateur Sortie d'extension - Onglet Sortie de graphique

L'onglet **Sortie de graphique** affiche tous les graphiques générés lors de l'exécution du script R ou Python for Spark sur l'onglet **Syntaxe** du noeud Sortie d'extension. Par exemple, si le script contient un appel à la fonction R `plot`, le graphique résultant s'affiche dans cet onglet.

Programmes externes de IBM SPSS Statistics

Si une version compatible de IBM SPSS Statistics est installée sur votre ordinateur et que vous disposez d'une version sous licence, vous pouvez configurer IBM SPSS Modeler pour traiter les données avec la fonctionnalité IBM SPSS Statistics à l'aide des noeuds Transformation Statistiques, Modèle Statistiques, Sortie Statistiques, ou Export Statistiques.

Pour des informations sur la compatibilité du produit avec la version actuelle de IBM SPSS Modeler, consultez le site de support technique de la société à l'adresse *http://www.ibm.com/support*.

Pour configurer IBM SPSS Modeler afin de l'utiliser avec IBM SPSS Statistics et d'autres applications, choisissez :

Outils > **Options** > **Applications externes**

IBM SPSS Statistics Interactive. Saisissez le chemin d'accès et le nom complet de la commande (par exemple, *C:\Program Files\IBM\SPSS\Statistics\<nn>\stats.exe*) à utiliser lorsque vous lancez IBM SPSS Statistics directement sur un fichier de données produit par le noeud Exporter Statistics. Pour plus d'informations, voir «Noeud Exporter Statistics», à la page 382.

Connexion. Si le serveur IBM SPSS Statistics est situé sur le même ordinateur qu'IBM SPSS Modeler Server, vous pouvez, à des fins d'efficacité, activer une connexion entre les deux applications. Lors de l'analyse, les données restent ainsi sur le serveur. Sélectionnez **Serveur** pour activer l'option **Port** ci-après. L'option par défaut est **Local**.

Port. Spécifiez le port du serveur IBM SPSS Statistics.

IBM SPSS Statistics Location Utility. Afin d'activer IBM SPSS Modeler pour pouvoir utiliser les noeuds Transformation Statistiques, Modèle Statistiques, et Sorties Statistiques, vous devez disposer d'une copie de IBM SPSS Statistics installée et de la licence correspondante sur l'ordinateur sur lequel le flux est exécuté.

- Si vous exécutez IBM SPSS Modeler en mode local (autonome), la copie sous licence de IBM SPSS Statistics doit se trouver sur l'ordinateur local. Cliquez sur ce bouton pour indiquer l'emplacement de l'installation IBM SPSS Statistics locale que vous souhaitez utiliser pour la licence.
- En outre, si vous exécutez le logiciel en mode distribué en utilisant IBM SPSS Modeler Server distant, vous devez également exécuter un utilitaire sur l'hôte IBM SPSS Modeler Server afin de créer le fichier *statistics.ini*, qui indique à IBM SPSS Statistics le chemin d'installation d'IBM SPSS Modeler Server. Pour ce faire, à partir de l'invite de commande, accédez au répertoire IBM SPSS Modeler Server *bin* et sous Windows, exécutez :

  ```
  statisticsutility -location=<IBM SPSS Statistics_installation_path>/
  ```

  Sous UNIX, exécutez :

  ```
  ./statisticsutility -location=<IBM SPSS Statistics_installation_path>/bin
  ```

Si vous ne disposez pas d'une copie sous licence de IBM SPSS Statistics sur votre ordinateur local, vous pouvez quand même exécuter le noeud Fichier Statistiques sur un serveur IBM SPSS Statistics, mais les tentatives d'exécution d'autres noeuds IBM SPSS Statistics entraîneront l'affichage d'un message d'erreur.

Commentaires

Si vous ne parvenez pas à exécuter correctement un noeud Commande IBM SPSS Statistics, suivez les conseils ci-dessous :

- Si les noms de champ utilisés dans IBM SPSS Modeler dépassent huit caractères (pour les versions antérieures à IBM SPSS Statistics 12.0), 64 caractères (pour IBM SPSS Statistics 12.0 et les versions suivantes) ou contiennent des caractères incorrects, il est nécessaire de les renommer ou de les tronquer avant de les lire dans IBM SPSS Statistics. Pour plus d'informations, voir «Changement du nom ou filtrage des champs pour IBM SPSS Statistics», à la page 383.
- Si IBM SPSS Statistics a été installé après IBM SPSS Modeler, il vous faudra indiquer l'emplacement de IBM SPSS Statistics. Voir ci-dessus.

Chapitre 7. Noeuds d'exportation

Présentation des noeuds d'exportation

Les noeuds d'exportation permettent d'exporter les données dans divers formats, afin de pouvoir les utiliser avec d'autres logiciels.

Les noeuds d'exportation disponibles sont les suivants :

Le noeud Export SGBD écrit des données dans une source de données relationnelles compatible ODBC. Pour que cette opération puisse être effectuée, la source de données ODBC doit exister et vous devez y avoir accès en écriture.

Le noeud Export Fichier plat génère des données dans un fichier texte délimité. Elles peuvent ainsi être lues par d'autres logiciels d'analyse ou par des tableurs.

Le noeud Export Statistics génère des données au format IBM SPSS Statistics *.sav* ou *.zsav*. Les fichiers *.sav* ou *.zsav* peuvent être lus par IBM SPSS Statistics Base et d'autres produits. Ce format est également utilisé pour les fichiers cache IBM SPSS Modeler.

Le noeud Export Data Collection génère des données au format utilisé par les logiciels d'étude de marché Data Collection. Pour pouvoir utiliser ce noeud, vous devez avoir installé auparavant Data Collection Data Library.

Le noeud IBM Cognos Export exporte des données dans un format qui peut être lu par les bases de données Cognos.

Le noeud IBM Cognos TM1 Export exporte des données dans un format qui peut être lu par les bases de données Cognos TM1.

Le noeud Export SAS permet d'obtenir des données de sortie au format SAS afin qu'elles puissent être lues par SAS ou par un logiciel compatible. Trois formats de fichier SAS sont disponibles : SAS pour Windows/OS2, SAS pour UNIX ou SAS version 7/8.

Le noeud Export Excel génère une sortie de données au format de fichier .xlsx de Microsoft Excel. Si vous le souhaitez, vous pouvez choisir de lancer Excel automatiquement et d'ouvrir le fichier exporté lors de l'exécution du noeud.

Le noeud Export XML génère une sortie de données dans un fichier au format XML. Vous pouvez également créer un noeud source XML pour lire de nouveau les données exportées dans le flux.

Noeud Export SGBD

Vous pouvez utiliser les noeuds SGBD pour écrire des données à des bases de données relationnelles compatibles ODBC, qui sont explicitées dans le noeud SGBD source. Pour plus d'informations, voir «noeud source de base de données», à la page 18.

Pour écrire des données dans une base de données, utilisez la procédure générale suivante :

1. Installez un pilote ODBC et configurez une source de données pour la base de données à utiliser.
2. Dans l'onglet Exporter du noeud SGBD, indiquez la source de données et la table où écrire. Vous pouvez créer une table ou insérer des données dans une table existante.
3. Indiquez d'autres options si nécessaire.

Cette procédure est détaillée dans les rubriques suivantes.

Noeud SGBD - Onglet Exporter

Remarque : Certaines des bases de données vers lesquelles vous exportez peuvent ne pas prendre en charge les noms de colonne dans les tables de plus de de 30 caractères de longueur. Si vous voyez s'afficher un message d'erreur indiquant que votre table a un nom de colonne incorrect, réduisez la taille du nom à moins de 30 caractères..

Source de données. Source de données sélectionnée. Entrez directement son nom ou sélectionnez-le dans la liste déroulante. Si la base de données souhaitée n'apparaît pas dans la liste, sélectionnez **Ajouter une nouvelle connexion à la base de données** et localisez votre base de données dans la boîte de dialogue Connexions de base de données. Pour plus d'informations, voir «Ajout d'une connexion de base de données», à la page 20.

Nom de la table - Entrez le nom de la table vers laquelle envoyer les données. Si vous sélectionnez l'option **Insérer dans la table**, vous pouvez choisir une table existante dans la base de données en cliquant sur le bouton **Sélectionner**.

Créer une table. Sélectionnez cette option pour créer une nouvelle table de base de données ou écraser une table de base de données existante.

Insérer dans la table. Sélectionnez cette option pour insérer les données dans de nouvelles lignes d'une table de base de données existante.

Fusionner la table. (Le cas échéant) Sélectionnez cette option pour mettre à jour les colonnes de la base de données sélectionnées avec des valeurs de champs de données source correspondants. Sélectionner cette option active le bouton **Fusionner**, qui affiche une boîte de dialogue dans laquelle vous pouvez mapper les champs de données source sur les colonnes de la base de données.

Supprimer la table existante. Sélectionnez cette option pour supprimer, le cas échéant, une table existante du même nom que la table créée.

Supprimer des lignes existantes. Sélectionnez cette option pour supprimer les lignes existantes de la table avant l'exportation, lors de l'insertion dans une table.

Remarque : Si vous sélectionnez l'une des deux options ci-dessus, vous recevez le message **Avertissement d'écrasement** lors de l'exécution du noeud. Pour que ces avertissements n'apparaissent plus, désélectionnez l'option **Avertir lorsqu'un noeud écrase une table de base de données** dans l'onglet Notifications de la boîte de dialogue Options utilisateur.

Taille de chaîne par défaut. Les champs marqués comme étant "sans type" dans un noeud Typer en amont sont écrits dans la base de données sous forme de champs de type chaîne. Indiquez la taille des chaînes à utiliser pour les champs sans type.

Cliquez sur **Schéma** pour ouvrir une boîte de dialogue dans laquelle vous pouvez définir diverses options d'exportation (pour les bases de données prenant en charge cette fonctionnalité), définir des types de données SQL pour vos champs et spécifier la clé primaire en vue de l'indexation de base de données. Pour plus d'informations, voir «Options de schéma d'exportation de base de données», à la page 368.

Cliquez sur **Index** pour définir les options d'indexation de la table exportée, afin d'améliorer les performances de la base de données. Pour plus d'informations, voir «Export SGBD - Options de l'index», à la page 371.

Cliquez sur **Options avancées** pour spécifier les options de chargement en bloc et de validation de base de données. Pour plus d'informations, voir «Options avancées d'exportation de base de données», à la page 372.

Entourer de guillemets les noms des tables et colonnes. Sélectionnez les options à utiliser lors de l'envoi d'une instruction `CREATE TABLE` à la base de données. Les tableaux ou colonnes comportant des espaces ou des caractères spéciaux doivent être mis entre guillemets.

- **Selon les besoins.** Sélectionnez cette option pour qu'IBM SPSS Modeler détermine automatiquement, au cas par cas, la nécessité d'utiliser des guillemets.
- **Toujours.** Sélectionnez cette option pour que les noms de tableau et de colonne soient systématiquement mis entre guillemets.
- **Jamais.** Sélectionnez cette option pour désactiver l'utilisation des guillemets.

Générer un noeud source pour ces données. Sélectionnez cette option pour générer un noeud source SGBD pour les données lors de leur exportation dans la table et la source de données spécifiées. Dès l'exécution, ce noeud est ajouté à l'espace de travail de flux.

Export SGBD - Options de fusion

Cette boîte de dialogue vous permet de mapper des champs à partir des données source dans des colonnes du tableau de la base de données cible. Lorsqu'un champ de données source est mappé sur une colonne de la base de données, la valeur de cette colonne est remplacée par la valeur des données source lorsque le flux est exécuté. Les champs source non mappés ne sont pas modifiés dans la base de données.

Mapper les champs. C'est ici que vous indiquez le mappage entre les champs de données source et les colonnes de la base de données. Les champs de données source avec le même nom que les colonnes de la base de données sont automatiquement mappés.

- **Mapper.** Mappe un champ de données source sélectionné dans la liste des champs à la gauche du bouton sur une colonne de la base de données sélectionnée dans la liste de droite. Vous pouvez mapper plusieurs champs en même temps mais le nombre d'entrées sélectionnées dans les deux listes doit être le même.

- **Démapper.** Supprime le mappage pour une ou plusieurs colonnes sélectionnées de la base de données. Ce bouton est activé lorsque vous sélectionnez un champ ou une colonne de base de données dans la table située à droite de la boîte de dialogue.
- **Ajouter.** Ajoute un ou plusieurs champs de données source sélectionnés dans la liste des champs à gauche du bouton, à la liste de droite prête pour le mappage. Ce bouton est activé lorsque vous sélectionnez un champ dans la liste de gauche et qu'aucun champ portant ce nom n'existe dans la liste de droite. En cliquant sur ce bouton, vous mappez le champ sélectionné sur une nouvelle colonne de la base de données portant le même nom. Le mot **<NEW>** est affiché après le nom de la colonne de base de données pour indiquer que le champ est nouveau.

Fusionner les lignes. Vous utilisez un champ-clé, tel que *ID transaction*, pour fusionner les enregistrements ayant une valeur identique dans ce champ. Cette opération est équivalente à une "équi-jointure" de base de données. Les valeurs des clés doivent être celles des clés principales ; c'est-à-dire qu'elles doivent être uniques et ne peuvent pas contenir de valeurs nulles.

- **Clés possibles.** Affiche tous les champs trouvés dans toutes les sources de données d'entrée. Sélectionnez un ou plusieurs champs de cette liste et utilisez la flèche pour les ajouter comme champs-clés pour la fusion des enregistrements. Tout champ de mappage avec une colonne de base de données mappée correspondante est disponible comme champ-clé, sauf que les champs ajoutés en tant que nouvelles colonnes de la base de données (indiqués par un **<NEW>** après le nom) ne sont pas disponibles.
- **Clés pour fusion.** Affiche tous les champs utilisés pour fusionner les enregistrements de toutes les sources de données d'entrée, sur la base des valeurs des champs-clés. Pour supprimer une clé de la liste, sélectionnez-la et utilisez la flèche pour la renvoyer dans la liste Clés possibles. Lorsque plusieurs champs-clés sont sélectionnés, l'option ci-dessous est activée.
- **Inclure uniquement les enregistrements qui existent dans la base de données.** Effectue une jointure partielle ; si l'enregistrement se trouve dans la base de données et dans le flux, les champs mappés seront mis à jour.
- **Ajouter les enregistrements à la base de données.** Effectue une jointure externe ; tous les enregistrements dans le flux seront fusionnés (si le même enregistrement existe dans la base de données) ou ajoutés (si l'enregistrement n'existe pas encore dans la base de données).

Pour mapper un champ de données source sur une nouvelle colonne de la base de données

1. Cliquez sur le nom du champ source dans la liste de gauche, sous **Mapper les champs**.
2. Cliquez sur le bouton **Ajouter** pour terminer le mappage.

Pour mapper un champ de données source sur une colonne existante de la base de données

1. Cliquez sur le nom du champ source dans la liste de gauche, sous **Mapper les champs**.
2. Cliquez sur le nom de la colonne sous **Colonne de la base de données** à droite.
3. Cliquez sur le bouton **Mapper** pour terminer le mappage.

Pour supprimer un mappage

1. Dans la liste de droite, dans Champ, cliquez sur le nom du champ dont vous souhaitez supprimer le mappage.
2. Cliquez sur le bouton **Démapper**.

Pour annuler la sélection d'un champ de l'une des listes

Maintenez enfoncée la touche CTRL et cliquez sur le nom du champ.

Options de schéma d'exportation de base de données

Dans la boîte de dialogue du schéma d'exportation de base de données, vous pouvez définir des options d'exportation de base de données (pour les bases de données prenant en charge ces options), définir des

types de données SQL pour vos champs, indiquer les champs qui constituent des clés primaires et personnaliser l'instruction CREATE TABLE générée lors de l'exportation.

Cette boîte de dialogue comprend plusieurs parties :

- La partie supérieure (si elle est visible) contient des options d'exportation vers une base de données prenant en charge ces options. Cette section n'apparaît pas si vous n'êtes pas connecté à une base de données de cette catégorie.
- Le champ de texte, dans la partie centrale, affiche le modèle utilisé pour générer la commande CREATE TABLE, qui suit par défaut le format ci-après :

  ```
  CREATE TABLE <table-name> <(table columns)>
  ```

- Le tableau, dans la partie inférieure, vous permet de définir le type de données SQL de chaque champ et d'indiquer les champs qui constituent des clés primaires, comme indiqué ci-dessous. La boîte de dialogue génère automatiquement les valeurs des paramètres <table-name> et <(table columns)> à partir des spécifications figurant dans le tableau.

Définition des options d'exportation de base de données

Si cette section apparaît, vous pouvez spécifier un certain nombre de paramètres pour l'exportation vers la base de données. Les bases de donnés prenant en charge cette fonctionnalité sont les suivantes.

- Editions SQL Server Enterprise et Developer. Pour plus d'informations, voir «Options pour SQL Server», à la page 370.
- Editions Oracle Enterprise ou Personal. Pour plus d'informations, voir «Options pour Oracle», à la page 370.

Personnalisation d'instructions CREATE TABLE

A l'aide du champ de texte de cette boîte de dialogue, vous pouvez ajouter d'autres options spécifiques aux bases de données à l'instruction CREATE TABLE.

1. Cochez la case **Personnaliser la commande CREATE TABLE** pour activer la fenêtre de texte.
2. Ajoutez des options de base de données à l'instruction. Veillez à conserver les paramètres de texte <table-name> et (<table-columns>), puisqu'IBM SPSS Modeler les remplace ensuite par les définitions de nom et de colonne réelles de la table .

Définition de types de données SQL

Par défaut, IBM SPSS Modeler permet au serveur de base de données d'affecter automatiquement des types de données SQL. Pour remplacer le type affecté automatiquement à un champ, recherchez la ligne correspondant au champ et sélectionnez le type voulu dans la liste déroulante de la colonne *Type* du tableau de la boîte de dialogue Schéma. Vous pouvez utiliser Maj-clic pour sélectionner plusieurs lignes.

Dans le cas des types dotés d'un argument de longueur, de précision ou d'échelle (BINARY, VARBINARY, CHAR, VARCHAR, NUMERIC et NUMBER), il vaut mieux spécifier une longueur plutôt que de laisser le serveur de base de données définir une longueur automatiquement. Par exemple, si vous spécifiez une valeur probable, comme VARCHAR(25), pour la longueur, le type de stockage dans IBM SPSS Modeler sera écrasé si telle est votre intention. Pour remplacer l'affectation automatique, sélectionnez **Spécifier** dans la liste déroulante Type et remplacez la définition du type par l'instruction de définition de type SQL souhaitée.

La méthode la plus simple consiste à sélectionner d'abord le type le plus proche de la définition souhaitée, puis à choisir **Spécifier** afin de modifier cette définition. Par exemple, pour définir le type de données SQL sur VARCHAR(25), paramétrez d'abord le type sur **VARCHAR(length)** dans la liste déroulante Type, puis sélectionnez **Spécifier** et remplacez la longueur de texte par la valeur 25.

Clés primaires

Si une ou plusieurs colonnes de la table exportée doivent comporter une valeur ou une combinaison de valeurs unique pour chaque ligne, vous pouvez l'indiquer en cochant la case **Clé primaire** correspondant à chaque champ concerné. La plupart des bases de données n'autorisent aucune modification de la table qui invaliderait une contrainte de clé primaire et créent automatiquement un index en fonction de la clé primaire pour appliquer cette restriction. (Si vous le souhaitez, vous pouvez créer des index pour d'autres champs dans la boîte de dialogue Index. Pour plus d'informations, voir «Export SGBD - Options de l'index», à la page 371.)

Options pour SQL Server

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression de. Choisissez le niveau de compression.

- **Ligne.** Active la compression au niveau des lignes (par exemple, l'équivalent de `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW);` dans SQL).
- **Page.** Active la compression au niveau des pages (par exemple `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);` dans SQL).

Options pour Oracle

Paramètres d'Oracle - Option basique

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression de. Choisissez le niveau de compression.

- **Par défaut.** Active la compression par défaut (par exemple `CREATE TABLE MYTABLE(...) COMPRESS;` dans SQL). Dans ce cas, cela a le même effet que l'option **Basique**.
- **De base.** Active la compression de base (par exemple `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` dans SQL).

Paramètres d'Oracle - Option avancée

Utiliser la compression. Si cette option est sélectionnée, des tables à exporter avec la compression sont créées.

Compression de. Choisissez le niveau de compression.

- **Par défaut.** Active la compression par défaut (par exemple `CREATE TABLE MYTABLE(...) COMPRESS;` dans SQL). Dans ce cas, cela a le même effet que l'option **Basique**.
- **De base.** Active la compression de base (par exemple `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` dans SQL).
- **OLTP.** Active la compression OLTP (par exemple `CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP;` dans SQL).
- **Requête faible/élevée.** (Serveurs Exadata uniquement) Active la compression Exadata Hybrid Columnar Compression pour les requêtes (par exemple `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW;` ou `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH;` dans SQL). La compression des requêtes est utile dans les environnements d'entreposage de données ; `HIGH` fournit un rapport de compression plus grand que `LOW`.
- **Archive faible/élevée.** (Serveurs Exadata uniquement) Active la compression Exadata Hybrid Columnar Compression pour les archives (par exemple `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW;` ou `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH;` dans SQL). La compression des archives est utile pour compresser des données qui seront stockées pendant de longues périodes ; `HIGH` fournit un rapport de compression plus grand que `LOW`.

Export SGBD - Options de l'index

La boîte de dialogue Index vous permet de créer des index sur les tables de base de données exportées depuis IBM SPSS Modeler. Vous pouvez indiquer les ensembles de champs à inclure et personnaliser la commande CREATE INDEX, selon les besoins.

Cette boîte de dialogue comprend deux parties :

- Le champ de texte figurant dans la partie supérieure affiche un modèle qui permet de générer une ou plusieurs commandes CREATE INDEX, qui suivent par défaut au format ci-après :

  ```
  CREATE INDEX <index-name> ON <table-name>
  ```

- Le tableau figurant dans la partie inférieure de la boîte de dialogue vous permet d'ajouter des spécifications pour chaque index à créer. Indiquez, pour chaque index, son nom ainsi que les champs ou les colonnes à inclure. La boîte de dialogue génère automatiquement les valeurs des paramètres <index-name> et <table-name> en conséquence.

  Par exemple, le code SQL généré pour un index simple réalisé sur les champs *empid* et *deptid* peut utiliser la syntaxe suivante :

  ```
  CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)
  ```

  Vous pouvez ajouter plusieurs lignes pour créer plusieurs index. Une autre commande CREATE INDEX est générée pour chaque ligne.

Personnalisation de la commande CREATE INDEX

Si vous le souhaitez, vous pouvez personnaliser la commande CREATE INDEX pour tous les index ou pour un index précis uniquement. Vous disposez ainsi d'une marge de manoeuvre pour vous adapter à des options ou des exigences de base de données particulières et pour appliquer des personnalisations à tous les index ou à certains index uniquement, si nécessaire.

- Sélectionnez **Personnaliser la commande CREATE INDEX** en haut de la boîte de dialogue afin de modifier le modèle utilisé pour tous les index ajoutés dès à présent. Notez que ces changements ne s'appliquent pas automatiquement aux index déjà ajoutés à la table.
- Sélectionnez une ou plusieurs lignes de la table, puis cliquez sur **Mettre à jour les index sélectionnés** en haut de la boîte de dialogue pour appliquer les personnalisations actuelles à toutes les lignes sélectionnées.
- Cochez la case **Personnaliser** sur chaque ligne pour modifier le modèle de commande de l'index uniquement.

La boîte de dialogue génère automatiquement les valeurs des paramètres <index-name> et <table-name> à partir des spécifications de la table ; ces valeurs ne peuvent pas être éditées directement.

BITMAP KEYWORD. Si vous utilisez une base de données Oracle, vous pouvez personnaliser le modèle afin de créer un index bitmap et non un index standard :

```
CREATE BITMAP INDEX <index-name> ON <table-name>
```

Les index bitmap peuvent s'avérer utiles pour indexer les colonnes contenant un nombre limité de valeurs distinctes. Le code SQL résultant peut être semblable à ce qui suit :

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

Mot-clé UNIQUE. La plupart des bases de données prennent en charge le mot-clé UNIQUE dans la commande CREATE INDEX. Il est ainsi possible d'appliquer à la table sous-jacente une contrainte d'unicité semblable à une contrainte de clé primaire.

```
CREATE UNIQUE INDEX <index-name> ON <table-name>
```

Pour les champs désignés comme clés primaires, cette spécification n'est pas nécessaire. La plupart des bases de données créent automatiquement un index pour les champs définis comme champs de clé primaire dans la commande CREATE TABLE. Par conséquent, la création explicite d'index sur ces champs est superflue. Pour plus d'informations, voir «Options de schéma d'exportation de base de données», à la page 368.

Mot-clé FILLFACTOR. Certains paramètres physiques de l'index peuvent être affinés. Par exemple, SQL Server permet à l'utilisateur de compenser les coûts de maintenance par la taille de l'index (après sa création initiale), lors des modifications ultérieures apportées à la table.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

Autres commentaires

- Si un index du même nom existe déjà, la création d'index échoue. Les échecs sont considérés initialement comme des avertissements (ce qui permet la création des index suivants), puis ils sont ensuite signalés comme des erreurs dans le journal des messages une fois que le système a essayé de créer tous les index.
- Pour optimiser les performances, les index doivent être créés une fois les données chargées dans la table. Les index doivent contenir au moins une colonne.
- Avant d'exécuter le noeud, vous pouvez prévisualiser le code SQL généré dans le journal des messages.
- Pour les tables temporaires écrites dans la base de données (c'est-à-dire lorsque la mise en cache des noeuds est activée), les options permettant de définir des clés primaires et des index ne sont pas disponibles. Toutefois, si cela s'avère nécessaire, le système peut créer des index à partir de la table temporaire en fonction du mode d'utilisation des données dans les noeuds en aval. Par exemple, si les données mises en cache sont ensuite liées par la colonne *DEPT*, il semble alors judicieux d'indexer sur cette colonne la table mise en cache.

Index et optimisation des requêtes

Dans certains systèmes de gestion de base de données, une fois la table de base de données créée, chargée et indexée, une autre étape est nécessaire pour que l'optimiseur puisse utiliser les index et accélérer l'exécution des requêtes sur la nouvelle table. Par exemple, dans Oracle, l'optimiseur de requêtes basé sur le coût exige qu'une table soit d'abord analysée avant que ses index puissent être utilisés pour l'optimisation des requêtes. Le fichier de propriétés ODBC interne pour Oracle (non visible par l'utilisateur) comporte une option pour que cela se produise, comme suit :

```
# Définit le code SQL à exécuter une fois qu'une table et les index associés
# ont été créés et renseignés
table_analysis_sql, 'ANALYZE TABLE <table-name> COMPUTE STATISTICS'
```

Cette étape est exécutée chaque fois qu'une table est créée dans Oracle (que des clés primaires ou des index soient définis ou non). Si nécessaire, le fichier de propriétés ODBC des bases de données supplémentaires peut être personnalisé de la même manière : contactez l'assistance technique.

Options avancées d'exportation de base de données

Lorsque vous cliquez sur Options avancées dans la boîte de dialogue du noeud d'exportation de base de données, une nouvelle boîte de dialogue apparaît, qui vous permet de spécifier les détails techniques d'exportation des résultats dans une base de données.

Utiliser la validation par lots. Sélectionnez cette option afin de désactiver la validation ligne à ligne dans la base de données.

Taille de lot. Indique le nombre d'enregistrements à envoyer à la base de données avant validation dans la mémoire. Si vous choisissez une valeur faible, l'intégrité des données est mieux préservée mais la vitesse de transfert moins rapide. Vous pouvez modifier cette valeur afin d'utiliser au mieux votre base de données.

Utiliser le chargement en bloc. Spécifie la méthode de chargement en bloc des données vers la base de données directement à partir d'IBM SPSS Modeler. Vous devrez peut-être effectuer des tests pour sélectionner les options de chargement en bloc adaptées à un scénario particulier.

- **Via ODBC.** Sélectionnez cette option afin d'utiliser l'API ODBC pour exécuter des insertions de plusieurs lignes. Cette méthode est plus efficace qu'une simple exportation vers la base de données. Parmi les options ci-après, optez pour un lien par ligne ou par colonne.
- **Via le module de chargement externe.** Sélectionnez cette option afin d'utiliser un programme de module de chargement en bloc personnalisé propre à votre base de données. Les options ci-dessous sont alors automatiquement activées. Pour AIX, si vous rencontrez des problèmes de données UTF-8, vous devrez peut-être ajouter `locale, en_US.UDF-8` dans `options.cfg`.

Options ODBC avancées. Ces options ne sont disponibles que lorsque vous sélectionnez **Via ODBC**. Notez que tous les pilotes ODBC ne prennent pas en charge ces fonctions.

- **Par ligne.** Sélectionnez le lien par ligne afin d'utiliser `SQLBulkOperations` pour charger les données dans la base de données. Le lien par ligne permet d'obtenir une vitesse plus importante que les insertions configurées qui insèrent les données de chaque enregistrement séparément.
- **Par colonne.** Sélectionnez cette option afin d'utiliser le lien par colonne pour charger les données dans la base de données. Le lien par colonne permet d'obtenir de meilleures performances car il relie chaque colonne de la base de données (dans une instruction `INSERT` configurée) à un ensemble de valeurs *N*. Si vous exécutez l'instruction `INSERT` une fois, *N* lignes sont insérées dans la base de données. Cette méthode permet d'obtenir de bien meilleures performances.

Options du module de chargement externe. Lorsque vous choisissez **Via le module de chargement externe**, de nombreuses options apparaissent : elles permettent d'exporter le jeu de données dans un fichier, et de spécifier et d'exécuter un programme de module de chargement personnalisé pour charger les données de ce fichier vers la base de données. IBM SPSS Modeler fonctionne avec les chargeurs externes d'un grand nombre de systèmes de base de données connus. Plusieurs scripts ont été inclus dans le logiciel ; ils se trouvent, avec la documentation technique, dans le sous-répertoire `scripts`. Notez que pour utiliser cette fonctionnalité, Python 2.7 doit être installé sur le même ordinateur qu'IBM SPSS Modeler ou IBM SPSS Modeler Server, et le paramètre `python_exe_path` doit être défini dans le fichier `options.cfg`. Pour plus d'informations, voir «Programmation de chargement en bloc», à la page 374.

- **Utiliser le délimiteur.** Spécifie le délimiteur à utiliser dans le fichier exporté. Sélectionnez **Tabulation** afin d'utiliser la tabulation comme délimiteur et **Espace** pour choisir l'espace. Sélectionnez **Autre** pour choisir un autre caractère, comme une virgule (,).
- **Spécifier le fichier de données.** Sélectionnez cette option afin de saisir l'emplacement de destination du fichier de données lors du chargement en bloc. Par défaut, un fichier temporaire est créé dans le répertoire temporaire du serveur.
- **Spécifier le programme du module de chargement.** Sélectionnez cette option pour spécifier le programme de chargement en bloc à utiliser. Par défaut, le logiciel recherche dans le sous-répertoire `scripts` du dossier d'installation IBM SPSS Modeler, le script Python à exécuter pour une base de données spécifique. Plusieurs scripts ont été inclus dans le logiciel ; ils se trouvent, avec la documentation technique, dans le sous-répertoire `scripts`.
- **Générer le fichier journal.** Sélectionnez cette option afin de générer un fichier journal dans le répertoire spécifié. Ce fichier journal contient les informations relatives aux erreurs. Il est particulièrement utile en cas d'échec du chargement en bloc.
- **Vérifier la taille de la table.** Sélectionnez cette option afin de vérifier les tables dans le but de garantir que l'augmentation de la taille des tables correspond au nombre de lignes exportées à partir d'IBM SPSS Modeler.

- **Options supplémentaires du module de chargement.** Spécifie les arguments supplémentaires servant au programme du module de chargement. Pour les arguments contenant des espaces, utilisez des guillemets doubles.

Pour utiliser des guillemets doubles dans les arguments facultatifs, accompagnez-les d'une barre oblique inverse. Prenons l'exemple de l'option `-comment "Ceci est un \"commentaire\""` qui comprend l'indicateur `-comment` et le commentaire à proprement parler, à savoir `Ceci est un "commentaire"`.

Vous pouvez utiliser une barre oblique inverse à condition de l'accompagner d'une autre barre oblique inverse. Prenons l'exemple de l'option `-specialdir "C:\\Test Scripts\\"` qui comprend l'indicateur `-specialdir` et le répertoire, à savoir `C:\Test Scripts\`.

Programmation de chargement en bloc

Le noeud d'exportation de base de données comporte des options de chargement en bloc dans la boîte de dialogue Options avancées. Les programmes de module de chargement en bloc permettent de charger les données d'un fichier texte dans une base de données.

L'option **Utiliser le chargement en bloc - Via le module de chargement externe** configure l'application IBM SPSS Modeler de sorte qu'elle exécute les trois opérations suivantes :

- Création des tables de base de données requises.
- Exportation des données vers un fichier texte.
- Appel d'un programme de module de chargement en bloc pour charger les données du fichier dans la table de base de données.

En général, le programme de module de chargement en bloc ne correspond pas à l'utilitaire de chargement de base de données proprement dit (par exemple, l'utilitaire sqlldr d'Oracle) ; il s'agit en fait d'un petit script ou programme qui créé les arguments corrects et les fichiers auxiliaires propres aux bases de données (comme un fichier de contrôle), puis appelle l'utilitaire de chargement de base de données. Les sections suivantes vous expliquent comment éditer un module de chargement en bloc existant.

Vous pouvez aussi écrire votre propre programme de chargement en bloc. Pour plus d'informations, voir la rubrique «Développement de programmes de chargement en bloc», à la page 378. Notez que cette opération n'est pas couverte dans le cadre d'un contrat de support technique standard et que vous devez prendre contact avec un technicien de maintenance IBM pour de l'aide.

Scripts destinés au chargement en bloc.

IBM SPSS Modeler est fourni avec plusieurs programmes de chargement en bloc qui correspondent aux différentes bases de données implémentées utilisant les scripts Python. Lorsque vous exécutez un flux contenant un noeud d'exportation de base de données et que l'option **Via le module de chargement externe** est sélectionnée, IBM SPSS Modeler crée la table de base de données (si nécessaire) via ODBC, exporte les données vers un fichier temporaire sur l'hôte exécutant IBM SPSS Modeler Server, puis invoque le script de chargement en bloc. Ensuite, ce script exécute des utilitaires fournis par le fournisseur SGBD afin de charger les données des fichiers temporaires vers la base de données.

Remarque : L'installation d'IBM SPSS Modeler ne comprend pas d'interpréteur d'exécution Python, par conséquent une installation distincte de Python est nécessaire. Pour plus d'informations, voir la rubrique «Options avancées d'exportation de base de données», à la page 372.

Les scripts fournis (disponibles dans le dossier *\scripts* du répertoire d'installation IBM SPSS Modeler) pour les bases de données figurent dans le tableau suivant.

Tableau 47. Scripts de chargement en bloc fournis.

Base de données	Nom du script	Informations complémentaires
IBM Db2	`db2_loader.py`	Pour plus d'informations, voir la rubrique «Chargement en bloc de données vers les bases de données IBM Db2».
IBM Netezza	`netezza_loader.py`	Pour plus d'informations, voir la rubrique «Chargement en bloc de données vers les bases de données IBM Netezza», à la page 376.
Oracle	`oracle_loader.py`	Pour plus d'informations, voir la rubrique «Chargement en bloc de données vers les bases de données Oracle», à la page 376.
SQL Server	`mssql_loader.py`	Pour plus d'informations, voir la rubrique «Chargement en bloc de données vers les bases de données SQL Server», à la page 377.
Teradata	`teradata_loader.py`	Pour plus d'informations, voir la rubrique «Chargement en bloc de données vers les bases de données Teradata», à la page 378.

Chargement en bloc de données vers les bases de données IBM Db2

Les points suivants peuvent vous aider à configurer le chargement en bloc à partir d'IBM SPSS Modeler vers une base de données IBM Db2 à l'aide de l'option Module de chargement externe située dans la boîte de dialogue Export SGBD - Options avancées.

Vérifiez que l'utilitaire du processeur de ligne de commande Db2 (CLP) est installé

Le script `db2_loader.py` appelle la commande Db2 LOAD. Vérifiez que le processeur de ligne de commande (`db2` sous UNIX, `db2cmd` sous Windows) est installé sur le serveur qui procède à l'exécution de `db2_loader.py` (généralement, l'hôte exécutant IBM SPSS Modeler Server).

Vérifiez que le nom d'alias de la base de données locale est le même que le nom réel de la base de données

Le nom d'alias de la base de données locale Db2 est le nom utilisé par le logiciel client Db2 pour faire référence à une base de données sur une instance DB2 locale ou distante. Si l'alias de la base de données locale est différent du nom de la base de données distante, utilisez l'option du module de chargement supplémentaire :

```
-alias <local_database_alias>
```

Par exemple, la base de données distante est nommée STARS sur l'hôte GALAXY mais l'alias de la base de données locale Db2 sur l'hôte exécutant IBM SPSS Modeler Server est STARS_GALAXY. Utilisez l'option du module de chargement supplémentaire

```
-alias STARS_GALAXY
```

Codage des données de caractères non-ASCII

Si vous chargez en bloc des données dont le format n'est pas ASCII, vous devez vous assurer que la variable codepage de la section de configuration de `db2_loader.py` est correctement définie sur votre système.

Chaînes vides

Les chaînes vides sont exportées vers la base de données en tant que valeurs NULL.

Chargement en bloc de données vers les bases de données IBM Netezza

Les points suivants peuvent vous aider à configurer le chargement en bloc à partir d'IBM SPSS Modeler vers une base de données IBM Netezza à l'aide de l'option Module de chargement externe située dans la boîte de dialogue Export SGBD - Options avancées.

Vérifiez que l'utilitaire Netezza nzload est installé

Le script *netezza_loader.py* invoque l'utilitaire Netezza *nzload*. Vérifiez que *nzload* est installé et correctement configuré sur le serveur qui va exécuter *netezza_loader.py*.

Exportation de données non-ASCII

Si votre exportation contient des données dont le format n'est pas ASCII, vous devrez peut-être ajouter `-encoding UTF8` au champ **Options supplémentaires du module de chargement** de la boîte de dialogue Export SGBD - Options avancées. Ceci garantit que les données non-ASCII sont correctement chargées.

Données aux formats de date, d'heure et d'horodatage

Dans les propriétés du flux, définissez le format de date sur **JJ-MM-AAAA** et le format d'heure sur **HH:MM:SS**.

Chaînes vides

Les chaînes vides sont exportées vers la base de données en tant que valeurs NULL.

Ordres des colonnes du flux et de la table cible différents lors de l'insertion de données dans une table existante

Si l'ordre des colonnes du flux est différent de celui de la table cible, les valeurs des données ne seront pas insérées dans les bonnes colonnes. Utilisez un noeud Re-trier pour garantir que l'ordre des colonnes du flux correspond à l'ordre de la table cible. Pour plus d'informations, voir «Noeud Re-trier», à la page 192.

Suivi de la progression de nzload

Lorsque IBM SPSS Modeler est exécuté en mode local, ajoutez `-sts` au champ **Options supplémentaires du module de chargement** dans la boîte de dialogue Export SGBD - Option avancées, afin d'afficher des messages indiquant le statut de progression toutes les 1000 lignes dans la fenêtre de commande ouverte par l'utilitaire *nzload*.

Chargement en bloc de données vers les bases de données Oracle

Les points suivants peuvent vous aider à configurer le chargement en bloc à partir d'IBM SPSS Modeler vers une base de données Oracle à l'aide de l'option Module de chargement externe située dans la boîte de dialogue Export SGBD - Options avancées.

Vérifiez que l'utilitaire Oracle sqlldr est installé

Le script *oracle_loader.py* invoque l'utilitaire Oracle *sqlldr*. Remarque : l'utilitaire *sqlldr* n'est pas inclus de façon automatique dans le client Oracle. Vérifiez que *sqlldr* est installé sur le serveur qui va exécuter *oracle_loader.py*.

Indiquez le SID de la base de données ou le nom du service

Si vous exportez des données vers un serveur Oracle non local ou que votre serveur local Oracle comprend plusieurs bases de données, vous devrez spécifier les éléments suivants dans le champ

Options supplémentaires du module de chargement situé dans la boîte de dialogue Export SGBD - Options avancées afin de transmettre le SID ou le nom du service :

```
-database <SID>
```

Modification de la section configuration dans oracle_loader.py

Sous UNIX (et éventuellement sous Windows), modifiez la section configuration située au début du script *oracle_loader.py*. Ici, les valeurs pour les variables d'environnement ORACLE_SID, NLS_LANG, TNS_ADMIN et ORACLE_HOME peuvent être spécifiées le cas échéant, de même que le chemin d'accès complet de l'utilitaire *sqlldr*.

Données aux formats de date, d'heure et d'horodatage

Dans les propriétés du flux, définissez le format de date sur **AAAA-MM-JJ** et le format d'heure sur **HH:MM:SS**.

Si vous avez besoin d'utiliser un format de date et d'heure différent de celui indiqué ci-dessus, consultez votre documentation oracle et modifiez le fichier du script *oracle_loader.py*.

Codage des données de caractères non-ASCII

Si vous chargez en bloc des données dont le format n'est pas ASCII, vous devez vous assurer que la variable d'environnement NLS_LANG est correctement définie sur votre système. Cette variable est lue par l'utilitaire Oracle de chargement *sqlldr*. Par exemple, la valeur correcte de NLS_LANG pour Shift-JIS sous Windows est Japanese_Japan.JA16SJIS. Pour plus d'informations sur NLS_LANG, consultez votre documentation Oracle.

Chaînes vides

Les chaînes vides sont exportées vers la base de données en tant que valeurs NULL.

Chargement en bloc de données vers les bases de données SQL Server

Les points suivants peuvent vous aider à configurer le chargement en bloc à partir d'IBM SPSS Modeler vers une base de données SQL Server à l'aide de l'option Module de chargement externe située dans la boîte de dialogue Exportation de la base de données : options avancées.

Vérifiez que l'utilitaire SQL Server bcp.exe est installé
Le script *mssql_loader.py* invoque l'utilitaire SQL Server *bcp.exe*. Vérifiez que *bcp.exe* est installé sur le serveur qui va exécuter *mssql_loader.py*.

L'utilisation d'espaces comme délimiteur ne fonctionne pas
Evitez de choisir l'espace comme délimiteur dans la boîte de dialogue Exportation de la base de données : options avancées.

Option Vérifier la taille de la table recommandée
Il est recommandé d'activer l'option **Vérifier la taille de la table** dans la boîte de dialogue Exportation de la base de données : options avancées. Les échecs dans le processus de chargement en bloc ne sont pas toujours détectés et l'activation de cette option permet de procéder à une vérification supplémentaire pour déterminer si le nombre correct de lignes a été chargé.

Chaînes vides
Les chaînes vides sont exportées vers la base de données en tant que valeurs NULL.

Spécifiez l'instance nommée du serveur SQL complète
Parfois, lorsque SPSS Modeler ne peut pas accéder à SQL Server car le nom d'hôte n'est pas qualifié, le message suivant peut s'afficher :

```
Erreur rencontrée lors de l'exécution du programme de chargement en bloc externe. Le fichier journal est susceptibl
détails.
```

Pour corriger cette erreur, ajoutez la chaîne suivante, avec les guillemets, dans le champ **Options supplémentaires du module de chargement** :

```
"-S mhreboot.spss.com\SQLEXPRESS"
```

Chargement en bloc de données vers les bases de données Teradata

Les points suivants peuvent vous aider à configurer le chargement en bloc à partir d'IBM SPSS Modeler vers une base de données Teradata à l'aide de l'option Module de chargement externe située dans la boîte de dialogue Export SGBD - Options avancées.

Vérifiez que l'utilitaire Teradata fastload est installé

Le script *teradata_loader.py* invoque l'utilitaire Teradata *fastload*. Vérifiez que *fastload* est installé et correctement configuré sur le serveur qui va exécuter *teradata_loader.py*.

Chargement en bloc des donnés possible uniquement vers des tables vides

Seules des tables vides peuvent être utilisées comme cibles d'un chargement en bloc. Si une table cible contient déjà des données avant le chargement en bloc, l'opération échoue.

Données aux formats de date, d'heure et d'horodatage

Dans les propriétés du flux, définissez le format de date sur **AAAA-MM-JJ** et le format d'heure sur **HH:MM:SS**.

Chaînes vides

Les chaînes vides sont exportées vers la base de données en tant que valeurs NULL.

ID de processus Teradata (tdpid)

Par défaut, *fastload* exporte les données vers le système Teradata avec `tdpid=dbc`. Généralement, il existe une entrée dans le fichier HOSTS qui associe `dbccop1` à l'adresse IP du serveur Teradata. Pour utiliser un serveur différent, spécifiez les éléments suivants dans le champ **Options supplémentaires du module de chargement** de la boîte de dialogue Export SGBD - Options avancées, afin de transmettre le `tdpid` du serveur :

```
-tdpid <id>
```

Espaces dans les noms des tables et colonnes

Si les noms des tables et colonnes contiennent des espaces, l'opération de chargement en bloc échoue. Si possible, renommez les tables ou colonnes pour supprimer les espaces.

Développement de programmes de chargement en bloc

Cette section explique comment développer un programme de module de chargement en bloc pouvant être exécuté dans IBM SPSS Modeler pour charger des données à partir d'un fichier texte vers une base de données. Notez que cette opération n'est pas couverte dans le cadre d'un contrat de support technique standard et que vous devez prendre contact avec un technicien de maintenance IBM pour de l'aide.

Utilisation de Python pour créer des programmes de module de chargement en bloc

Par défaut, IBM SPSS Modeler recherche un programme de module de chargement en bloc en fonction du type de la base de données. Voir tableau 47, à la page 375.

Le script *test_loader.py* peut pour vous aider dans le développement de programmes de module de chargement en bloc. Pour plus d'informations, voir la rubrique «Test des programmes de module de chargement en bloc», à la page 380.

Objets transmis au programme de module de chargement en bloc

IBM SPSS Modeler crée deux fichiers qui sont transmis au programme de module de chargement en bloc.

- **Fichier de données.** Il contient les donnés à charger au format texte.
- **Fichier de schéma.** Ce fichier est au format XML. Il décrit les noms et types des colonnes et fournit des informations sur le format des données (par exemple, quel caractère sert de délimiteur entre les champs).

En outre, IBM SPSS Modeler transmet d'autres informations telles que le nom de la table, le nom et le mot de passe utilisateur sous forme d'arguments lors de l'invocation du programme de module de chargement en bloc.

Remarque : pour indiquer la réussite de l'opération à IBM SPSS Modeler, le programme de module de chargement en bloc doit supprimer le fichier de schéma.

Arguments transmis au programme de module de chargement en bloc

Les arguments transmis au programme sont répertoriés dans le tableau ci-après.

Tableau 48. Arguments transmis au module de chargement en bloc.

Argument	Description
`schemafile`	Chemin du fichier de schéma.
`data file`	Chemin du fichier de données.
`servername`	Nom du serveur DBMS ; peut être vide.
`databasename`	Nom de la base de données sur le serveur DBMS ; peut être vide.
`username`	Nom d'utilisateur servant à la connexion à la base de données.
`password`	Mot de passe servant à la connexion à la base de données.
`tablename`	Nom de la table à charger.
`ownername`	Nom du propriétaire de la table (aussi appelé nom du schéma).
`logfilename`	Nom du fichier journal (s'il est laissé vide, aucun fichier journal n'est créé).
`rowcount`	Nombre de lignes dans le jeu de données.

Toutes les options spécifiées dans le champ **Options supplémentaires du module de chargement** de la boîte de dialogue Export SGBD - Options avancées, sont transmises au programme de module de chargement en bloc après ces arguments standard.

Format du fichier de données.

Les données sont écrites dans le fichier de données au format texte, chaque champ étant séparé par un caractère délimiteur spécifié dans la boîte de dialogue Export SGBD - Options avancées. L'exemple suivant indique la manière dont un fichier de données séparé par des tabulations doit apparaître.

```
48  F  HIGH    NORMAL  0.692623  0.055369  drugA
15  M  NORMAL  HIGH    0.678247  0.040851  drugY
37  M  HIGH    NORMAL  0.538192  0.069780  drugA
35  F  HIGH    HIGH    0.635680  0.068481  drugA
```

Le fichier est codé à l'aide du codage local utilisé par IBM SPSS Modeler Server (ou IBM SPSS Modeler si le système n'est pas relié à IBM SPSS Modeler Server). Une partie du format est contrôlée par les paramètres de flux IBM SPSS Modeler.

Format du fichier de schéma.

Le fichier de schéma est un fichier XML qui décrit le fichier de données. Voici un exemple du fichier de schéma qui accompagnerait le fichier de données précédent.

```
<?xml version="1.0" encoding="UTF-8" ?>
<DBSCHEMA version="1.0">
 <table delimiter="\t" commit_every="10000" date_format="YYYY-MM-DD" time_format="HH:MM:SS"
append_existing="false" delete_datafile="false">
  <column name="Age" encoded_name="416765" type="integer"/>
  <column name="Sex" encoded_name="536578" type="char" size="1"/>
  <column name="BP" encoded_name="4250" type="char" size="6"/>
  <column name="Cholesterol" encoded_name="43686F6C65737465726F6C" type="char" size="6"/>
  <column name="Na" encoded_name="4E61" type="real"/>
  <column name="K" encoded_name="4B" type="real"/>
  <column name="Drug" encoded_name="44727567" type="char" size="5"/>
 </table>
</DBSCHEMA>
```

Les deux tableaux suivants répertorient les attributs des éléments `<table>` and `<column>` du fichier de schéma.

Tableau 49. Attributs de l'élément <table>.

Attribut	Description
`delimiter`	Le caractère délimiteur des champs (TAB est représenté par \t).
`commit_every`	L'intervalle de taille du lot (tel qu'indiqué dans la boîte de dialogue Export SGBD - Options avancées).
`date_format`	Le format utilisé pour représenter les dates.
`time_format`	Le format utilisé pour représenter les heures.
`append_existing`	`true` si la table chargée contient déjà des données ; sinon `false`
`delete_datafile`	`true` si le programme de chargement en bloc doit supprimer le fichier de données après la fin du chargement.

Tableau 50. Attributs de l'élément <column>.

Attribut	Description
`name`	Nom de la colonne.
`encoded_name`	Le nom de la colonne converti dans le même codage que celui du fichier de données et affiché sous la forme d'une série de nombres hexadécimaux à deux chiffres.
`type`	Type de données de la colonne : un parmi `integer`, `real`, `char`, `time`, `date`, et `datetime`.
`size`	Pour le type de données `char` , la largeur maximale de la colonne en caractères.

Test des programmes de module de chargement en bloc

Vous pouvez tester le chargement en bloc à l'aide d'un script de test *test_loader.py* disponible dans le dossier *\scripts* du répertoire d'installation IBM SPSS Modeler. Il est utile de procéder à un test lors du développement, du débogage ou du dépannage de programmes de chargement en bloc ou de scripts à utiliser avec IBM SPSS Modeler.

Pour utiliser le script de test, suivez la procédure suivante.

1. Exécutez le script *test_loader.py* pour copier les fichiers de schéma et de données vers les fichiers *schema.xml* et *data.txt*, et créez un fichier de commande Windows (*test.bat*).
2. Modifiez le fichier *test.bat* pour sélectionner le programme de module de chargement en bloc ou le script à tester.
3. Exécutez *test.bat* depuis un shell de commande pour tester le programme de module de chargement en bloc ou le script choisi.

Remarque : l'exécution de *test.bat* ne charge pas réellement les données vers la base de données.

Noeud d'exportation Fichier à plat

Le noeud d'exportation Fichier à plat vous permet d'écrire des données dans un fichier texte délimité. Il est utile pour exporter des données pouvant aussi être lues par d'autres logiciels d'analyse ou par des tableurs.

Si vos données contiennent des informations géospatiales, vous pouvez les exporter sous forme de fichier à plat et, si vous générez un noeud source Délimité en vue d'une utilisation avec le même flux, toutes les métadonnées de stockage, de mesure et géospatiales sont conservées sur le nouveau noeud source. Toutefois, si vous exportez les données, puis les importez dans un flux différent, vous devez effectuer des opérations supplémentaires pour définir les données géospatiales sur le nouveau noeud source. Pour plus d'informations, voir la rubrique «Noeud Délimité», à la page 27.

Remarque : Vous ne pouvez pas écrire de fichiers dans l'ancien format de cache car IBM SPSS Modeler ne l'utilise plus pour les fichiers cache. Les fichiers cache IBM SPSS Modeler sont désormais sauvegardés au format IBM SPSS Statistics *.sav*, que vous pouvez écrire à l'aide d'un noeud d'exportation Statistiques. Pour plus d'informations, voir la rubrique «Noeud Exporter Statistics», à la page 382.

Noeud Fichier plat - Onglet Exporter

Exporter le fichier. Indique le nom du fichier. Entrez directement le nom ou cliquez sur le sélecteur de fichiers pour accéder à l'emplacement du fichier.

Mode écriture. Si vous sélectionnez **Remplacer**, les éventuelles données présentes dans le fichier indiqué seront écrasées. Si vous sélectionnez **Ajouter**, la sortie sera ajoutée à la fin du fichier et les données existantes conservées.

- **Inclure les noms des champs.** Si vous sélectionnez cette option, les noms des champs figurent sur la première ligne du fichier de sortie. Cette option est disponible uniquement avec le mode d'écriture **Remplacer**.

Nouvelle ligne après chaque enregistrement. Si vous sélectionnez cette option, chaque enregistrement est écrit sur une nouvelle ligne dans le fichier de sortie.

Séparateur de champs. Indique le caractère à insérer entre les valeurs des champs dans le fichier texte généré. Les options sont les suivantes : **Virgule**, **Tabulation**, **Espace** et **Autre**. Si vous sélectionnez **Autre**, entrez les caractères de séparation souhaités dans la zone de texte.

Guillemets. Indique le type de guillemet à utiliser pour les valeurs des champs symboliques. Les options disponibles sont les suivantes : **Aucun** (valeurs non accompagnées de guillemets), **Simple (')**, **Double (")** et **Autre**. Si vous sélectionnez **Autre**, entrez le type de guillemet souhaité dans la zone de texte.

Codage. Indique la méthode de codage de texte employée. Vous pouvez choisir la valeur par défaut du système, la valeur par défaut du flux ou UTF-8.

- Si le système est exécuté en mode réparti, sa valeur par défaut est spécifiée dans le Panneau de configuration de Windows de l'ordinateur serveur.
- La valeur par défaut du flux est spécifiée dans la boîte de dialogue Propriétés du flux.

Symbole décimal. Indique le mode de représentation des décimales dans les données.

- **Flux par défaut.** Le séparateur décimal défini par défaut pour le flux actuel est utilisé. Il s'agit généralement du séparateur décimal défini dans les paramètres régionaux de l'ordinateur.
- **Poinr (.).** Le point est utilisé comme séparateur décimal.
- **Virgule (,).** La virgule est utilisée comme séparateur décimal.

Générer un noeud source pour ces données. Sélectionnez cette option afin de générer automatiquement un noeud source Délimité pour lire le fichier de données exporté. Pour plus d'informations, voir «Noeud Délimité», à la page 27.

Noeud Exporter Statistics

Le noeud Exporter Statistics vous permet d'exporter les données au format IBM SPSS Statistics *.sav*. Les fichiers IBM SPSS Statistics *.sav* peuvent être lus par IBM SPSS Statistics Base et d'autres modules. Ce format est également utilisé pour les fichiers cache IBM SPSS Modeler.

Il est possible que le mappage des noms de champ IBM SPSS Modeler à des noms de variable IBM SPSS Statistics génère des erreurs, car ces noms de variable IBM SPSS Statistics sont limités à 64 caractères et ne peuvent pas inclure certains caractères, comme l'espace, le signe dollar ($), et le tiret (–). Vous pouvez ajuster ces restrictions de deux façons :

- Vous pouvez renommer les champs en respectant les conventions de dénomination des variables IBM SPSS Statistics en cliquant sur l'onglet Filtrer. Pour plus d'informations, voir «Changement du nom ou filtrage des champs pour IBM SPSS Statistics», à la page 383.
- Exportez les noms et libellés de champ à partir d'IBM SPSS Modeler.

Remarque : IBM SPSS Modeler écrit les fichiers *.sav* au format Unicode UTF-8. IBM SPSS Statistics prend en charge uniquement les fichiers au format Unicode UTF-8 de la version 16.0 et versions supérieures. Pour limiter les risques de corruption des données, les fichiers *.sav* enregistrés avec le codage Unicode ne doivent pas être utilisés avec les versions de IBM SPSS Statistics antérieures à 16.0. Pour plus d'informations, reportez-vous à l'aide de IBM SPSS Statistics.

Ensembles de réponses multiples. Tous les ensembles de réponses multiples définis dans le flux seront automatiquement préservés lors de l'exportation du fichier. Vous pouvez afficher et modifier les ensembles de réponses multiples dans n'importe quel noeud avec un onglet Filtrer. Pour plus d'informations, voir «Modification des ensembles de réponses multiples», à la page 161.

Noeud Export Statistics - Onglet Exporter

Exporter le fichier Indique le nom du fichier. Entrez directement le nom ou cliquez sur le sélecteur de fichiers pour accéder à l'emplacement du fichier.

Type de fichier Sélectionnez cette option si le fichier doit être enregistré au format normal (*.sav*) ou compressé (*.zsav*).

Chiffrer le fichier avec mot de passe Cochez cette case pour protéger le fichier à l'aide d'un mot de passe ; vous êtes ensuite invité à saisir et confirmer le **Mot de passe** dans une boîte de dialogue distincte.

Remarque : Les fichiers protégés par mot de passe ne peuvent être ouverts que par SPSS Modeler version 16 ou supérieure, ou par SPSS Statistics version 21 ou supérieure.

Exporter les noms de champ en tant que variables Indique une méthode de gestion des noms et libellés de variable lors de l'exportation d'SPSS Modeler vers un fichier SPSS Statistics *.sav* ou *.zsav*.

- **Noms et libellés de variable** Sélectionnez cette option pour exporter les noms et les libellés de champs SPSS Modeler. Les noms sont exportés en tant que noms de variable SPSS Statistics, alors que les libellés le sont en tant que libellés de variable SPSS Statistics.

- **Noms en tant que libellés de variable** Sélectionnez cette option pour utiliser les noms de champ SPSS Modeler en tant que libellés de variable dans SPSS Statistics. SPSS Modeler autorise dans les noms de champ des caractères non valides dans les noms de variable SPSS Statistics. Pour éviter de créer des noms SPSS Statistics incorrects, sélectionnez **Libellés** ou utilisez l'onglet Filtre pour ajuster les noms de champ.

Lancer Application Si SPSS Statistics est installé sur votre ordinateur, vous pouvez sélectionner cette option pour exécuter l'application directement sur le fichier de données enregistré. Les options de lancement de cette application doivent être indiquées dans la boîte de dialogue Programmes externes. Pour plus d'informations, voir «Programmes externes de IBM SPSS Statistics», à la page 363. Pour créer simplement un fichier SPSS Statistics *.sav* ou *.zsav* sans ouvrir de programme externe, désélectionnez cette option.

Remarque : Si vous exécutez SPSS Modeler et SPSS Statistics ensemble en mode serveur (distribué), l'écriture de données en sortie et le lancement d'une session SPSS Statistics n'ouvre pas automatiquement un client SPSS Statistics affichant le fichier lu dans le fichier actif. La solution consiste à ouvrir manuellement le fichier de données dans le client SPSS Statistics une fois que ce dernier a été lancé.

Générer un noeud source pour ces données Sélectionnez cette option afin de générer automatiquement un noeud source Fichier statistiques pour lire le fichier de données exporté. Pour plus d'informations, voir «Noeud Statistics», à la page 33.

Changement du nom ou filtrage des champs pour IBM SPSS Statistics

Avant d'exporter ou de déployer des données d'IBM SPSS Modeler vers des applications externes telles que IBM SPSS Statistics, vous pouvez être amené à renommer ou à ajuster des noms de champ. Les boîtes de dialogue Transformation Statistics, Sortie Statistics et Exporter Statistics contiennent un onglet Filtrer pour faciliter ce processus.

Vous trouverez dans une autre rubrique une brève description de l'onglet Filtrer. Pour plus d'informations, voir «Définition des options de filtrage», à la page 159. Cette rubrique fournit des astuces concernant la lecture des données dans IBM SPSS Statistics.

Pour ajuster les noms de fichiers afin de se conformer aux conventions de dénomination de IBM SPSS Statistics :

1. Dans l'onglet Filtrer, cliquez sur le bouton de la barre d'outils du menu des options de filtrage (le premier de la barre d'outils).
2. Sélectionnez Renommer pour IBM SPSS Statistics.
3. Dans la boîte de dialogue Renommer pour IBM SPSS Statistics, vous pouvez choisir de remplacer les caractères non valides des noms de fichier soit par un caractère **dièse (#)**, soit parc un **caractère de soulignement (_)**.

Renommer des ensembles à réponses multiples. Sélectionnez cette option si vous souhaitez ajuster le nom de plusieurs ensembles a réponses multiples, lesquels peuvent être importés dans IBM SPSS Modeler à l'aide d'un noeud source Statistics. Ils sont utilisés pour enregistrer des données qui peuvent comporter plus d'une valeur pour chaque cas, telles que les réponses à une enquête.

Noeud d'exportation Data Collection

Le noeud Export Data Collection enregistre des données au format utilisé par les logiciels d'étude de marché Data Collection, en fonction du modèle Data Collection Data Model. Ce format fait la distinction entre les données d'observation (réponses réelles fournies à des questions et collectées au cours d'une enquête) et les métadonnées qui décrivent le mode de collecte et d'organisation des données d'observation. Les métadonnées consistent en des informations diverses : texte des questions, nom et

description de variables, ensemble de réponses multiples, traduction des différents textes et définition de la structure des données d'observation. Pour plus d'informations, voir «Data Collection Noeud», à la page 34.

Fichier de métadonnées. Indique le nom du fichier de définition du questionnaire (*.mdd*) dans lequel les métadonnées exportées seront enregistrées. Un questionnaire par défaut est créé en fonction des informations de type de champ. Par exemple, un champ nominal (ensemble) peut être représenté sous la forme d'une question unique, avec la description du champ utilisée comme texte de la question et une case à cocher distincte pour chaque valeur définie.

Fusionner les métadonnées. Indique si les métadonnées remplaceront les versions existantes ou seront fusionnées avec les métadonnées existantes. Si l'option de fusion est sélectionnée, une nouvelle version est créée à chaque exécution du flux. Ceci permet le suivi des versions d'un questionnaire au fil des changements. Chaque version peut être considérée comme un instantané des métadonnées utilisées pour collecter un ensemble précis de données d'observation.

Activer les variables système. Indique si les variables système sont incluses dans le fichier *.mdd* exporté. Il s'agit de variables telles que *Respondent.Serial*, *Respondent.Origin*, et *DataCollection.StartTime*.

Paramètres des données d'observation. Indique le fichier de données IBM SPSS Statistics (*.sav*) dans lequel les données d'observation sont exportées. Notez que toutes les restrictions sur les noms de variable et de valeur s'appliquent ici, vous pouvez avoir besoin de basculer vers l'onglet Filtrer et d'utiliser l'option "Renommer pour IBM SPSS Statistics" du menu des options de filtrage pour corriger les caractères non valides dans les noms de champ.

Générer un noeud source pour ces données. Sélectionnez cette option afin de générer automatiquement un noeud source Data Collection pour lire le fichier de données exporté.

Ensembles de réponses multiples. Tous les ensembles de réponses multiples définis dans le flux seront automatiquement préservés lors de l'exportation du fichier. Vous pouvez afficher et modifier les ensembles de réponses multiples dans n'importe quel noeud avec un onglet Filtrer. Pour plus d'informations, voir «Modification des ensembles de réponses multiples», à la page 161.

Noeud d'exportation Analytic Server

Le noeud d'exportation Analytic Server vous permet d'écrire des données provenant de votre analyse dans une source de données Analytic Server existante. Il peut s'agir par exemple de fichiers texte sur Hadoop Distributed File System (HDFS) ou d'une base de données.

En général, un flux avec un noeud d'exportation Analytic Server commence également avec des noeuds de source Analytic Server. Il est envoyé au Analytic Server et exécuté sur HDFS. Un flux avec des sources de données "locales" peut également terminer avec un noeud d'exportation Analytic Server pour exporter des jeux de données relativement réduits (pas plus de 100 000 enregistrements) pour une utilisation avec le Analytic Server.

Si vous souhaitez utiliser votre propre connexion Analytic Server à la place de la connexion par défaut définie par votre administrateur, désélectionnez **Utiliser le serveur Analytic Server par défaut** et sélectionnez votre connexion. Pour des informations sur la configuration de plusieurs connexions Analytic Server, voir Connexion à Analytic Server.

Source de données. Sélectionnez une source de données contenant les données que vous souhaitez utiliser. Une source de données contient les fichiers et métadonnées associés à cette source. Cliquez sur **Sélectionner** pour afficher une liste des sources de données disponibles. Pour plus d'informations, reportez-vous à la rubrique «Sélection d'une source de données», à la page 13.

Si vous avez besoin de créer une nouvelle source de données ou d'éditer une source de données existante, cliquez sur **Launch Data Source Editor...**.

Mode. Sélectionnez **Ajouter** pour ajouter des données à la source de données existante ou **Remplacer** pour remplacer le contenu de la source de données.

Générer un noeud d'importation pour ces données. Sélectionnez cette option pour générer un noeud de source pour les données, comme exporté vers la source de données spécifiée. Ce noeud est ajouté au canevas de flux.

Notez que l'utilisation de plusieurs connexions Analytic Server peut s'avérer utile pour contrôler le flux de données. Par exemple, si vous utilisez les noeuds Source et Exportation d'Analytic Server, vous pouvez utiliser des connexions Analytic Server dans différentes branches d'un flux afin que lorsque chaque branche est exécutée, elle utilise son propre serveur Analytic Server, sans extraction de données dans IBM SPSS Modeler Server. Notez que si une branche contient plusieurs connexions Analytic Server, les données seront extraites des serveurs Analytic Server vers IBM SPSS Modeler Server. Pour plus d'informations, notamment sur les restrictions, voir Propriétés du flux Analytic Server.

Noeud d'exportation IBM Cognos

Le noeud d'exportation IBM Cognos permet d'exporter des données d'un flux IBM SPSS Modeler vers Cognos Analytics, au format UTF-8. Ainsi, Cognos peut utiliser des données transformées ou évaluées d'IBM SPSS Modeler. Par exemple, vous pouvez utiliser Cognos Report Studio pour créer un rapport basé sur les données exportées, qui contient les prévisions et les valeurs de confiance. Le rapport peut ensuite être enregistré sur le serveur Cognos et distribué aux utilisateurs de Cognos.

Remarque : Vous pouvez uniquement exporter des données relationnelles et pas de données OLAP.

Pour exporter des données vers Cognos, vous devez spécifier les paramètres suivants :

- Connexion Cognos - la connexion au serveur Cognos Analytics (version 11 ou ultérieure prise en charge)
- Connexion ODCB - la connexion au serveur de données Cognos que le serveur Cognos utilise

Dans la connexion Cognos, vous spécifiez une source de données Cognos à utiliser. Cette source de données doit utiliser la même connexion que la source de données ODBC.

Vous exportez les données de flux vers le serveur de données et les métadonnées du package vers le serveur Cognos.

Comme avec n'importe quel autre noeud d'exportation, vous pouvez également utiliser l'onglet Publier de la boîte de dialogue du noeud pour publier le flux et le déployer à l'aide de IBM SPSS Modeler Solution Publisher.

Remarque : Le noeud source Cognos ne prend en charge que les packages Cognos CQM. Les packages DQM ne sont pas pris en charge.

Connexion Cognos

Vous spécifiez à cet endroit la connexion au serveur Cognos Analytics (version 11 ou ultérieure prise en charge) que vous souhaitez utiliser pour l'exportation. Cette procédure se compose de l'exportation des métadonnées vers un nouveau package sur le serveur Cognos pendant que les données de flux sont exportées vers le serveur de données Cognos.

Connexion. Cliquez sur le bouton **Modifier** pour afficher une boîte de dialogue dans laquelle vous pourrez définir l'URL et les autres informations sur le serveur Cognos vers lequel vous souhaitez exporter

les données. Si vous êtes déjà connecté à un serveur Cognos via IBM SPSS Modeler, vous pouvez également modifier les détails de la connexion actuelle. Consultez «Connexions Cognos», à la page 42 pour plus d'informations.

Source de données. Le nom de la source de données Cognos (généralement une base de données) vers laquelle vous exportez les données. La liste déroulante indique toutes les sources de données Cognos auxquelles vous pouvez accéder à partir de la connexion actuelle. Cliquez sur le bouton **Rafraîchir** pour mettre à jour la liste.

Dossier. Le chemin d'accès et le nom du dossier du serveur Cognos où le package d'exportation doit être créé.

Nom du package. Le nom du package dans le dossier spécifié qui doit contenir les métadonnées exportées. Il doit s'agir d'un nouveau package avec un seul objet de requête ; l'exportation ne peut pas se faire vers un package existant.

Mode. Spécifie les modalités de l'exportation :

- **Publier le package maintenant.** (par défaut) Effectue l'exportation dès que vous cliquez sur **Exécuter**.
- **Exporter le script d'action.** Crée un script XML que vous pourrez exécuter ultérieurement (par exemple, à l'aide de Framework Manager) pour effectuer l'exportation. Saisissez le chemin d'accès et le nom du fichier du script dans le champ **Fichier** ou utilisez le bouton **Modifier** pour spécifier le nom et l'emplacement du fichier du script.

Générer un noeud source pour ces données. Sélectionnez cette option pour générer un noeud source pour les données lors de leur exportation dans la table et la source de données spécifiées. Dès que vous cliquez sur **Exécuter**, ce noeud est ajouté à l'espace de travail de flux.

Connexion ODBC

C'est ici que vous spécifiez la connexion au serveur de données Cognos (c'est-à-dire la base de données) vers lequel les données de flux vont être exportées.

Remarque : Vous devez vous assurer que la source des données spécifiée ici pointe vers la même source que celle spécifiée dans le volet **Connexions Cognos**. Vous devez également vous assurer que la source de données de connexion Cognos utilise la même source de données que la source de données ODBC.

Source de données. Source de données sélectionnée. Entrez directement son nom ou sélectionnez-le dans la liste déroulante. Si la base de données souhaitée n'apparaît pas dans la liste, sélectionnez **Ajouter une nouvelle connexion à la base de données** et localisez votre base de données dans la boîte de dialogue Connexions de base de données. Pour plus d'informations, voir «Ajout d'une connexion de base de données», à la page 20.

Nom de la table - Entrez le nom de la table vers laquelle envoyer les données. Si vous sélectionnez l'option **Insérer dans la table**, vous pouvez choisir une table existante dans la base de données en cliquant sur le bouton **Sélectionner**.

Créer une table. Sélectionnez cette option pour créer une nouvelle table de base de données ou écraser une table de base de données existante.

Insérer dans la table. Sélectionnez cette option pour insérer les données dans de nouvelles lignes d'une table de base de données existante.

Fusionner la table. (Le cas échéant) Sélectionnez cette option pour mettre à jour les colonnes de la base de données sélectionnées avec des valeurs de champs de données source correspondants. Sélectionner cette option active le bouton **Fusionner**, qui affiche une boîte de dialogue dans laquelle vous pouvez mapper les champs de données source sur les colonnes de la base de données.

Supprimer la table existante. Sélectionnez cette option pour supprimer, le cas échéant, une table existante du même nom que la table créée.

Supprimer des lignes existantes. Sélectionnez cette option pour supprimer les lignes existantes de la table avant l'exportation, lors de l'insertion dans une table.

Remarque : Si vous sélectionnez l'une des deux options ci-dessus, vous recevez le message **Avertissement d'écrasement** lors de l'exécution du noeud. Pour que ces avertissements n'apparaissent plus, désélectionnez l'option **Avertir lorsqu'un noeud écrase une table de base de données** dans l'onglet Notifications de la boîte de dialogue Options utilisateur.

Taille de chaîne par défaut. Les champs marqués comme étant "sans type" dans un noeud Typer en amont sont écrits dans la base de données sous forme de champs de type chaîne. Indiquez la taille des chaînes à utiliser pour les champs sans type.

Cliquez sur **Schéma** pour ouvrir une boîte de dialogue dans laquelle vous pouvez définir diverses options d'exportation (pour les bases de données prenant en charge cette fonctionnalité), définir des types de données SQL pour vos champs et spécifier la clé primaire en vue de l'indexation de base de données. Pour plus d'informations, voir «Options de schéma d'exportation de base de données», à la page 368.

Cliquez sur **Index** pour définir les options d'indexation de la table exportée, afin d'améliorer les performances de la base de données. Pour plus d'informations, voir «Export SGBD - Options de l'index», à la page 371.

Cliquez sur **Options avancées** pour spécifier les options de chargement en bloc et de validation de base de données. Pour plus d'informations, voir «Options avancées d'exportation de base de données», à la page 372.

Entourer de guillemets les noms des tables et colonnes. Sélectionnez les options à utiliser lors de l'envoi d'une instruction `CREATE TABLE` à la base de données. Les tableaux ou colonnes comportant des espaces ou des caractères spéciaux doivent être mis entre guillemets.

- **Selon les besoins.** Sélectionnez cette option pour qu'IBM SPSS Modeler détermine automatiquement, au cas par cas, la nécessité d'utiliser des guillemets.
- **Toujours.** Sélectionnez cette option pour que les noms de tableau et de colonne soient systématiquement mis entre guillemets.
- **Jamais.** Sélectionnez cette option pour désactiver l'utilisation des guillemets.

Générer un noeud source pour ces données. Sélectionnez cette option pour générer un noeud source pour les données lors de leur exportation dans la table et la source de données spécifiées. Dès que vous cliquez sur **Exécuter**, ce noeud est ajouté à l'espace de travail de flux.

Noeud d'exportation IBM Cognos TM1

Le noeud d'exportation IBM Cognos permet d'exporter des données d'un flux SPSS Modeler vers Cognos TM1. Ainsi, Cognos Analytics peut utiliser des données transformées ou évaluées d'SPSS Modeler.

Remarque : Vous pouvez exporter des mesures mais vous ne pouvez pas exporter de données de dimension contextuelle ; vous pouvez également ajouter de nouveaux éléments au cube.

Pour exporter des données vers Cognos Analytics (version 11 ou ultérieure prise en charge), vous devez spécifier les paramètres suivants :

- La connexion au serveur Cognos TM1.
- Le cube vers lequel les données seront exportées.
- Le mappage des noms de données SPSS vers les dimensions et mesures TM1 équivalentes.

Remarque : L'utilisateur TM1 doit disposer des droits suivants : privilège d'écriture de cubes, privilège de lecture des dimensions et privilège d'écriture d'éléments de dimension. En outre, IBM Cognos TM1 10.2, groupe de correctifs 3 ou ultérieur, est requis pour que SPSS Modeler puisse importer et exporter des données Cognos TM1. Les flux existants basés sur des versions antérieures continueront de fonctionner.

Les informations d'identification de l'administrateur ne sont pas requises pour ce noeud. Elles le sont toutefois si vous utilisez toujours l'ancien noeud TM1 existant antérieur à la version 17.1.

Comme avec n'importe quel autre noeud d'exportation, vous pouvez également utiliser l'onglet Publier de la boîte de dialogue du noeud pour publier le flux et le déployer à l'aide de IBM SPSS Modeler Solution Publisher.

Remarque : Pour pouvoir utiliser les noeuds source ou d'exportation TM1 dans SPSS Modeler, vous devez vérifier certains paramètres dans le fichier `tm1s.cfg`. Il s'agit du fichier de configuration du serveur TM1 dans le répertoire racine du serveur TM1.

- HTTPPortNumber - définissez un numéro de port valide, généralement 1 à 65535. Notez qu'il ne s'agit pas du numéro de port que vous avez par la suite spécifié dans la connexion dans le noeud, mais d'un port interne utilisé par TM1, qui est désactivé par défaut. Si nécessaire, contactez votre administrateur afin qu'il vous indique le paramètre valide pour ce port.
- UseSSL - si vous le définissez sur *True*, HTTPS est utilisé comme protocole de transport. Dans ce cas, vous devez importer la certification TM1 vers l'environnement d'exécution Java SPSS Modeler Server.

Connexion à un cube IBM Cognos TM1 pour l'exportation des données

La première étape d'exportation des données dans une base de données IBM Cognos TM1 consiste à sélectionner l'hôte d'administration TM1 pertinent, ainsi que le serveur et le cube associés, dans l'onglet **Connexion** de la boîte de dialogue IBM Cognos TM1.

Remarque : Seules les valeurs "null" réelles seront supprimées lors de l'exportation des données vers TM1. Les valeurs zéro (0) seront exportées en tant que valeurs valides. Notez également que seuls les champs dont le type de stockage est *chaîne* peuvent être mappés à des dimensions dans l'onglet Mappage. Avant de procéder à l'exportation vers TM1, vous devez utiliser le client IBM SPSS Modeler pour convertir les types de données autres que chaîne en types de données chaîne.

Hôte admin Entrez l'adresse URL de l'hôte d'administration où le serveur TM1 auquel vous voulez vous connecter est installé. L'hôte d'administration est défini en tant qu'adresse URL unique pour tous les serveurs TM1. A partir de cette adresse URL, tous les serveurs IBM Cognos TM1 installés et s'exécutant dans votre environnement peuvent être reconnus et sont accessibles.

Serveur TM1 Une fois que vous avez établi la connexion à l'hôte d'administration,sélectionnez le serveur qui contient les données à importer et cliquez sur **Connexion**. Si vous ne vous êtes pas déjà connecté à ce serveur, vous êtes invité à entrer votre **Nom d'utilisateur** et votre **Mot de passe**. Sinon, vous pouvez rechercher les détails de connexion déjà entrés que vous avez sauvegardés en tant que **Données d'identification stockées**.

Sélectionnez un cube TM1 à exporter Affiche le nom des cubes du serveur TM1 vers lesquels vous pouvez exporter des données.

Pour choisir les données à exporter, sélectionnez le cube et cliquez sur la flèche droite pour déplacer le cube dans le champ **Exporter vers un cube**. Une fois que vous avez sélectionné le cube, utilisez l'onglet Mappage afin de mapper les dimensions et les mesures TM1 aux champs SPSS pertinents ou à une valeur fixe (opération *Sélectionner*).

Mappage de données IBM Cognos TM1 pour l'exportation

Après avoir sélectionné l'hôte d'administration TM1 ainsi que le cube et le serveur TM1 associés, utilisez l'onglet Mappage de la boîte de dialogue Exportation IBM Cognos TM1 pour mapper les dimensions et les mesures TM1 à des champs SPSS ou définir des dimensions TM1 en les associant à une valeur fixe.

Remarque : Seuls les champs dont le type de stockage est *chaîne* peuvent être mappés à des dimensions. Avant de procéder à l'exportation vers TM1, vous devez utiliser le client IBM SPSS Modeler pour convertir les types de données autres que chaîne en types de données chaîne.

Champs Répertorie les noms de champ de données figurant dans le fichier de données SPSS qui sont disponibles pour l'exportation.

TM1 Dimensions Affiche le cube TM1 sélectionné dans l'onglet Connexion, avec ses dimensions de mesure, ses dimensions ordinaires et les éléments de la dimension de mesure sélectionnée. Sélectionnez le nom de la dimension TM1 ou sa mesure pour la mapper au champ de données SPSS.

Les options ci-dessous sont disponibles dans l'onglet Mappage.

Sélectionner une dimension de mesure Sélectionnez une dimension de mesure dans la liste des dimensions du cube sélectionné.

Lorsque vous sélectionnez une dimension, sauf la dimension de mesure, et cliquez sur **Sélectionner**, une boîte de dialogue affichant les membres feuilles de la dimension sélectionnée s'ouvre. Vous ne pouvez sélectionner que les éléments feuilles. Les éléments sélectionnés ont le libellé **S**.

Mappe Mappe le champ de données SPSS sélectionné à la mesure ou à la dimension TM1 sélectionnée (une dimension ordinaire ou une mesure spécifique de la dimension de mesure). Les champs mappés sont associés à la lettre **M**.

Démapper Annule le mappage du champ de données SPSS sélectionné à la mesure ou la dimension TM1. Vous ne pouvez annuler le mappage que d'un seul mappage à la fois. Le champ de données SPSS non mappé revient dans la colonne de gauche.

Créer Crée une mesure dans la dimension de mesure TM1. Une boîte de dialogue s'affiche dans laquelle vous entrez le nouveau **nom de mesure TM1**. Cette option est disponible seulement pour les dimensions de mesure, et non pour les dimensions ordinaires.

Pour plus d'informations sur TM1, voir la documentation d'IBM Cognos TM1 à l'adresse http://www-01.ibm.com/support/knowledgecenter/SS9RXT_10.2.2/com.ibm.swg.ba.cognos.ctm1.doc/welcome.html.

Noeud Export SAS

Cette fonction est disponible dans SPSS Modeler Professional et SPSS Modeler Premium.

Le noeud Export SAS permet d'écrire les données au format SAS afin qu'elles puissent être lues par SAS ou par un logiciel compatible. Vous pouvez exporter dans trois formats de fichier SAS : SAS pour Windows/OS2, SAS pour UNIX ou SAS.

Noeud Export SAS - Onglet Exporter

Exporter le fichier. Indiquez le nom du fichier. Entrez directement le nom ou cliquez sur le sélecteur de fichiers pour accéder à l'emplacement du fichier.

Exporter. Indiquez le format du fichier d'exportation. Les options disponibles sont les suivantes : **SAS pour Windows/OS2**, **SAS pour UNIX** ou **SAS Version 7/8/9**.

Exporter les noms de champ en tant que variables. Sélectionnez les options d'exportation des noms et des libellés de champ depuis IBM SPSS Modeler pour leur utilisation avec SAS.

- **Noms et libellés de variable.** Sélectionnez cette option pour exporter les noms et les libellés de champs IBM SPSS Modeler. Les noms sont exportés en tant que noms de variable SAS, alors que les libellés le sont en tant que libellés de variable SAS.
- **Noms en tant que libellés de variable.** Sélectionnez cette option pour utiliser les noms de champ IBM SPSS Modeler en tant que libellés de variable dans SAS. Les noms de champ IBM SPSS Modeler prennent en charge des caractères non valides dans les noms de variable SAS. Pour éviter de créer des noms SAS incorrects, sélectionnez **noms** à la place.

Générer un noeud source pour ces données. Sélectionnez cette option afin de générer automatiquement un noeud source SAS pour lire le fichier de données exporté. Pour plus d'informations, voir «Noeud source SAS», à la page 45.

Remarque : La longueur maximum autorisée d'une chaîne est de 255 octets. Si une chaîne dépasse 255 octets, elle sera tronquée lors de l'exportation.

Noeud Export Excel

Le noeud Export Excel génère une sortie de données au format de fichier Microsoft Excel (.xlsx). Si vous le souhaitez, vous pouvez choisir de lancer Excel automatiquement et d'ouvrir le fichier exporté lors de l'exécution du noeud.

Noeud Excel - Onglet Exporter

Nom du fichier. Entrez directement le nom ou cliquez sur le sélecteur de fichiers pour accéder à l'emplacement du fichier. Le nom de fichier par défaut est *excelexp.xlsx*.

Type de fichier. Le format de fichier Excel .xlsx est pris en charge.

Créer un fichier. Crée un nouveau fichier Excel.

Insérer dans un fichier existant. Le contenu est remplacé dès le début de la cellule désignée par le champ **Démarrer dans la cellule**. Les autres cellules de la feuille de calcul sont laissées avec leur contenu d'origine.

Inclure les noms des champs. Indique si les noms de champ doivent être inclus dans la première ligne de la feuille de calcul.

Démarrer dans la cellule. L'emplacement des cellules est utilisé pour le premier enregistrement d'exportation (ou le premier nom de champ si **Inclure les noms de champ** est sélectionné). Les données sont remplies à droite et en bas de la cellule d'origine.

Choisir une feuille de calcul. Spécifie la feuille de calcul dans laquelle vous souhaitez exporter les données. Vous pouvez identifier la feuille de calcul, via un index ou un nom.

- **Par index** Si vous créez un nouveau fichier, spécifiez un nombre de 0 à 9 pour identifier la feuille de calcul dans laquelle vous souhaitez exporter, en commençant par 0 pour la première feuille de calcul, 1 pour la seconde feuille de calcul, etc. Vous pouvez utiliser des valeurs de 10 ou plus uniquement si une feuille de calcul existe déjà à cette position.
- **Par nom.** Si vous créez un nouveau fichier, spécifiez le nom utilisé pour la feuille de calcul. Si vous effectuez une insertion dans un fichier existant, les données sont insérées dans cette feuille de calcul si elle existe, ou une nouvelle feuille de calcul avec ce nom est créée.

Lancer Excel. Indique si Excel est lancé automatiquement sur le fichier exporté lors de l'exécution du noeud. Dans le cas d'une exécution en mode réparti dans IBM SPSS Modeler Server, la sortie est enregistrée dans le système de fichiers du serveur et Excel est lancé sur le client avec une copie du fichier exporté.

Générer un noeud source pour ces données. Sélectionnez cette option afin de générer automatiquement un noeud source Excel pour lire le fichier de données exporté. Pour plus d'informations, voir «Noeud source Excel», à la page 46.

---

noeud Exportation d'extension

Avec le noeud Exportation d'extension, vous pouvez exécuter des scripts R ou Python for Spark pour exporter des données.

Noeud Exportation d'extension - Onglet Syntaxe

Sélectionnez le type de syntaxe – **R** ou **Python for Spark**. Consultez les sections suivantes pour plus d'informations. Lorsque votre syntaxe est prête, vous pouvez cliquer sur **Exécuter** pour exécuter le noeud Exportation d'extension.

Syntaxe R

Syntaxe R. Permet d'entrer ou de coller la syntaxe de script R personnalisé en vue de l'analyse des données dans ce champ.

Convertir les champs indicateurs. Indique comment sont traités les champs indicateurs. Deux options sont disponibles : **Chaînes en facteur, Entiers et Réels en double** et **Valeurs logiques (True, False)**. Si vous sélectionnez **Valeurs logiques (True, False)**, les valeurs originales des champs indicateurs sont perdues. Par exemple, si un champ a les valeurs Mâle et Femelle, elles sont remplacées par True et False.

Convertir les valeurs manquantes en valeur R 'non disponible' (NA). Lorsque cette option est sélectionnée, toute valeur manquante est convertie en valeur R `NA`. La valeur `NA` est utilisée par R pour identifier les valeurs manquantes. Des fonctions R que vous utilisez peuvent comporter un argument par le biais duquel il est possible de contrôler le comportement des fonctions lorsque les données contiennent `NA`. Par exemple, la fonction peut vous permettre de choisir d'exclure automatiquement les enregistrements qui contiennent `NA`. Si cette option n'est pas sélectionnée, les valeurs manquantes sont transmises à R en l'état et peuvent entraîner des erreurs lors de l'exécution du script R.

Convertir les champs date/heure en classes R avec contrôle spécial pour les fuseaux horaires. Lorsque cette option est sélectionnée, les variables de format de date et de date/heure sont converties en objets R date/heure. Vous devez sélectionner l'une des options suivantes :

- **R POSIXct.** Les variables de format de date ou de date/heure sont converties en objets R `POSIXct`.
- **R POSIXlt (liste).** Les variables de format de date ou de date/heure sont converties en objets R `POSIXlt`.

Remarque : Les formats POSIX sont des options avancées. Utilisez-les uniquement si le script R spécifie que les champs date/heure sont traités de telle manière que ces formats sont requis. Les formats POSIX ne s'appliquent pas aux variables de format horaire.

Syntaxe Python

Syntaxe Python. Permet d'entrer ou de coller la syntaxe de script Python personnalisé en vue de l'analyse des données dans ce champ. Pour plus d'informations relatives à Python for Spark, voir Python for Spark et Scriptage avec Python for Spark.

Noeud Exportation d'extension - Onglet Sortie de la console

L'onglet **Sortie de la console** contient les sorties reçues lorsque le script R ou le script Python for Spark de l'onglet Syntaxe est exécuté (par exemple, si un script R est utilisé, il affiche la sortie reçue de la console R lorsque le script R du champ **Syntaxe R** de l'onglet **Syntaxe** est exécuté). La sortie peut contenir des messages d'erreur ou d'avertissement R ou Python générés lors de l'exécution du script R ou Python. Cette sortie permet essentiellement de déboguer le script. L'onglet **Sortie de la console** contient également le script du champ **Syntaxe R** ou **Syntaxe Python**.

A chaque exécution du script Exportation d'extension, le contenu de l'onglet **Sortie de la console** est écrasé par la sortie reçue de la console R ou Python for Spark. La sortie ne peut pas être éditée.

Noeud Export XML

Le noeud Export XML vous permet d'exporter des données au format XML à l'aide de l'encodage UTF-8. Vous pouvez également créer un noeud source XML pour lire de nouveau les données exportées dans le flux.

Exporter le fichier XML. Le chemin complet et le nom du fichier XML dans lequel vous souhaitez exporter les données.

Utiliser un schéma XML. Cochez cette case si vous souhaitez utiliser un schéma ou DTD pour contrôler la structure des données exportées. Si vous la cochez, vous activez le bouton **Mapper**, décrit ci-dessous.

Si vous n'utilisez pas de schéma ou DTD, la structure par défaut suivante est utilisée pour les données exportées :

```
<records>
     <record>
          <fieldname1>value</fieldname1>
          <fieldname2>value</fieldname2>
      :
          <fieldnameN>value</fieldnameN>
     </record>
     <record>
      :
      :
     </record>
      :
      :
</records>
```

Les espaces dans un nom de champ sont remplacés par des caractères de soulignement ; par exemple, "Mon champ" devient `<Mon_champ>`.

Mappe. Si vous avez choisi d'utiliser un schéma XML, ce bouton ouvre une boîte de dialogue dans laquelle vous pouvez spécifier la partie de la structure XML à utiliser pour commencer chaque nouvel enregistrement. Pour plus d'informations, voir «Mappage XML - Options Enregistrements», à la page 393.

Champs mappés. Indique le nombre de champs ayant été mappés.

Générer un noeud source pour ces données. Sélectionnez cette option afin de générer automatiquement un noeud source XML pour lire le fichier de données exporté. Pour plus d'informations, voir «Noeud source XML», à la page 47.

Ecrire des données XML

Lorsqu'un élément XML est spécifié, la valeur du champ est placée à l'intérieur de la balise de l'élément :

```
<element>value</element>
```

Lorsqu'un attribut est mappé, la valeur du champ est placée en tant que valeur pour l'attribut :

```
<element attribute="value">
```

Si un champ est mappé sur un élément au-dessus de l'élément `<records>`, le champ n'est écrit qu'une seule fois et représente une constante pour tous les enregistrements. La valeur de cet élément provient du premier enregistrement.

Si une valeur nulle doit être écrite, ceci est réalisé en spécifiant un contenu vide. Pour les éléments, il s'agit de :

```
<element></element>
```

Pour les attributs, il s'agit de :

```
<element attribute="">
```

Mappage XML - Options Enregistrements

L'onglet Enregistrements vous permet de spécifier la partie de la structure XML à utiliser pour commencer chaque nouvel enregistrement. Afin de procéder correctement au mappage sur un schéma, vous devez spécifier le délimiteur d'enregistrement.

Structure XML. Un arbre hiérarchique montrant la structure du schéma XML spécifié dans l'écran précédent.

Enregistrements (expression XPath). Pour définir le délimiteur d'enregistrement, sélectionnez un élément dans la structure XML et cliquez sur le bouton de la flèche droite. A chaque fois que cet élément est rencontré dans les données source, un nouvel enregistrement est créé dans le fichier de résultat.

Remarque : Si vous sélectionnez l'élément racine de la structure XML, un seul enregistrement peut être écrit, et tous les autres enregistrements sont ignorés.

Mappage XML - Options Champs

L'onglet Champs permet de mapper des champs du jeu de données sur des éléments ou des attributs dans la structure XML lorsqu'un fichier de schéma est utilisé.

Les noms de fichier qui correspondent à un nom d'élément ou d'attribut sont automatiquement mappés tant que le nom d'élément ou d'attribut est unique. Par conséquent, s'il y a un élément et un attribut portant le nom `champ1`, il n'y a pas de mappage automatique. S'il n'y a qu'un seul élément de la structure nommé `champ1`, un champ portant ce nom dans le flux est automatiquement mappé.

Champs. La liste des champs utilisés dans le modèle. Sélectionnez un ou plusieurs champs comme partie source du mappage. Vous pouvez utiliser les boutons situés en bas de la liste pour sélectionner tous les fichiers ou tous les champs ayant un niveau de mesure particulier.

Structure XML. Sélectionnez un élément de la structure XML en tant que cible de mappage. Pour créer le mappage, cliquez sur Mapper. Le mappage s'affiche alors. Le nombre de champs ayant été mappés de cette manière s'affiche en dessous de cette liste.

Pour supprimer un mappage, sélectionnez l'élément dans la liste de la structure XML et cliquez sur **Démapper**.

Afficher les attributs. Affiche ou masque les attributs, le cas échéant, des éléments XML dans la structure XML.

Mappage XML - Aperçu

Dans l'onglet Aperçu, cliquez sur **Mettre à jour** pour voir un aperçu du XML qui sera écrit.

Si le mappage n'est pas correct, revenez à l'onglet Enregistrements ou Champs pour corriger les erreurs et cliquez de nouveau sur **Mettre à jour** pour voir le résultat.

Onglets communs aux noeuds d'exportation

Les options suivantes peuvent être spécifiées pour tous les noeuds d'exportation en cliquant sur l'onglet correspondant :

- **Onglet Publier.** Permet de publier les résultats d'un flux.
- **Onglet Annotations.** Utilisé pour tous les noeuds, cet onglet propose des options permettant de renommer les noeuds, de créer des info-bulles personnalisées et de stocker de longues annotations.

Publication de flux

La publication des flux s'effectue directement à partir d'IBM SPSS Modeler à l'aide de n'importe quel noeud d'exportation standard : Base de données, Fichier à plat, Exportation Statistiques, Exportation d'extension, Exportation de collection de données, Exportation SAS, Excel et Exportation XML. Le type de noeud d'exportation détermine le format des résultats à écrire chaque fois que le flux publié est exécuté à l'aide d'IBM SPSS Modeler Solution Publisher Runtime ou d'une application externe. A titre d'exemple, pour enregistrer vos résultats dans une base de données chaque fois que le flux publié est exécuté, utilisez un noeud exportation de base de données.

Pour publier un flux

1. Ouvrez ou créez un flux, puis reliez un noeud d'exportation à la fin du flux.
2. dans l'onglet Publier du noeud d'exportation, spécifiez un nom racine pour les fichiers publiés (à savoir, le nom de fichier auquel les extensions `.pim`, `.par` et `.xml` seront ajoutées).
3. Cliquez sur **Publier** pour publier le flux ou sélectionnez **Publier le flux** pour publier automatiquement le flux chaque fois que le noeud est exécuté.

Nom publié. Spécifiez le nom racine de l'image publiée et les fichiers de paramètres.

- Le **fichier image** (`*.pim`) fournit toutes les informations nécessaires à l'exécution par Runtime du flux publié, exactement tel qu'il était au moment de l'exportation. Si vous ne pensez pas devoir changer de paramètres du flux (comme la source des données d'entrée ou le fichier des données de sortie), vous pouvez vous contenter de déployer le fichier image.
- Le **fichier de paramètres** (`*.par`) contient des informations configurables sur les sources de données, les fichiers de sortie et les options d'exécution. Pour pouvoir contrôler l'entrée ou la sortie du flux sans republier ce dernier, vous avez besoin du fichier de paramètres et du fichier image.
- Le **fichier de métadonnées** (`*.xml`) décrit les entrées et les sorties de l'image et de leurs modèles de données. Il est conçu pour être utilisé par des applications qui incorporent la bibliothèque d'exécution et qui ont besoin de connaître la structure des données d'entrée et de sortie.

  Remarque : ce fichier n'est généré que si vous sélectionnez l'option **Publier les métadonnées**.

Publication des paramètres. Si nécessaire, vous pouvez inclure les paramètres de flux dans le fichier `*.par`. Vous pouvez modifier les valeurs des paramètres de flux lorsque vous exécutez l'image soit en modifiant le fichier `*.par` ou en utilisant l'API d'exécution.

Cette option active le bouton **Paramètres**. La boîte de dialogue Publier les paramètres s'affiche lorsque vous cliquez sur ce bouton.

Choisissez les paramètres à inclure dans l'image publiée en sélectionnant l'option appropriée dans la colonne **Publier**.

Exécution dans le flux. Indique si le flux est publié automatiquement lorsque le noeud est exécuté.

- **Exporter les données.** Exécute le noeud d'exportation de manière standard, sans publier le flux. (Essentiellement, le noeud est exécuté dans IBM SPSS Modeler de la même manière que si IBM SPSS Modeler Solution Publisher n'était pas disponible.) Si vous sélectionnez cette option, le flux n'est pas publié à moins que vous ne le publiez explicitement en cliquant sur **Publier** dans la boîte de dialogue du noeud d'exportation. Vous pouvez également publier le flux en cours à l'aide de l'outil Publier de la barre d'outils ou à l'aide d'un script.
- **Publier le flux.** Publie le flux en vue de le déployer à l'aide de IBM SPSS Modeler Solution Publisher. Sélectionnez cette option pour publier automatiquement le flux chaque fois qu'il est exécuté.

Remarque :

- Si vous comptez exécuter le flux publié avec de nouvelles données ou des données mises à jour, n'oubliez pas que l'ordre des champs du fichier d'entrée doit être identique à celui des champs du fichier d'entrée du noeud source spécifié dans le flux publié.
- Lorsque vous publiez des données vers des applications externes, vous pouvez être amené à filtrer les champs superflus ou à renommer des champs pour respecter les exigences de saisie. Pour ce faire, utilisez le noeud Filtrer avant le noeud d'exportation.

Chapitre 8. IBM SPSS Statistics Noeuds

Noeuds IBM SPSS Statistics - Présentation

Pour compléter IBM SPSS Modeler et ses capacités d'exploration de données, IBM SPSS Statistics vous permet d'effectuer des analyses statistiques et une gestion de données plus avancées.

Lorsque vous avez installé une copie compatible de IBM SPSS Statistics avec sa licence, vous pouvez le connecter à partir d'IBM SPSS Modeler et effectuer une manipulation et des analyses de données complexes et en plusieurs étapes qu'IBM SPSS Modeler ne prend habituellement pas en charge. Pour les utilisateurs avancés, il existe également une option permettant de modifier l'analyse en utilisant la syntaxe de commande. Consulter les notes de version pour obtenir des informations concernant la compatibilité de version.

S'ils sont disponibles, les noeuds IBM SPSS Statistics apparaissent dans une partie spécifique de la palette des noeuds.

Remarque : Nous vous recommandons d'instancier vos données dans un noeud Typer avant d'utiliser les noeuds Transformer, Modèle ou Sortie de IBM SPSS Statistics. Ceci est aussi nécessaire lors de l'utilisation de la commande de syntaxe AUTORECODE.

La palette IBM SPSS Statistics comporte les noeuds suivants :

Le noeud Fichier Statistics lit les données du format de fichier *.sav* ou *.zsav* utilisé par IBM SPSS Statistics, ainsi que des fichiers cache enregistrés dans IBM SPSS Modeler, qui utilisent le même format.

Le noeud Transformation exécute une sélection de commandes de syntaxe IBM SPSS Statistics en fonction des sources de données dans IBM SPSS Modeler. Ce noeud requiert une copie avec licence de IBM SPSS Statistics.

Le noeud Modèle Statistics vous permet d'analyser et de travailler avec vos données en exécutant des procédures IBM SPSS Statistics qui produisent un PMML. Ce noeud requiert une copie avec licence de IBM SPSS Statistics.

Le noeud Sortie Statistics vous permet d'appeler une procédure IBM SPSS Statistics pour analyser les données IBM SPSS Modeler. De nombreuses procédures d'analyses IBM SPSS Statistics sont disponibles. Ce noeud requiert une copie avec licence de IBM SPSS Statistics.

Le noeud Export Statistics génère des données au format IBM SPSS Statistics *.sav* ou *.zsav*. Les fichiers *.sav* ou *.zsav* peuvent être lus par IBM SPSS Statistics Base et d'autres produits. Ce format est également utilisé pour les fichiers cache IBM SPSS Modeler.

Remarque : Si votre copie de SPSS Statistics possède une licence mono-utilisateur et que vous exécutez un flux avec deux branches ou plus, chacune d'elles contenant un noeud SPSS Statistics, vous pourriez recevoir une erreur de licence. Cette erreur survient lorsque la session SPSS Statistics d'une branche n'est pas terminée alors que la session d'une autre branche tente de démarrer. Si possible, revoyez la conception du flux de manière à ce qu'il ne présente pas plusieurs branches contenant des noeuds SPSS Statistics et s'exécutant en parallèle.

Noeud Statistics

Vous pouvez utiliser le noeud Fichier Statistics pour lire des données directement à partir d'un fichier IBM SPSS Statistics enregistré (*.sav* ou *.zsav*). Ce format remplace le format de fichier cache qui était utilisé dans les versions précédentes d'IBM SPSS Modeler. Si vous souhaitez importer un fichier cache enregistré, vous devez utiliser un noeud IBM SPSS Statistics.

Importer le fichier. Indiquez le nom du fichier. Vous pouvez entrer un nom de fichier ou cliquer sur le bouton représentant des points de suspension (**...**) pour sélectionner un fichier. Le chemin d'accès du fichier apparaît une fois le fichier sélectionné.

Le fichier est chiffré par mot de passe. Cochez cette case si vous savez que le fichier est protégé par mot de passe ; vous êtes ensuite invité à saisir le **Mot de passe**. Si le fichier est protégé par mot de passe et que vous ne le saisissez pas, un message d'avertissement s'affiche lorsque vous tentez d'activer un autre onglet, d'actualiser les données, de prévisualiser le contenu du noeud ou d'exécuter un flux contenant le noeud.

Remarque : Les fichiers protégés par mot de passe ne peuvent être ouverts que par IBM SPSS Modeler version 16 ou supérieure.

Noms des variables. Sélectionnez une méthode de gestion des noms et des libellés de variable lors de l'importation d'un fichier IBM SPSS Statistics *.sav* ou *.zsav*. Les métadonnées que vous incluez sont conservées tout au long de votre travail dans IBM SPSS Modeler. Vous pouvez également les réexporter pour les utiliser dans IBM SPSS Statistics.

- **Lire les noms et les libellés.** Sélectionnez cette option afin de lire les noms et les libellés de variable dans IBM SPSS Modeler. Par défaut, cette option est sélectionnée et les noms de variable affichés dans le noeud Typer. Les libellés peuvent apparaître dans les graphiques, les navigateurs de modèle et d'autres types de sortie, selon les options spécifiées dans la boîte de dialogue des propriétés du flux. Par défaut, l'affichage de libellés dans la sortie est désactivé.
- **Lire les libellés sous forme de nom.** Sélectionnez cette option pour lire les libellés de variable descriptifs du fichier IBM SPSS Statistics *.sav* ou *.zsav* au lieu des noms de champ abrégés, puis utilisez ces libellés en tant que noms de variable dans IBM SPSS Modeler.

Valeurs. Sélectionnez une méthode de gestion des noms et des libellés lors de l'importation d'un fichier IBM SPSS Statistics *.sav* ou *.zsav*. Les métadonnées que vous incluez sont conservées tout au long de votre travail dans IBM SPSS Modeler. Vous pouvez également les réexporter pour les utiliser dans IBM SPSS Statistics.

- **Lire les données et les libellés.** Choisissez cette option pour les valeurs réelles et les libellés de valeur dans IBM SPSS Modeler. Par défaut, cette option est sélectionnée et les valeurs proprement dites

apparaissent dans le noeud Typer. Les libellés de valeur peuvent être affichées dans le Générateur de formules, les navigateurs de modèle et d'autres types de sortie, selon les options spécifiées dans la boîte de dialogue des propriétés du flux.

- **Lire les libellés sous forme de données.** Choisissez cette option si vous préférez utiliser les libellés de valeurs du fichier *.sav* ou *.zsav* plutôt que les codes numériques ou symboliques utilisés pour représenter les valeurs. Par exemple, si vous sélectionnez cette option pour les données dont le champ indiquant le genre a pour valeur 1 et 2 (représentant respectivement *masculin* et *féminin*), le champ sera converti en chaîne, et importera *masculin* et *féminin* comme valeurs réelles.

  Il est important de prendre en compte les valeurs manquantes dans vos données IBM SPSS Statistics avant de choisir cette option. Par exemple, si un champ numérique utilise des libellés uniquement pour les valeurs manquantes (0 = *Pas de réponse*, 99 = *Inconnu*) et que vous sélectionnez l'option ci-dessus, seules les libellés de valeurs *Pas de réponse* et *Inconnu* sont importées, et le champ est converti en chaîne. Dans ce cas, vous devez importer les valeurs elles-mêmes et définir les valeurs manquantes dans un noeud Typer.

Utilisez les informations de formats de champ pour déterminer le stockage. Si cette case est désélectionnée, les valeurs de champs formatées dans le fichier *.sav* en tant qu'entiers (c'est à dire des champs spécifiés comme F*n*.0 dans la Vue des variables d'IBM SPSS Statistics) sont importées à l'aide du stockage d'entier. Toutes les autres valeurs de champ, à l'exception des chaînes, sont importées en tant que nombres réels.

Si cette case est cochée (par défaut), toutes les valeurs de champ, à l'exception des chaînes, sont importées en tant que nombres réels, qu'elles soient formatées dans le fichier *.sav* sous la forme d'entiers ou non.

Ensembles de réponses multiples. Tout ensemble de réponses multiples défini dans le fichier IBM SPSS Statistics est automatiquement conservé lors de l'importation du fichier. Vous pouvez afficher et modifier les ensembles de réponses multiples dans n'importe quel noeud avec un onglet Filtrer. Pour plus d'informations, voir «Modification des ensembles de réponses multiples», à la page 161.

Noeud Transformation Statistics

Le noeud Transformation Statistics vous permet de procéder à des transformations de données grâce à la syntaxe des commandes IBM SPSS Statistics. Il est ainsi possible de procéder à certaines transformations non prises en charge par IBM SPSS Modeler et d'automatiser des transformations complexes en plusieurs étapes, y compris la création d'un certain nombre de champs à partir d'un noeud unique. Ce noeud ressemble au noeud Sortie Statistics à ceci près que les données sont renvoyées à IBM SPSS Modeler pour analyse complémentaire, alors que, dans le noeud Sortie SPSS, les données sont renvoyées sous forme d'objets de sortie requis, tels que des graphiques ou des tableaux.

Vous devez disposer d'une version compatible de IBM SPSS Statistics installée sur votre ordinateur et en détenir la license d'utilisation pour utiliser ce noeud. Pour plus d'informations, voir «Programmes externes de IBM SPSS Statistics», à la page 363. Consulter les notes de version pour obtenir des informations concernant la compatibilité.

Si nécessaire, vous pouvez utiliser l'onglet Filtrer pour filtrer ou renommer des champs afin qu'ils soient conformes aux conventions de dénomination IBM SPSS Statistics. Pour plus d'informations, voir «Changement du nom ou filtrage des champs pour IBM SPSS Statistics», à la page 383.

Référence de la syntaxe des commandes. Pour plus d'informations sur les procédures IBM SPSS Statistics spécifiques, consultez le *IBM SPSS Statisticsguide de référence de la syntaxe des commandes*, fourni avec votre copie du logiciel IBM SPSS Statistics. Pour consulter le guide dans l'onglet Syntaxe, choisissez l'option **Editeur de syntaxe** et cliquez sur le bouton Lancer l'aide syntaxe IBM SPSS Statistics.

Remarque : ce noeud ne prend pas en charge la totalité de la syntaxe IBM SPSS Statistics.. Pour plus d'informations, voir «Syntaxe autorisée», à la page 400.

Noeud Transformation Statistics - Onglet Syntaxe

Option de la boîte de dialogue IBM SPSS Statistics

Si vous n'êtes pas habitué à la syntaxe IBM SPSS Statistics d'une procédure, la façon la plus simple de créer une syntaxe dans IBM SPSS Modeler est de choisir l'option **Boîte de dialogue IBM SPSS Statistics**, de sélectionner la boîte de dialogue de la procédure, suivre ses instructions et cliquer sur OK. Cela vous permet de placer la syntaxe dans l'onglet Syntaxe du noeud IBM SPSS Statistics utilisé dans IBM SPSS Modeler. Vous pouvez ensuite exécuter le flux afin d'obtenir les résultats de la procédure.

Option de l'éditeur de syntaxe IBM SPSS Statistics

Vérifier. Une fois que vous avez saisi vos commandes de syntaxe dans la partie supérieure de la boîte de dialogue, utilisez ce bouton pour valider vos entrées. Toute syntaxe incorrecte est mise en évidence dans la partie inférieure de la boîte de dialogue.

Pour garantir que le processus de vérification n'est pas trop long, lorsque vous validez la syntaxe, une comparaison est effectuée avec un échantillon représentatif de vos données, plutôt qu'avec la totalité du jeu de données, afin d'assurer la validité de vos entrées.

Syntaxe autorisée

Si votre syntaxe est en grande partie héritée de IBM SPSS Statistics ou si vous connaissez les fonctions de préparation des données de IBM SPSS Statistics, vous pouvez utiliser le noeud Transformation Statistics pour exécuter un grand nombre des transformations existantes. En tant qu'instruction, le noeud vous permet de transformer les données de façon prévisible, par exemple en exécutant des commandes en boucle ou en modifiant, ajoutant, triant, filtrant ou sélectionnant des données.

Vous trouverez ci-dessous des exemples de commandes pouvant être exécutées :

- Calculer des nombres aléatoires d'après une loi binomiale :

  ```
  COMPUTE newvar = RV.BINOM(10000,0.1)
  ```

- Recoder une variable en une nouvelle variable :

  ```
  RECODE Age (Lowest thru 30=1) (30 thru 50=2) (50 thru Highest=3) INTO AgeRecoded
  ```

- Remplacer des valeurs manquantes :

```
RMV Age_1=SMEAN(Age)
```

La syntaxe IBM SPSS Statistics prise en charge par le noeud Transformation Statistics est répertoriée ci-après :

Nom de la commande

```
ADD VALUE LABELS
APPLY DICTIONARY
AUTORECODE
BREAK
CD
CLEAR MODEL PROGRAMS
CLEAR TIME PROGRAM
CLEAR TRANSFORMATIONS
COMPUTE
COUNT
CREATE
```

Nom de la commande

DATE

DEFINE-!ENDDEFINE

DELETE VARIABLES

DO IF

DO REPEAT

ELSE

ELSE IF

END CASE

END FILE

END IF

END INPUT PROGRAM

END LOOP

END REPEAT

EXECUTE

FILE HANDLE

FILE LABEL

FILE TYPE-END FILE TYPE

FILTER

FORMATS

IF

INCLUDE

INPUT PROGRAM-END INPUT PROGRAM

INSERT

LEAVE

LOOP-END LOOP

MATRIX-END MATRIX

MISSING VALUES

N OF CASES

NUMERIC

PERMISSIONS

PRESERVE

RANK

RECODE

RENAME VARIABLES

RESTORE

RMV

SAMPLE

SELECT IF

SET

SORT CASES

STRING

Nom de la commande

SUBTITLE

TEMPORARY

TITLE

UPDATE

V2C

VALIDATEDATA

VALUE LABELS

VARIABLE ATTRIBUTE

VARSTOCASES

VECTOR

Noeud Modèle Statistics

Le noeud Modèle Statistics vous permet d'analyser et de travailler avec vos données en exécutant des procédures IBM SPSS Statistics qui produisent un PMML. Les nuggets de modèle que vous créez peuvent ensuite être utilisés de la façon habituelle dans les flux IBM SPSS Modeler pour l'évaluation, etc.

Vous devez disposer d'une version compatible de IBM SPSS Statistics installée sur votre ordinateur et en détenir la license d'utilisation pour utiliser ce noeud. Pour plus d'informations, voir «Programmes externes de IBM SPSS Statistics», à la page 363. Consulter les notes de version pour obtenir des informations concernant la compatibilité.

Les procédures d'analyse IBM SPSS Statistics disponibles dépendent du type de licence que vous possédez.

Noeud Modèle Statistics - Onglet Modèle

Nom du modèle Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

Sélectionnez une boîte de dialogue. Cliquez pour afficher une liste des procédures IBM SPSS Statistics disponibles que vous pouvez sélectionner et exécuter. Cette liste ne contient que les procédures qui produisent PMML et pour laquelle vous disposez d'une licence, et ne contient pas de procédures écrites par l'utilisateur.

1. Cliquez sur la procédure requise ; la boîte de dialogue IBM SPSS Statistics correspondante s'affiche.
2. Dans la boîte de dialogue IBM SPSS Statistics , saisissez les détails de la procédure.
3. Cliquez sur **OK** pour revenir au noeud Modèle Statistics; la syntaxe IBM SPSS Statistics apparaît dans l'onglet Modèle.
4. Pour revenir à la boîte de dialogue IBM SPSS Statistics à tout moment, par exemple pour modifier votre demande, cliquez sur le bouton d'affichage de la boîte de dialogue IBM SPSS Statistics à droite du bouton de sélection des procédures.

Noeud de modèle Statistics - Récapitulatif du nugget de modèle

Lorsque vous exécutez le noeud Modèle Statistics, il exécute la procédure IBM SPSS Statistics associée et crée un nugget de modèle que vous pouvez utiliser dans les flux IBM SPSS Modeler pour l'évaluation.

L'onglet Récapitulatif d'un nugget de modèle affiche des informations sur les champs, les paramètres de création et le processus d'estimation du modèle. Les résultats sont présentés dans un arbre que vous pouvez développer ou réduire en cliquant sur des éléments précis.

Le bouton **Afficher le modèle** affiche les résultats sous une forme modifiée du visualiseur de sortie Output Viewer de IBM SPSS Statistics. Pour des informations supplémentaires sur cet visualiseur, consultez la documentation de IBM SPSS Statistics.

Les options standard d'exportation et d'impression sont disponibles dans le menu Fichier. Pour plus d'informations, voir «Affichage de la sortie», à la page 321.

Noeud Sortie Statistics

Le noeud Sortie Statistics vous permet d'appeler une procédure IBM SPSS Statistics pour analyser les données IBM SPSS Modeler. Vous pouvez visualiser les résultats dans une fenêtre de navigateur ou les enregistrer au format de fichier de sortie IBM SPSS Statistics. IBM SPSS Modeler permet d'accéder à de nombreuses procédures d'analyses IBM SPSS Statistics.

Vous devez disposer d'une version compatible de IBM SPSS Statistics installée sur votre ordinateur et en détenir la license d'utilisation pour utiliser ce noeud. Pour plus d'informations, voir «Programmes externes de IBM SPSS Statistics», à la page 363. Consulter les notes de version pour obtenir des informations concernant la compatibilité.

Si nécessaire, vous pouvez utiliser l'onglet Filtrer pour filtrer ou renommer des champs afin qu'ils soient conformes aux conventions de dénomination IBM SPSS Statistics. Pour plus d'informations, voir «Changement du nom ou filtrage des champs pour IBM SPSS Statistics», à la page 383.

Référence de la syntaxe des commandes. Pour plus d'informations sur les procédures IBM SPSS Statistics spécifiques, consultez le *IBM SPSS Statisticsguide de référence de la syntaxe des commandes*, fourni avec votre copie du logiciel IBM SPSS Statistics. Pour consulter le guide dans l'onglet Syntaxe, choisissez l'option **Editeur de syntaxe** et cliquez sur le bouton Lancer l'aide syntaxe IBM SPSS Statistics.

Noeud Sortie Statistics - Onglet Syntaxe

Utilisez cet onglet pour créer une syntaxe pour la procédure SPSS Statistics que vous souhaitez utiliser pour analyser vos données. La syntaxe se compose de deux éléments : une **instruction** et des **options** associées. L'instruction indique l'analyse ou l'opération à effectuer, et les champs à utiliser. Les options décrivent les autres aspects de l'analyse, tels que les statistiques à afficher, les champs calculés à enregistrer, etc.

Option de la boîte de dialogue SPSS Statistics

Si vous n'êtes pas habitué à la syntaxe IBM SPSS Statistics d'une procédure, la façon la plus simple de créer une syntaxe dans IBM SPSS Modeler est de choisir l'option **Boîte de dialogue IBM SPSS Statistics**, de sélectionner la boîte de dialogue de la procédure, suivre ses instructions et cliquer sur OK. Cela vous permet de placer la syntaxe dans l'onglet Syntaxe du noeud IBM SPSS Statistics utilisé dans IBM SPSS Modeler. Vous pouvez ensuite exécuter le flux afin d'obtenir les résultats de la procédure.

Vous avez la possibilité de générer un noeud source Statistics pour importer les données obtenues. Cela peut être utile, par exemple, si une procédure écrit des champs tels que des scores dans le jeu de données actif en plus d'afficher les résultats.

Remarque :

- Lors de la génération de sorties dans d'autres langues que l'anglais, il est conseillé de spécifier la langue dans la syntaxe.
- L'option de style de sortie n'est pas prise en charge dans le noeud Sorties statistiques.

Pour créer la syntaxe :

1. Cliquez sur le bouton **Sélectionner une boîte de dialogue**.
2. Choisissez une de ces options :
   - **Analyse** Répertorie le contenu du menu Analyse de SPSS Statistics ; sélectionnez la procédure que vous souhaitez utiliser.
   - **Autre** Si elle apparaît, répertorie les boîtes de dialogue créées par Custom Dialog Builder dans SPSS Statistics, ainsi que toutes les autres boîtes de dialogue de SPSS Statistics qui n'apparaissent pas dans le menue Analyse et pour lesquelles vous disposez d'une licence. Si aucune boîte de dialogue n'est concernée, cette option n'apparaît pas.

Remarque : Les boîtes de dialogue Préparation automatique des données n'apparaissent pas.

Si vous disposez d'une boîte de dialogue personnalisée SPSS Statistics qui crée de nouveaux champs, ces champs ne peuvent pas être utilisés dans SPSS Modeler parce que le noeud Sortie Statistics est un noeud terminal.

Vous pouvez aussi cocher la case **Générer un noeud d'importation pour les données obtenues** pour créer un noeud source Statistics à utiliser pour importer les données obtenues vers un autre flux. Le noeud est placé sur l'espace de travail, avec les données contenues dans le fichier `.sav` spécifié dans le champ **Fichier** (l'emplacement par défaut est le répertoire d'installation SPSS Modeler).

Option de l'éditeur de syntaxe

Pour enregistrer la syntaxe créée pour une procédure fréquemment utilisée :

1. Cliquez sur le bouton **Options du fichier** (le premier de la barre d'outils).
2. Sélectionnez **Enregistrer** ou **Enregistrer sous** dans le menu.
3. Enregistrez le fichier en tant que fichier `.sps`.

Pour utiliser des fichiers de syntaxe créés préalablement, en remplaçant le contenu actuel, le cas échéant, de l'éditeur de syntaxe :

1. Cliquez sur le bouton Options du fichier (le premier de la barre d'outils).
2. Dans le menu, sélectionnez **Ouvrir**.
3. Sélectionnez un fichier `.sps` afin de coller son contenu dans l'onglet Syntaxe du noeud Sortie.

Pour insérer une syntaxe préalablement enregistrée sans remplacer le contenu actuel :

1. Cliquez sur le bouton Options du fichier (le premier de la barre d'outils).
2. Dans le menu, sélectionnez **Insérer**.
3. Sélectionnez un fichier `.sps` afin de coller son contenu dans le noeud Sortie au point spécifié par le curseur.

Vous pouvez aussi cocher la case **Générer un noeud d'importation pour les données obtenues** pour créer un noeud source Statistics à utiliser pour importer les données obtenues vers un autre flux. Le noeud est placé sur l'espace de travail, avec les données contenues dans le fichier `.sav` spécifié dans le champ **Fichier** (l'emplacement par défaut est le répertoire d'installation SPSS Modeler).

Lorsque vous cliquez sur **Exécuter**, les résultats sont affichés dans le Visualiseur de résultats SPSS Statistics. Pour plus d'informations sur le visualiseur, consultez la documentation de SPSS Statistics.

Remarque : la syntaxe des éléments suivants (et des options correspondantes dans la boîte de dialogue SPSS Statistics) n'est pas prise en charge. Ils n'ont aucun impact sur la sortie.

- `OUTPUT ACTIVATE`
- `OUTPUT CLOSE`

- OUTPUT DISPLAY
- OUTPUT EXPORT
- OUTPUT MODIFY
- OUTPUT NAME
- OUTPUT NEW
- OUTPUT OPEN
- OUTPUT SAVE

Noeud Sortie Statistics - Onglet Sortie

L'onglet Sortie vous permet de préciser le format et l'emplacement de la sortie. Vous pouvez choisir d'afficher les résultats à l'écran ou de les envoyer vers un des types de fichier disponibles.

Nom de la sortie - Spécifie le nom de la sortie générée lorsque le noeud est exécuté. L'option **Automatique** sélectionne un nom en fonction du noeud qui génère la sortie. Si vous le souhaitez, vous pouvez choisir **Personnalisé** pour indiquer un autre nom.

Sortie à l'écran (option par défaut)**.** Crée un objet de sortie à afficher en ligne. L'objet de sortie apparaît dans l'onglet Sorties de la fenêtre du gestionnaire lors de l'exécution du noeud de sortie.

Sortie dans un fichier. Enregistre la sortie dans un fichier lorsque vous exécutez le noeud. Si vous choisissez cette option, entrez un nom de fichier dans le champ **Nom de fichier** (ou parcourez l'arborescence et indiquez un nom de fichier à l'aide du sélecteur de fichiers), puis sélectionnez un type de fichier.

Type de fichier. Sélectionnez le type de fichier auquel vous souhaitez envoyer le résultat.

- **Document HTML (*.html).** Ecrit le résultat au format HTML.
- **Fichier du visualiseur IBM SPSS Statistics (*.spv).** Ecrit le résultat dans un format qui peut être lu par le visualiseur de résultat de IBM SPSS Statistics.
- **Fichier des Rapports Web IBM SPSS Statistics (*.spw).** Ecrit le résultat au format Web Reports de IBM SPSS Statistics qui peut être publié sur un référentiel IBM SPSS Collaboration and Deployment Services et consulté ultérieurement dans un navigateur Web. Pour plus d'informations, voir «Publier sur le Web», à la page 321.

Remarque : Si vous sélectionnez **Sortie à l'écran**, la directive OMS IBM SPSS Statistics `VIEWER=NO` n'a aucun effet ; de plus, les API de scriptage (*Basic* et *Python SpssClient* module) ne sont pas disponibles dans IBM SPSS Modeler.

Noeud Exporter Statistics

Le noeud Exporter Statistics vous permet d'exporter les données au format IBM SPSS Statistics *.sav*. Les fichiers IBM SPSS Statistics *.sav* peuvent être lus par IBM SPSS Statistics Base et d'autres modules. Ce format est également utilisé pour les fichiers cache IBM SPSS Modeler.

Il est possible que le mappage des noms de champ IBM SPSS Modeler à des noms de variable IBM SPSS Statistics génère des erreurs, car ces noms de variable IBM SPSS Statistics sont limités à 64 caractères et ne peuvent pas inclure certains caractères, comme l'espace, le signe dollar ($), et le tiret (–). Vous pouvez ajuster ces restrictions de deux façons :

- Vous pouvez renommer les champs en respectant les conventions de dénomination des variables IBM SPSS Statistics en cliquant sur l'onglet Filtrer. Pour plus d'informations, voir «Changement du nom ou filtrage des champs pour IBM SPSS Statistics», à la page 383.
- Exportez les noms et libellés de champ à partir d'IBM SPSS Modeler.

Remarque : IBM SPSS Modeler écrit les fichiers *.sav* au format Unicode UTF-8. IBM SPSS Statistics prend en charge uniquement les fichiers au format Unicode UTF-8 de la version 16.0 et versions supérieures. Pour limiter les risques de corruption des données, les fichiers *.sav* enregistrés avec le codage Unicode ne doivent pas être utilisés avec les versions de IBM SPSS Statistics antérieures à 16.0. Pour plus d'informations, reportez-vous à l'aide de IBM SPSS Statistics.

Ensembles de réponses multiples. Tous les ensembles de réponses multiples définis dans le flux seront automatiquement préservés lors de l'exportation du fichier. Vous pouvez afficher et modifier les ensembles de réponses multiples dans n'importe quel noeud avec un onglet Filtrer. Pour plus d'informations, voir «Modification des ensembles de réponses multiples», à la page 161.

Noeud Export Statistics - Onglet Exporter

Exporter le fichier Indique le nom du fichier. Entrez directement le nom ou cliquez sur le sélecteur de fichiers pour accéder à l'emplacement du fichier.

Type de fichier Sélectionnez cette option si le fichier doit être enregistré au format normal (*.sav*) ou compressé (*.zsav*).

Chiffrer le fichier avec mot de passe Cochez cette case pour protéger le fichier à l'aide d'un mot de passe ; vous êtes ensuite invité à saisir et confirmer le **Mot de passe** dans une boîte de dialogue distincte.

Remarque : Les fichiers protégés par mot de passe ne peuvent être ouverts que par SPSS Modeler version 16 ou supérieure, ou par SPSS Statistics version 21 ou supérieure.

Exporter les noms de champ en tant que variables Indique une méthode de gestion des noms et libellés de variable lors de l'exportation d'SPSS Modeler vers un fichier SPSS Statistics *.sav* ou *.zsav*.

- **Noms et libellés de variable** Sélectionnez cette option pour exporter les noms et les libellés de champs SPSS Modeler. Les noms sont exportés en tant que noms de variable SPSS Statistics, alors que les libellés le sont en tant que libellés de variable SPSS Statistics.
- **Noms en tant que libellés de variable** Sélectionnez cette option pour utiliser les noms de champ SPSS Modeler en tant que libellés de variable dans SPSS Statistics. SPSS Modeler autorise dans les noms de champ des caractères non valides dans les noms de variable SPSS Statistics. Pour éviter de créer des noms SPSS Statistics incorrects, sélectionnez **Libellés** ou utilisez l'onglet Filtre pour ajuster les noms de champ.

Lancer Application Si SPSS Statistics est installé sur votre ordinateur, vous pouvez sélectionner cette option pour exécuter l'application directement sur le fichier de données enregistré. Les options de lancement de cette application doivent être indiquées dans la boîte de dialogue Programmes externes. Pour plus d'informations, voir «Programmes externes de IBM SPSS Statistics», à la page 363. Pour créer simplement un fichier SPSS Statistics *.sav* ou *.zsav* sans ouvrir de programme externe, désélectionnez cette option.

Remarque : Si vous exécutez SPSS Modeler et SPSS Statistics ensemble en mode serveur (distribué), l'écriture de données en sortie et le lancement d'une session SPSS Statistics n'ouvre pas automatiquement un client SPSS Statistics affichant le fichier lu dans le fichier actif. La solution consiste à ouvrir manuellement le fichier de données dans le client SPSS Statistics une fois que ce dernier a été lancé.

Générer un noeud source pour ces données Sélectionnez cette option afin de générer automatiquement un noeud source Fichier statistiques pour lire le fichier de données exporté. Pour plus d'informations, voir «Noeud Statistics», à la page 33.

Changement du nom ou filtrage des champs pour IBM SPSS Statistics

Avant d'exporter ou de déployer des données d'IBM SPSS Modeler vers des applications externes telles que IBM SPSS Statistics, vous pouvez être amené à renommer ou à ajuster des noms de champ. Les boîtes de dialogue Transformation Statistics, Sortie Statistics et Exporter Statistics contiennent un onglet Filtrer pour faciliter ce processus.

Vous trouverez dans une autre rubrique une brève description de l'onglet Filtrer. Pour plus d'informations, voir «Définition des options de filtrage», à la page 159. Cette rubrique fournit des astuces concernant la lecture des données dans IBM SPSS Statistics.

Pour ajuster les noms de fichiers afin de se conformer aux conventions de dénomination de IBM SPSS Statistics :

1. Dans l'onglet Filtrer, cliquez sur le bouton de la barre d'outils du menu des options de filtrage (le premier de la barre d'outils).
2. Sélectionnez Renommer pour IBM SPSS Statistics.
3. Dans la boîte de dialogue Renommer pour IBM SPSS Statistics, vous pouvez choisir de remplacer les caractères non valides des noms de fichier soit par un caractère **dièse (#)**, soit parc un **caractère de soulignement (_)**.

Renommer des ensembles à réponses multiples. Sélectionnez cette option si vous souhaitez ajuster le nom de plusieurs ensembles a réponses multiples, lesquels peuvent être importés dans IBM SPSS Modeler à l'aide d'un noeud source Statistics. Ils sont utilisés pour enregistrer des données qui peuvent comporter plus d'une valeur pour chaque cas, telles que les réponses à une enquête.

Chapitre 9. Super noeuds

Présentation des super noeuds

L'une des raisons pour laquelle l'interface de programmation graphique IBM SPSS Modeler est si facile à utiliser est que chaque noeud a une fonction clairement définie. Toutefois, un traitement complexe peut nécessiter une longue séquence de noeuds. Cela risque d'encombrer l'espace de travail de flux et de rendre difficile le suivi des diagrammes de flux. Vous pouvez éviter l'encombrement d'un flux long et complexe de deux manières :

- Vous pouvez partager une séquence de traitement en plusieurs flux qui s'auto-alimentent. Le premier flux, par exemple, crée un fichier de données que le deuxième utilise comme données d'entrée. Le deuxième crée un fichier que le troisième utilise également comme données d'entrée, et ainsi de suite. Vous pouvez gérer ces flux en les enregistrant dans un **projet**. Un projet permet d'organiser plusieurs flux, ainsi que leurs sorties. Cependant, un fichier de projet contient seulement une référence aux objets qu'il contient, et vous avez plusieurs fichiers de flux à gérer.
- Lorsque vous utilisez des processus de flux complexes, une alternative simple consiste à créer un **super noeud**.

Les super noeuds regroupent en un noeud unique plusieurs noeuds, en encapsulant les sections d'un flux de données. Le travail d'exploration de données est facilité grâce aux avantages suivants :

- Les flux sont plus nets et plus faciles à gérer.
- Les noeuds peuvent être regroupés en un super noeud propre à votre entreprise.
- Les super noeuds peuvent être exportés vers des bibliothèques pour être réutilisés dans plusieurs projets d'exploration de données.

Types de super noeuds

Les super noeuds sont représentés dans le flux de données par une icône en forme d'étoile. L'icône est partiellement hachurée pour indiquer le type de super noeud et le sens dans lequel le flux se déplace.

Il existe trois types de super noeuds :

- Super noeuds source
- Super noeuds d'exécution
- Super noeuds terminaux

Super noeuds source

A l'instar des noeuds source standard, les super noeuds source contiennent une source de données et peuvent être utilisés partout où le noeud source est utilisé. La partie gauche d'un super noeud source est hachurée pour indiquer qu'il n'est pas accessible à partir de la gauche et que les données doivent se déplacer en aval *à partir* d'un super noeud.

Les super noeuds source ne disposent que d'un seul point de connexion, à droite, indiquant que les données quittent le super noeud en direction du flux.

Super noeuds d'exécution

Les super noeuds d'exécution contiennent uniquement des noeuds d'exécution non hachurés pour indiquer que les données peuvent à la fois *entrer* et *sortir* de ce type de super noeud.

Les super noeuds d'exécution disposent de points de connexion à gauche et à droite, indiquant que les données pénètrent dans le super noeud et repartent dans le flux. Bien que les super noeuds puissent

contenir des fragments de flux supplémentaires, et même des flux supplémentaires, les deux points de connexion doivent circuler via un chemin d'accès unique reliant les points *A partir du flux* et *Vers le flux* entre eux.

Remarque : Les super noeuds d'exécution sont parfois appelés *super noeuds de manipulation*.

Super noeuds terminaux

Les super noeuds terminaux contiennent un ou plusieurs noeuds terminaux (Tracé, Table, etc.) et peuvent être utilisés de la même façon que des noeuds terminaux. La partie droite d'un super noeud terminal est hachurée pour indiquer qu'il n'est pas accessible à droite et que les données ne peuvent se déplacer que *vers* un super noeud terminal.

Les super noeuds terminaux ne disposent que d'un seul point de connexion, à gauche, indiquant que les données pénètrent dans le super noeud depuis le flux et finissent à l'intérieur du super noeud.

Les super noeuds terminaux peuvent également contenir des scripts utilisés pour indiquer l'ordre d'exécution de tous les noeuds terminaux au sein du super noeud. Pour plus d'informations, voir «Super noeuds et génération de scripts», à la page 416.

Création de super noeuds

La création d'un super noeud entraîne le "rétrécissement" du flux de données puisque plusieurs noeuds sont encapsulés dans un seul. Une fois que vous avez créé ou chargé un flux dans l'espace de travail, vous pouvez créer un super noeud de différentes manières.

Sélection multiple

La méthode la plus simple pour créer un super noeud consiste à sélectionner tous les noeuds que vous souhaitez encapsuler :

1. Utilisez la souris pour sélectionner plusieurs noeuds dans l'espace de travail du flux. Vous pouvez également utiliser la méthode Maj+clic pour sélectionner un flux ou la section d'un flux.

   Remarque : Vous devez sélectionner des noeuds provenant d'un flux continu ou bifurqué. Vous ne pouvez pas sélectionner des noeuds qui ne sont pas adjacents ou connectés.
2. Ensuite, encapsulez les noeuds sélectionnés en exécutant l'une des trois méthodes suivantes :
   - Cliquez sur l'icône du super noeud (en forme d'étoile) dans la barre d'outils.
   - Cliquez avec le bouton droit de la souris sur le super noeud et, dans le menu contextuel, sélectionnez :

     Créer un super noeud > **A partir de la sélection**
   - Dans le menu Super noeud, sélectionnez :

     Créer un super noeud > **A partir de la sélection**

Ces trois options encapsulent les noeuds dans un super noeud hachuré afin de refléter son type (source, exécution ou terminal) en fonction de son contenu.

Sélection unique

Vous pouvez également créer un super noeud en ne sélectionnant qu'un seul noeud et en utilisant les options du menu pour déterminer le début et la fin du super noeud, ou en encapsulant tous les noeuds se trouvant en aval du noeud sélectionné.

1. Cliquez sur le noeud qui détermine le départ de l'encapsulation.
2. Dans le menu Super noeud, sélectionnez :

   Créer un super noeud > **A partir d'ici**

Vous pouvez également créer des super noeuds de manière plus interactive, en sélectionnant le début et la fin de la section du flux pour encapsuler les noeuds :

1. Cliquez sur le premier ou le dernier noeud à ajouter au super noeud.
2. Dans le menu Super noeud, sélectionnez :

   Créer un super noeud > **Sélectionner...**
3. Vous pouvez également utiliser les options du menu contextuel en cliquant avec le bouton droit de la souris sur le noeud souhaité.
4. Le curseur prend la forme de l'icône de super noeud indiquant que vous devez sélectionner un autre endroit du flux. Déplacez-le vers le haut ou vers le bas, en direction de l'autre extrémité du fragment de super noeud et cliquez sur un noeud. Cette action remplace tous les noeuds situés entre les deux par l'icône en étoile du super noeud.

Remarque : Vous devez sélectionner des noeuds provenant d'un flux continu ou bifurqué. Vous ne pouvez pas sélectionner des noeuds qui ne sont pas adjacents ou connectés.

Imbrication des super noeuds

Les super noeuds peuvent s'imbriquer dans d'autres super noeuds. Les mêmes règles pour chaque type de super noeud (source, exécution et terminal) s'appliquent aux super noeuds imbriqués. Par exemple, un super noeud d'exécution avec imbrication doit comporter un flux de données continu à travers tous les super noeuds imbriqués afin de rester le super noeud d'exécution. Si l'un des super noeuds imbriqués est un super noeud terminal, les données ne circulent plus dans la hiérarchie.

Les super noeuds source et terminaux peuvent contenir d'autres types de super noeud imbriqué, mais les mêmes règles de base s'appliquent pour la création de super noeuds.

Verrouillage des super noeuds

Après avoir créé un super noeud, vous pouvez le verrouiller avec un mot de passe pour empêcher sa modification. Vous pouvez par exemple le faire si vous créez des flux, ou des parties de flux, en tant que modèles à valeur fixe destinés aux autres membres de votre organisation qui possèdent moins d'expérience dans la configuration d'enquêtes IBM SPSS Modeler.

Lorsqu'un super noeud est verrouillé, les utilisateurs peuvent toujours saisir des valeurs sur l'onglet Paramètres pour tous les paramètres qui ont été définis, et un super noeud verrouillé peut être exécuté sans saisir de mot de passe.

Remarque : Il est impossible d'effectuer un verrouillage ou un déverrouillage à l'aide de scripts.

Verrouillage et déverrouillage d'un super noeud

Remarque : Les mots de passe perdus ne sont pas récupérables.

Vous pouvez verrouiller ou déverrouiller un super noeud sur n'importe lequel de ces trois onglets.

1. Cliquez sur **Verrouiller un noeud**.
2. Entrez et confirmez le mot de passe.
3. Cliquez sur **OK**.

Un super noeud protégé par mot de passe est identifié sur l'espace de travail de flux par un petit symbole de cadenas en haut à gauche de l'icône du super noeud.

Déverrouillage d'un super noeud

1. Pour supprimer de manière permanente la protection par mot de passe, cliquez sur **Déverrouiller le noeud**. Vous êtes invité à saisir le mot de passe.

2. Saisissez le mot de passe et cliquez sur **OK**. Le super noeud n'est plus protégé par mot de passe et le symbole de cadenas n'apparaît plus en regard de l'icône dans le flux.

Pour un flux enregistré dans une version de SPSS Modeler entre 16 et 17.0 qui contient un super noeud verrouillé, lorsque vous ouvrez le flux dans un environnement différent tel que IBM SPSS Collaboration and Deployment Services ou sur un Mac où l'environnement d'exécution Java installé par SPSS Modeler est différent, vous devez d'abord l'ouvrir, le déverrouiller et réenregistrer à l'aide de la version 17.1 ou ultérieure sur l'ancien environnement où il a été sauvegardé pour la dernière fois.

Parfois une erreur de mot de passe incorrect sera affichée lors du déverrouillage d'un super noeud dans un flux antérieur à la version 18. Comme solution palliative, rouvrez et déverrouillez le noeud en utilisant la version exacte de IBM SPSS Modeler (ou une version plus récente) sur la même plateforme avec les mêmes paramètres locaux de système que ceux de sa dernière sauvegarde. Ouvrez-le dans la version 18 ou ultérieure. Verrouillez le noeud et sauvegardez de nouveau le flux.

Edition d'un super noeud verrouillé

Si vous essayez de définir des paramètres ou d'effectuer un zoom avant pour afficher un super noeud verrouillé, vous êtes invité à saisir le mot de passe.

Saisissez le mot de passe et cliquez sur **OK**.

Vous pouvez maintenant éditer les définitions de paramètre et effectuer un zoom avant ou arrière aussi souvent que nécessaire, jusqu'à ce que vous fermiez le flux dans lequel se trouve le super noeud.

Remarquez que cela ne supprime pas la protection par mot de passe, mais vous permet d'accéder au super noeud pour travailler dessus. Pour plus d'informations, voir «Verrouillage et déverrouillage d'un super noeud», à la page 411.

Edition de super noeuds

Après avoir créé un super noeud, vous pouvez l'analyser de plus près en effectuant un zoom avant ; si le super noeud est verrouillé, vous serez invité à saisir le mot de passe. Pour plus d'informations, voir «Edition d'un super noeud verrouillé».

Pour afficher le contenu d'un super noeud, vous pouvez utiliser l'icône de zoom avant située dans la barre d'outils IBM SPSS Modeler, ou la méthode suivante :

1. Cliquez avec le bouton droit de la souris sur un super noeud.
2. Dans le menu contextuel, choisissez **Zoom avant**.

Le contenu du super noeud sélectionné apparaît dans un environnement IBM SPSS Modeler légèrement différent ; des connecteurs affichent le flot de données circulant dans le flux ou le fragment de flux. Sur ce niveau de l'espace de travail, vous pouvez effectuer différentes tâches :

- Modifier le type—source du super noeud (source, exécution ou terminal).
- Créer des paramètres ou éditer les valeurs d'un paramètre. Les paramètres sont utilisés dans la génération de scripts et les expressions CLEM.
- Indiquer des options de mise en cache pour le super noeud et ses sous-noeuds.
- Créer ou modifier le script d'un super noeud (super noeuds terminaux uniquement).

Modification des types de super noeud

Il peut être utile, dans certains cas, de modifier le type d'un super noeud. Cette option n'est disponible que si le zoom avant est activé dans un super noeud et elle ne s'applique au super noeud qu'à ce niveau. Il existe trois types de super noeuds comme décrit le tableau ci-après.

Tableau 51. Types de super noeuds.

Type de super noeuds	Description
Super noeud source	Une connexion en sortie
Super noeud d'exécution	Deux connexions : une en entrée et une en sortie
Super noeud terminal	Une connexion en entrée

Pour modifier le type d'un super noeud

1. Assurez-vous que le zoom avant est activé dans le super noeud.
2. Dans le menu Super noeud, sélectionnez **Type de super noeud**, puis choisissez le type voulu.

Annotation et changement de nom des super noeuds

Vous pouvez renommer un super noeud apparaissant dans le flux et rédiger des annotations utilisées dans un projet ou un rapport. Pour accéder à ces propriétés :

- Cliquez avec le bouton droit de la souris sur un super noeud (zoom arrière activé) et sélectionnez **Renommer et annoter**.
- Vous pouvez également sélectionner **Renommer et annoter** dans le menu Super noeud. Cette option est disponible aussi bien en mode zoom avant qu'en mode zoom arrière.

Dans les deux cas, une boîte de dialogue apparaît avec l'onglet Annotations sélectionné. Utilisez ces options pour personnaliser le nom affiché dans l'espace de travail du flux et fournir des informations concernant les opérations du super noeud.

Utilisation des commentaires avec les super noeuds

Si vous créez un super noeud à partir d'un noeud ou nugget commenté, vous devez inclure le commentaire dans la sélection pour créer le super noeud et que le commentaire y apparaisse. Si vous n'avez pas inclus le commentaire dans la sélection, il restera dans le flux lors de la création du super noeud.

Lorsque vous développez un super noeud qui contenait des commentaires, ceux-ci sont restaurés à l'endroit où ils se trouvaient avant la création du super noeud.

Lorsque vous développez un super noeud qui contenait des objets commentés, mais que les commentaires n'étaient pas inclus dans le super noeud, les objets sont restaurés à l'endroit où ils se trouvaient mais les commentaires ne sont pas attachés à nouveau.

paramètres du super noeud

Dans IBM SPSS Modeler, vous pouvez indiquer des variables définies par l'utilisateur, telles que `Minvalue`, dont les valeurs peuvent être spécifiées lorsqu'elles sont employées dans un script ou dans des expressions CLEM. Ces variables sont appelées des **paramètres**. Vous pouvez définir des paramètres pour les flux, les sessions et les super noeuds. Tous les paramètres définis pour un super noeud sont disponibles lors de la création d'expressions CLEM dans ce super noeud ou dans n'importe quel noeud imbriqué. Les paramètres définis pour les super noeuds imbriqués ne sont pas disponibles pour leur super noeud parent.

Vous devez suivre deux étapes pour créer et définir les paramètres des super noeuds :

1. Définissez les paramètres du super noeud.
2. Indiquez ensuite la valeur de chaque paramètre du super noeud.

Vous pouvez ensuite utiliser ces paramètres dans des expressions CLEM pour n'importe quel noeud encapsulé.

Définitions des paramètres de super noeud

Les paramètres d'un super noeud peuvent être définis en mode zoom avant comme en mode zoom arrière. Les paramètres définis s'appliquent à tous les noeuds encapsulés. Pour définir les paramètres d'un super noeud, vous devez d'abord accéder à l'onglet Paramètres de la boîte de dialogue du super noeud. Pour ouvrir la boîte de dialogue, utilisez l'une des méthodes suivantes :

- Double-cliquez sur un super noeud du flux.
- Dans le menu Super noeud, sélectionnez **Définir les paramètres**.
- Si le zoom avant est activé dans le super noeud, vous pouvez également sélectionner **Définir les paramètres** dans le menu contextuel.

Dans la boîte de dialogue, l'onglet Paramètres affiche tous les paramètres définis précédemment.

Pour définir un nouveau paramètre

Cliquez sur le bouton **Définir les paramètres** pour ouvrir la boîte de dialogue.

Nom. Les noms des paramètres sont répertoriés ici. Vous pouvez créer un paramètre en entrant un nom dans ce champ. Par exemple, pour créer un paramètre relatif à la température minimale, vous pouvez saisir `Valeur min.`. N'insérez pas le préfixe `$P-`, qui indique un paramètre dans les expressions CLEM. Ce nom est également utilisé pour l'affichage dans le Générateur de formules de CLEM.

Nom complet. Répertorie le nom descriptif de chaque paramètre créé.

Stockage. Sélectionnez le type de stockage dans la liste. Indique le mode de stockage des valeurs de données dans le paramètre. Par exemple, si vous utilisez des valeurs commençant par des zéros à conserver (comme 008), vous devez sélectionner **Chaîne** comme type de stockage. Sinon, les zéros seront supprimés de la valeur. Les types de stockage disponibles sont les suivants : chaîne, entier, réel, temps, date et horodatage. Pour les paramètres de date, les valeurs doivent être définies à l'aide de la notation standard ISO telle qu'elle est présentée dans le paragraphe suivant.

Valeur. Indique la valeur actuelle du paramètre sélectionné. Modifiez ce paramètre selon les besoins. Pour les paramètres de date, les valeurs doivent être définies à l'aide de la notation standard ISO (soit, `YYYY-MM-DD`). Toute date définie dans un autre format est refusée.

Type (facultatif). Si vous prévoyez de déployer le flux vers une application externe, sélectionnez un niveau de mesure dans la liste. Sinon, il est conseillé de laisser la colonne *Type* en l'état. Si vous souhaitez spécifier des contraintes de valeur pour le paramètre, telles que des limites supérieures et inférieures d'un intervalle numérique, sélectionnez **Spécifier** dans la liste.

Vous ne pouvez définir les options de nom long, de stockage et de type pour les paramètres que dans l'interface utilisateur. Il est impossible de définir ces options à l'aide de scripts.

Cliquez sur les flèches à droite pour déplacer le paramètre sélectionné vers le haut ou le bas de la liste des paramètres disponibles. Utilisez le bouton de suppression (indiqué par un *X*) pour supprimer le paramètre sélectionné.

Définition des valeurs des paramètres de super noeud

Une fois que vous avez défini les paramètres d'un super noeud, vous pouvez spécifier les valeurs à l'aide des paramètres dans une expression CLEM ou un script.

Pour définir les paramètres d'un super noeud

1. Double-cliquez sur l'icône Super noeud pour ouvrir la boîte de dialogue Super noeud.
2. Vous pouvez également sélectionner **Définir les paramètres** dans le menu Super noeud.

3. Cliquez sur l'onglet **Paramètres**. *Remarque* : les champs de cette boîte de dialogue sont ceux qui ont été définis à l'aide du bouton **Définir les paramètres** de cet onglet.
4. Entrez une valeur dans la zone de texte pour chaque paramètre créé. Par exemple, vous pouvez définir la valeur *Valeur min.* sur un seuil d'intérêt particulier. Ce paramètre peut ensuite être utilisé dans de nombreuses opérations, telles que la sélection d'enregistrements au-delà ou en deçà de ce seuil pour une analyse plus approfondie.

Utilisation des paramètres de super noeud pour accéder aux propriétés du noeud

Vous pouvez également utiliser les paramètres de super noeud pour définir les propriétés de noeud (également appelées **paramètres d'emplacement**) pour les noeuds encapsulés. Par exemple, imaginons que vous souhaitiez spécifier qu'un super noeud forme un noeud R. neurones (Réseau de neurones) encapsulé pendant une durée déterminée en utilisant un échantillon aléatoire des données disponibles. Grâce aux paramètres, vous pouvez spécifier les valeurs concernant la durée et l'échantillon de pourcentage.

Supposons que le super noeud pris en exemple contienne un noeud Echantillon appelé *Echantillonner*; et un noeud Réseau de neurones appelé *Apprendre*. Vous pouvez utiliser les boîtes de dialogue du noeud pour définir le paramètre **Echantillonner** du noeud Echantillon sur **% aléatoire** et le paramètre **Critère d'arrêt** du noeud R. neurones (Réseau de neurones) sur **Temps**. Une fois ces options spécifiées, vous pouvez accéder aux propriétés du noeud avec les paramètres et indiquer des valeurs particulières pour le super noeud. Dans la boîte de dialogue du super noeud, cliquez sur **Définir les paramètres** et créez les paramètres indiqués dans le tableau ci-après.

Tableau 52. Paramètres à créer

Paramètre	Valeur	Nom complet
Durée.apprentissage	5	Durée d'apprentissage (minutes)
Echantillonnage.aléatoire	10	Pourcentage d'échantillonnage aléatoire

Remarque : les noms de paramètres tels que *Echantillonnage.aléatoire* utilisent une syntaxe correcte pour faire référence aux propriétés de noeud où *Echantillonnage* représente le nom du noeud et *aléatoire* une propriété de noeud.

Une fois que vous avez défini ces paramètres, vous pouvez facilement modifier les valeurs des propriétés des noeuds Echantillon et R. neurones (Réseau de neurones) sans rouvrir chaque boîte de dialogue. Au contraire, il vous suffit de sélectionner **Définir les paramètres** dans le menu du super noeud pour accéder à l'onglet Paramètres de la boîte de dialogue du super noeud, où vous pouvez choisir de nouvelles valeurs pour **% aléatoire** et **Temps**. Cette opération est particulièrement utile lorsque vous explorez les données pendant de nombreuses itérations de génération de modèle.

Super noeuds et mise en cache

A l'exception des noeuds terminaux, tous les noeuds peuvent être mis en cache depuis l'intérieur d'un super noeud. Pour effectuer une mise en cache, cliquez avec le bouton droit de la souris sur un noeud et sélectionnez une option dans le menu contextuel Cache. Cette option de menu est disponible depuis l'extérieur d'un super noeud, ainsi que pour les noeuds encapsulés au sein d'un super noeud.

Il existe plusieurs directives pour les caches de super noeud :

- Si la mise en cache est activée pour un noeud encapsulé, elle l'est également pour le super noeud.
- En désactivant le cache sur un super noeud, vous désactivez également le cache de *tous* les noeuds encapsulés.
- En activant le cache sur un super noeud, vous activez également le cache du dernier sous-noeud pouvant être mis en cache. En d'autres termes, si le dernier sous-noeud est un noeud Sélectionner, le cache sera activé pour ce noeud. Si le dernier sous-noeud est un noeud terminal (n'autorisant pas la mise en cache), le noeud suivant en amont prenant en charge la mise en cache est activé.

- Une fois que vous avez défini les caches pour les sous-noeuds d'un super noeud, toutes les activités en amont en provenance du noeud mis en cache, telles que l'ajout ou l'édition de noeuds, entraînent le vidage des caches.

Super noeuds et génération de scripts

Vous pouvez utiliser le langage de script SPSS Modeler pour écrire des programmes simples servant à manipuler et à exécuter le contenu d'un super noeud. Vous pouvez, par exemple, indiquer l'ordre d'exécution d'un flux complexe. Par exemple, si un super noeud contient un noeud Valeurs globales qui doit être exécuté avant un noeud Nuage, vous pouvez créer un script qui exécute d'abord le noeud Valeurs globales. Les valeurs calculées par ce noeud, telles que la moyenne ou l'écart-type, peuvent être utilisées lorsque le noeud Tracé est exécuté.

L'onglet Script de la boîte de dialogue du super noeud n'est disponible que pour les super noeuds terminaux.

Pour ouvrir la boîte de dialogue de script pour un super noeud terminal :

- Cliquez avec le bouton droit de la souris sur l'espace de travail du super noeud et sélectionnez **Script Super noeud**.
- Sinon, que ce soit en mode zoom avant ou zoom arrière, vous pouvez sélectionner **Script Super noeud** dans le menu du super noeud.

Remarque : Les scripts de super noeud sont exécutés uniquement avec le flux et le super noeud si vous avez sélectionné **Exécuter ce script** dans la boîte de dialogue.

Les options propres à la génération de scripts et son utilisation dans SPSS Modeler sont abordées dans le *Guide de génération de scripts et d'automatisation*, disponible sous forme de fichier PDF avec le téléchargement de votre produit.

Enregistrement et chargement des super noeuds

Les super noeuds peuvent être enregistrés et réutilisés dans d'autres flux. L'extension utilisée pour l'enregistrement et le chargement des super noeuds est `.slb`.

Pour enregistrer un super noeud

1. Effectuez un zoom avant dans le super noeud.
2. Dans le menu Super noeud, sélectionnez **Enregistrer le super noeud**.
3. Entrez un nom de fichier et un répertoire dans la boîte de dialogue.
4. Indiquez si vous souhaitez ajouter le super noeud enregistré au projet en cours.
5. Cliquez sur **Enregistrer**.

Pour charger un super noeud

1. Dans le menu Insertion de la fenêtre IBM SPSS Modeler, sélectionnez **Super noeud**.
2. Sélectionnez un fichier de super noeud (`.slb`) dans le répertoire ouvert ou accédez à un autre répertoire.
3. Cliquez sur **Charger**.

Remarque : Les valeurs de tous les paramètres des super noeuds importés sont celles par défaut. Pour modifier les paramètres, double-cliquez sur un super noeud dans l'espace de travail.

Remarques

Le présent document a été développé pour des produits et des services proposés aux Etats-Unis et peut être mis à disposition par IBM dans d'autres langues. Toutefois, il peut être nécessaire de posséder une copie du produit ou de la version du produit dans cette langue pour pouvoir y accéder.

Le présent document peut contenir des informations ou des références concernant certains produits, logiciels ou services IBM non annoncés dans ce pays. Pour plus de détails, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial IBM. Toute référence à un produit, programme ou service IBM n'implique pas que seul ce produit, programme ou service IBM puisse être utilisé. Tout produit, programme ou service fonctionnellement équivalent peut être utilisé s'il n'enfreint aucun droit de propriété intellectuelle d'IBM. Cependant l'utilisateur doit évaluer et vérifier l'utilisation d'un produit, programme ou service non IBM.

IBM peut détenir des brevets ou des demandes de brevet couvrant les produits mentionnés dans le présent document. L'octroi de ce document n'équivaut aucunement à celui d'une licence pour ces brevets. Vous pouvez envoyer par écrit des questions concernant la licence à :

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Pour toute demande au sujet des licences concernant les jeux de caractères codés sur deux octets (DBCS), contactez le service Propriété intellectuelle IBM de votre pays ou adressez vos questions par écrit à :

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

LE PRESENT DOCUMENT EST LIVRE "EN L'ETAT" SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DECLINE NOTAMMENT TOUTE RESPONSABILITE RELATIVE A CES INFORMATIONS EN CAS DE CONTREFACON AINSI QU'EN CAS DE DEFAUT D'APTITUDE A L'EXECUTION D'UN TRAVAIL DONNE. Certaines juridictions n'autorisent pas l'exclusion des garanties implicites, auquel cas l'exclusion ci-dessus ne vous sera pas applicable.

Ces informations peuvent contenir des erreurs techniques ou des erreurs typographiques. Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. IBM peut, à tout moment et sans préavis, modifier les produits et logiciels décrits dans ce document.

Les références à des sites Web non IBM sont fournies à titre d'information uniquement et n'impliquent en aucun cas une adhésion aux données qu'ils contiennent. Les éléments figurant sur ces sites Web ne font pas partie des éléments du présent produit IBM et l'utilisation de ces sites relève de votre seule responsabilité.

IBM pourra utiliser ou diffuser, de toute manière qu'elle jugera appropriée et sans aucune obligation de sa part, tout ou partie des informations qui lui seront fournies.

Les licenciés souhaitant obtenir des informations permettant : (i) l'échange des données entre des logiciels créés de façon indépendante et d'autres logiciels (dont celui-ci), et (ii) l'utilisation mutuelle des données ainsi échangées, doivent adresser leur demande à :

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Ces informations peuvent être disponibles, soumises à des conditions générales, et dans certains cas payantes.

Le programme sous licence décrit dans le présent document et tous les éléments sous licence disponibles s'y rapportant sont fournis par IBM conformément aux dispositions du Livret Contractuel IBM, des Conditions internationales d'utilisation des Logiciels IBM ou de tout autre contrat équivalent.

Les données de performances et les exemples de clients ne sont présentés qu'à des fins d'illustration. Les performances réelles peuvent varier en fonction des configurations et des conditions d'exploitation spécifiques.

Les informations concernant des produits non IBM ont été obtenues auprès des fournisseurs de ces produits, par l'intermédiaire d'annonces publiques ou via d'autres sources disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances ni leur compatibilité. Elle ne peut recevoir aucune réclamation concernant des produits non IBM. Toute question concernant les performances de produits non IBM doit être adressée aux fournisseurs de ces produits.

Toute instruction relative aux intentions d'IBM pour ses opérations à venir est susceptible d'être modifiée ou annulée sans préavis, et doit être considérée uniquement comme un objectif.

Ces informations contiennent des exemples de données et de rapports utilisés au cours d'opérations quotidiennes standard. Pour les illustrer le mieux possible, ces exemples contiennent des noms d'individus, d'entreprises, de marques et de produits. Toute ressemblance avec des noms de personnes, de sociétés ou des données réelles serait purement fortuite.

Marques

IBM, le logo IBM et ibm.com sont des marques d'International Business Machines dans de nombreux pays. Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés. La liste actualisée de toutes les marques d'IBM est disponible sur la page Web "Copyright and trademark information" à www.ibm.com/legal/copytrade.shtml.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques déposées ou des marques commerciales de Adobe Systems Incorporated aux Etats-Unis et/ou dans d'autres pays.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, et Pentium sont des marques commerciales ou des marques déposées de Intel Corporation ou de ses filiales aux Etats-Unis et dans d'autres pays.

Linux est une marque déposée de Linus Torvalds aux Etats-Unis et/ou dans d'autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques commerciales de Microsoft Corporation aux Etats-Unis et/ou dans d'autres pays.

UNIX est une marque déposée de The Open Group aux Etats-Unis et dans d'autres pays.

Les marques commerciales Java et basées sur Java ainsi que les logos sont des marques commerciales ou déposées de Oracle et/ou de ses filiales.

Dispositions applicables à la documentation du produit

Les droits d'utilisation relatifs à ces publications sont soumis aux dispositions suivantes.

Applicabilité

Les présentes dispositions viennent s'ajouter à toute autre condition d'utilisation applicable au site Web IBM.

Usage personnel

Vous pouvez reproduire ces publications pour votre usage personnel, non commercial, sous réserve que toutes les mentions de propriété soient conservées. Vous ne pouvez distribuer ou publier tout ou partie de ces publications ou en faire des oeuvres dérivées sans le consentement exprès d'IBM.

Usage commercial

Vous pouvez reproduire, distribuer et afficher ces publications uniquement au sein de votre entreprise, sous réserve que toutes les mentions de propriété soient conservées. Vous ne pouvez reproduire, distribuer, afficher ou publier tout ou partie de ces publications en dehors de votre entreprise, ou en faire des oeuvres dérivées, sans le consentement exprès d'IBM.

Droits

Excepté les droits d'utilisation expressément accordés dans ce document, aucun autre droit, licence ou autorisation, implicite ou explicite, n'est accordé pour ces publications ou autres informations, données, logiciels ou droits de propriété intellectuelle contenus dans ces publications.

IBM se réserve le droit de retirer les autorisations accordées ici si, à sa discrétion, l'utilisation des publications s'avère préjudiciable à ses intérêts ou que, selon son appréciation, les instructions susmentionnées n'ont pas été respectées.

Vous ne pouvez télécharger, exporter ou réexporter ces informations qu'en total accord avec toutes les lois et règlements applicables dans votre pays, y compris les lois et règlements américains relatifs à l'exportation.

IBM N'OCTROIE AUCUNE GARANTIE SUR LE CONTENU DE CES PUBLICATIONS. LES PUBLICATIONS SONT LIVREES EN L'ETAT SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DECLINE NOTAMMENT TOUTE RESPONSABILITE RELATIVE A CES PUBLICATIONS EN CAS DE CONTREFAÇON AINSI QU'EN CAS DE DEFAUT D'APTITUDE A L'EXECUTION D'UN TRAVAIL DONNE.

Glossaire

V

Covariance **:** Mesure d'association non normalisée entre deux variables, égale à la déviation des produits en croix divisée par N-1.

K

Kurtosis **:** Mesure de l'importance des valeurs extrêmes. Dans le cas d'une distribution normale, la valeur de la statistique d'aplatissement est égale à zéro. Un kurtosis positif indique qu'on observe dans les données plus de valeurs extrêmes que dans une distribution normale. Une valeur négative indique que les données comportent moins de valeurs extrêmes qu'une distribution normale.

M

maximum **:** La plus grande valeur d'une variable numérique.

Moyenne **:** Mesure de la tendance centrale. Moyenne arithmétique : somme divisée par le nombre d'observations.

Médiane **:** Valeur au-dessus et au-dessous de laquelle se trouvent la moitié des observations, le 50e percentile. Si le nombre de cellules est pair, la médiane correspond à la moyenne des deux cellules du milieu lorsqu'elles sont triées dans l'ordre croissant ou décroissant. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées.

minimum **:** La plus petite valeur d'une variable numérique.

Mode **:** Valeur qui revient le plus souvent. Si plusieurs valeurs partagent la plus grande fréquence d'occurrence, chacune d'elles constitue un mode.

R

Plage **:** Différence entre la valeur maximale et la valeur minimale d'une variable numérique (maximum – minimum).

S

Asymétrie **:** Mesure de l'asymétrie d'une distribution. La distribution normale est symétrique et présente une valeur de décalage de zéro. Une distribution avec un important décalage positif présente une longue queue vers la droite. Une distribution caractérisée par une importante asymétrie négative présente une extrémité gauche plus allongée. Pour simplifier, une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.

écart type **:** Mesure de la dispersion des valeurs autour de la moyenne, égale à la racine carrée de la variance. L'écart type est mesuré dans les mêmes unités que la variable d'origine.

écart type **:** Mesure de dispersion autour de la moyenne. Dans le cas d'une distribution normale, 68 % des observations se situent à l'intérieur d'un écart-type de la moyenne et 95 % se situent à l'intérieur de deux écarts-types. Par exemple, si l'âge moyen est 45 ans avec un écart type égal à 10, une distribution normale verra 95 % des observations se situer entre 25 et 65.

Erreur standard **:** Mesure du degré de variation des valeurs de statistiques de test d'un échantillon à l'autre. Il s'agit de l'écart-type de la distribution de l'échantillon pour des statistiques. Par exemple, l'erreur standard de la moyenne est l'écart type des moyennes d'échantillon.

erreur standard de Kurtosis **:** Rapport de kurtosis avec son erreur standard pouvant être utilisé comme test de normalité (autrement dit, vous pouvez conclure à une anormalité si ce rapport est inférieur à -2 ou supérieur à +2).

Une valeur de kurtosis positive élevée indique que les extrémités de la distribution sont plus longues que celles d'une distribution normale ; une valeur de kurtosis négative présente des extrémités plus courtes (semblables à celles d'une distribution uniforme sous forme de boîtes).

Erreur standard de la moyenne **:** Mesure de la variation de la valeur de la moyenne d'un échantillon à l'autre issus de la même distribution. Cette mesure permet de comparer approximativement la moyenne observée avec une valeur hypothétique (autrement dit, vous pouvez conclure que ces deux valeurs sont différentes si le rapport de la différence avec l'erreur standard est inférieur à -2 ou supérieur à +2).

erreur standard d'asymétrie **:** Rapport de l'asymétrie avec son erreur standard pouvant être utilisé comme test de normalité (autrement dit, vous pouvez conclure à une anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur positive élevée indique une extrémité allongée vers la droite ; une valeur négative extrême produit une extrémité allongée vers la gauche.

somme **:** Somme ou total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.

U

unique **:** Evalue tous les effets simultanément, en ajustant chaque effet à tous les autres effets d'un type donné.

V

valide **:** Observations valides qui ne comportent pas la valeur système manquante ni une valeur manquante définie par l'utilisateur.

Variance **:** Mesure de dispersion autour de la moyenne, égal à la somme des écarts au carré par rapport à la moyenne divisé par le nombre d'observations moins un. La variance se mesure en unités, qui sont égales au carré des unités de la variable.

Index

E

F

G

H

I

J

L

M

N

O

P

Q

R

S

T

U

V

Z

Imprimé en France