IBM SPSS Categories 20

Jacqueline J. Meulman

Willem J. Heiser

Remarque : Avant d'utiliser ces informations et le produit qu'elles concernent, lisez les informations générales sous Remarques sur p. 316.

Cette version s'applique à IBM® SPSS® Statistics 20 et à toutes les publications et modifications ultérieures jusqu'à mention contraire dans les nouvelles versions.

Les captures d'écran des produits Adobe sont reproduites avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran des produits Microsoft sont reproduites avec l'autorisation de Microsoft Corporation.

Préface

IBM® SPSS® Statistics est un système complet d'analyse de données. Le module complémentaire facultatif Categories fournit les techniques d'analyse supplémentaires décrites dans ce manuel. Le module complémentaire Categories doit être utilisé avec le système central SPSS Statistics auquel il est entièrement intégré.

A propos de IBM Business Analytics

Le logiciel IBM Business Analytics offre des informations complètes, cohérentes et précises permettant aux preneurs de décision d'améliorer leurs performances professionnelles. Un portefeuille complet de solutions de business intelligence, d'analyses prédictives, de performance financière et de gestion de la stratégie, et d'applications analytiques permet une connaissance claire et immédiate et offre des possibilités d'actions sur les performances actuelles et la capacité de prédire les résultats futurs. En combinant des solutions du secteur, des pratiques prouvées et des services professionnels, les entreprises de toute taille peuvent générer la plus grande productivité, automatiser les décisions en toute confiance et apporter de meilleurs résultats.

Dans le cadre de ce portefeuille, le logiciel IBM SPSS Predictive Analytics aide les entreprises à prédire des événements futurs et à agir de manière proactive en fonction de ces prédictions pour apporter de meilleurs résultats. Des clients dans les domaines commerciaux, gouvernementaux et académiques se servent de la technologie IBM SPSS comme d'un avantage concurrentiel pour attirer ou retenir des clients, tout en réduisant les risques liés à l'incertitude et à la fraude. En intégrant le logiciel IBM SPSS à leurs opérations quotidiennes, les entreprises peuvent effectuer des prévisions, et sont capables de diriger et d'automatiser leurs décisions afin d'atteindre leurs objectifs commerciaux et d'obtenir des avantages concurrentiels mesurables. Pour plus d'informations ou pour contacter un représentant, visitez le site *http://www.ibm.com/spss*.

Support technique

Un support 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, visitez le site IBM Corp. à l'adresse *http://www.ibm.com/support*. Votre nom, celui de votre société, ainsi que votre contrat d'assistance vous seront demandés.

Support technique pour les étudiants

Si vous êtes un étudiant qui utilise la version pour étudiant, personnel de l'éducation ou diplômé d'un produit logiciel IBM SPSS, veuillez consulter les pages Solutions pour l'éducation (*http://www.ibm.com/spss/rd/students/*) consacrées aux étudiants. Si vous êtes un étudiant utilisant une copie du logiciel IBM SPSS fournie par votre université, veuillez contacter le coordinateur des produits IBM SPSS de votre université.

Service clients

Si vous avez des questions concernant votre livraison ou votre compte, contactez votre bureau local. Veuillez préparer et conserver votre numéro de série à portée de main pour l'identification.

Séminaires de formation

IBM Corp. propose des séminaires de formation, publics et sur site. Tous les séminaires font appel à des ateliers de travaux pratiques. Ces séminaires seront proposés régulièrement dans les grandes villes. Pour plus d'informations sur ces séminaires, accédez au site *http://www.ibm.com/software/analytics/spss/training*.

Documents supplémentaires

Les ouvrages *SPSS Statistics : Guide to Data Analysis*, *SPSS Statistics : Statistical Procedures Companion*, et *SPSS Statistics : Advanced Statistical Procedures Companion*, écrits par Marija Norušis et publiés par Prentice Hall, sont suggérés comme documentation supplémentaire. Ces publications présentent les procédures statistiques des modules SPSS Statistics Base, Advanced Statistics et Regression. Que vous soyez novice dans les analyses de données ou prêt à utiliser des applications plus avancées, ces ouvrages vous aideront à exploiter au mieux les fonctionnalités offertes par IBM® SPSS® Statistics. Pour obtenir des informations supplémentaires y compris le contenu des publications et des extraits de chapitres, visitez le site web de l'auteur : *http://www.norusis.com*

Remerciements

Les procédures de codage optimal et leur mise en oeuvre dans IBM® SPSS® Statistics ont été développées par le groupe DTSS (Data Theory Scaling System Group), composé de membres des départements d'enseignement et de psychologie de la Faculté des sciences sociales et du comportement de l'Université de Leyde (Pays-Bas).

Willem Heiser, Jacqueline Meulman, Gerda van den Berg et Patrick Groenen ont apporté leur contribution à la création des procédures initiales, en 1990. Jacqueline Meulman et Peter Neufeglise ont participé au développement des procédures de régression nominale, d'analyse des correspondances, d'analyse en composantes principales qualitatives et de positionnement multidimensionnel. En outre, Anita van der Kooij a spécialement contribué aux procédures CATREG, CORRESPONDENCE et CATPCA. Willem Heiser, Jacques Commandeur, Frank Busing, Gerda van den Berg et Patrick Groenen ont participé au développement de la procédure PROXSCAL. Frank Busing, Willem Heiser, Patrick Groenen et Peter Neufeglise ont participé au développement de la procédure PREFSCAL.

Contenu

Partie I: Guide de l'utilisateur

Partie II: Exemples

15 Dépliage multidimensionnel 269

Annexes

Partie I:
Guide de l'utilisateur

Chapitre
1

Introduction aux procédures de codage optimal pour les données qualitatives

Les procédures de modalité font appel au codage optimal pour analyser les données dont l'analyse, par le biais des procédures statistiques standard, est complexe, voire impossible. Ce chapitre décrit le fonctionnement de chacune des procédures, les circonstances dans lesquelles leur utilisation est la plus favorable, les relations entre les différentes procédures et les relations de ces dernières avec les procédures statistiques standard.

Remarque : Ces procédures et leur mise en oeuvre IBM® SPSS® Statistics ont été développées par le groupe DTSS (Data Theory Scaling System), composé de membres des départements d'enseignement et de psychologie de la Faculté des sciences sociales et du comportement de l'Université de Leyde (Pays-Bas).

Définition du codage optimal

Le codage optimal consiste à associer des quantifications numériques aux modalités de chaque variable. Ainsi, les procédures standard peuvent être utilisées pour obtenir une solution portant sur les variables quantifiées.

Les valeurs d'échelle optimale sont attribuées aux modalités de chaque variable selon le critère d'optimisation de la procédure utilisée. A la différence des étiquettes d'origine des variables nominales ou ordinales de l'analyse, ces valeurs d'échelle ont des propriétés métriques.

Dans la plupart des procédures de modalité, la quantification optimale de chaque variable codée est obtenue via une méthode itérative appelée **moindres carrés alternés**. Dans cette méthode, les quantifications actuelles, une fois utilisées pour chercher une solution, sont mises à jour à l'aide de cette solution. Les quantifications mises à jour permettent alors de chercher une autre solution, utilisée pour mettre à jour ces quantifications, jusqu'à ce que le critère signalant la fin du processus soit satisfait.

Raisons de l'utilisation du codage optimal

En général, les données qualitatives sont utilisées dans le cadre d'une recherche commerciale, d'un sondage ou d'une recherche liée aux sciences sociales et du comportement. En fait, nombreux sont les chercheurs qui travaillent exclusivement avec ce type de données.

Alors que les adaptations de la plupart des modèles standard sont disponibles notamment pour l'analyse des données qualitatives, leur utilisation ne convient pas aux ensembles de données contenant :

- Un nombre d'observations insuffisant

- Un nombre de variables excessif
- Un nombre de valeurs par variable excessif

En quantifiant les modalités, les méthodes de codage optimal évitent tout problème dans ces cas-là. En outre, elles s'avèrent utiles même si des méthodes spécifiques sont appropriées.

Habituellement, l'interprétation de résultats de codage optimal repose sur des graphiques, plutôt que sur des estimations de paramètres. Les méthodes de codage optimal fournissent d'excellentes analyses exploratoires qui complètent bien les autres modèles IBM® SPSS® Statistics. Si vous affinez votre recherche, la visualisation des données codées de façon optimale peut servir de base à une analyse centrée sur l'interprétation de paramètres de modèle.

Niveau de codage optimal et niveau de mesure

Ce concept peut fortement prêter à confusion lorsque vous utilisez les procédures de modalité pour la première fois. Si vous spécifiez le niveau, il ne s'agit pas du niveau de *mesure* des variables, mais de leur niveau de *codage*. L'idée est la suivante : les variables à quantifier peuvent avoir des relations non linéaires, quelle que soit la manière dont elles sont mesurées.

On distingue trois niveaux de mesure de base pour les modalités :

- Le niveau **nominal** signifie que les valeurs d'une variable représentent des modalités non classées. Voici quelques exemples de variables pouvant être considérées comme nominales : les modalités de région, de code postal, d'appartenance religieuse et à choix multiples.
- Le niveau **ordinal** signifie que les valeurs d'une variable représentent des modalités classées. En voici quelques exemples : les échelles d'attitude représentant le degré de satisfaction ou de confiance, et les échelles d'évaluation des préférences.
- Le niveau **numérique** signifie que les valeurs d'une variable représentent des modalités classées avec une mesure significative, de sorte que les comparaisons de distance entre les modalités soient adéquates. L'âge en années et le revenu en milliers de dollars constituent des exemples.

Par exemple, supposons que les variables *région*, *travail* et *âge* sont codées comme l'indique le tableau suivant.

Table 1-1
Système de codage de la région, du travail et de l'âge

Code de région	Valeur de région.	Code de tâche	Valeur de tâche	Age
1	Nord	1	stagiaire	20
2	Sud	2	représentant	22
3	Est	3	Gestionnaire	25
4	Ouest			27

Les valeurs mentionnées représentent les modalités de chaque variable. *Région* est une variable nominale. On distingue quatre modalités de *région* sans ordre intrinsèque. Les valeurs 1 à 4 représentent simplement ces quatre modalités. Le système de codage est totalement arbitraire. En revanche, la variable *travail* peut être considérée comme une variable ordinale. Les modalités d'origine représentent une progression du statut de stagiaire à celui de responsable. Plus les codes

sont élevés, plus ils font référence à fonction élevée dans la hiérarchie de l'entreprise. Toutefois, seules les informations relatives à l'ordre sont connues, mais aucun élément d'information ne peut être fourni concernant la distance entre les modalités adjacentes. En revanche, la variable *âge* peut être considérée comme une variable numérique. Dans le cas de la variable *âge*, les distances entre les valeurs sont intrinsèquement explicites. La distance entre 20 et 22 est identique à celle entre 25 et 27, alors que la distance entre 22 et 25 est supérieure à ces deux distances.

Sélection du niveau de codage optimal

Il est important de comprendre qu'aucune propriété intrinsèque de variable ne prédéfinit automatiquement le niveau de codage optimal que vous devez indiquer. Vous pouvez explorer les données de manière cohérente et simplifiant l'interprétation. En analysant une variable numérique au niveau ordinal, par exemple, une transformation non linéaire autorise une solution dans un nombre inférieur de dimensions.

Les deux exemples suivants illustrent le fait que le niveau de mesure « évident » n'est peut-être pas le niveau de codage optimal. Supposez qu'une variable répartit les objets dans les différents groupes d'âge. Bien que la variable âge puisse être codée en tant que variable numérique, il s'avère parfois que, pour les jeunes de moins de 25 ans, la sécurité a un rapport positif avec l'âge alors que, pour les personnes de plus de 60 ans, ce rapport est négatif. Dans ce cas, mieux vaut peut-être considérer âge comme une variable nominale.

Autre exemple : une variable triant les personnes par préférence politique semble avant tout être nominale. Toutefois, si vous triez les partis politiques de gauche à droite, il se peut que leur quantification doive respecter cet ordre. Dans ce cas, vous devrez utiliser un niveau d'analyse ordinal.

Même s'il n'existe aucune propriété prédéfinie de variable la transformant exclusivement en tel ou tel niveau, l'utilisateur débutant peut s'aider des règles générales suivantes. Dans la quantification nominale simple, vous ne connaissez pas en général l'ordre des modalités, mais l'analyse doit en imposer un. Si l'ordre des modalités est connu, vous devez faire appel à la quantification ordinale. Si les modalités ne peuvent pas être classées, vous pouvez utiliser la quantification nominale multiple.

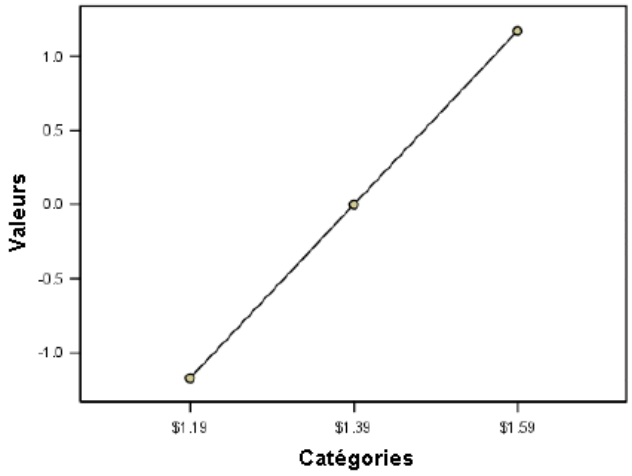
Diagrammes de transformation

Les différents niveaux auxquels chaque variable peut être codée imposent plusieurs restrictions dans les quantifications. Les diagrammes de transformation illustrent la relation entre les quantifications et les modalités d'origine résultant du niveau de codage optimal sélectionné. Par exemple, un diagramme de transformation linéaire est obtenu lorsqu'une variable est considérée comme numérique. Les variables considérées comme ordinales entraînent la création d'un diagramme de transformation non décroissant. Les diagrammes de transformation de variables considérées comme nominales, en forme de U (ou l'inverse), affichent une relation quadratique. Ces variables peuvent également créer des diagrammes de transformation sans tendance apparente en changeant complètement l'ordre des modalités. La figure suivante représente un exemple de diagramme de transformation.

Les diagrammes de transformation conviennent particulièrement à la définition du mode de fonctionnement du niveau de codage optimal sélectionné. Si plusieurs modalités reçoivent des quantifications similaires, la fusion de ces modalités en une seule modalité peut être garantie.

Si une variable considérée comme nominale reçoit des quantifications affichant une tendance croissante, une transformation ordinale peut également entraîner un ajustement similaire. Si cette tendance est linéaire, il peut être approprié de considérer la variable comme numérique. Toutefois, si la fusion des modalités ou la modification des niveaux de codage est garantie, l'analyse ne varie pas de façon significative.

Figure 1-1
Diagramme de transformation de prix (numérique)



Codes de modalité

Soyez vigilant lorsque vous codez des variables qualitatives, car certains systèmes de codage peuvent générer des résultats indésirables ou des analyses incomplètes. Les systèmes de codage applicables à la variable *travail* sont répertoriés dans le tableau suivant.

Table 1-2
Autres systèmes de codage de la variable travail

Modalité	A	B	C	D
stagiaire	1	1	5	1
représentant	2	2	6	5
Gestionnaire	3	7	7	3

Certaines procédures de modalité exigent que la plage de valeurs de chaque variable soit définie. Toute valeur en dehors de cette plage est considérée comme manquante. La valeur minimale de modalité est toujours égale à 1. La valeur maximale de modalité, quant à elle, est fournie par l'utilisateur. Cette valeur ne représente pas le *nombre* de modalités d'une variable.— Il s'agit de la valeur *maximale* de modalité. Par exemple, dans ce tableau, le système de codage A est doté d'une modalité maximale égale à 3, et le système de codage B, d'une valeur maximale de modalité égale à 7. Toutefois, ces deux systèmes codent les trois mêmes modalités.

La plage de variables détermine les modalités qui sont omises de l'analyse. Les modalités ayant des codes en dehors de la plage définie sont également omises de cette analyse. Cette méthode est certes simple pour omettre des modalités, mais elle peut entraîner des analyses indésirables. Une modalité mal définie peut omettre des modalités *valides* de l'analyse. Par exemple, pour le système de codage B, définir la valeur maximale de modalité sur 3 signifie que la variable

travail possède des modalités codées de 1 à 3. La modalité de *responsable* est considérée comme manquante. Aucune modalité n'ayant été réellement codée 3, la troisième modalité de l'analyse ne contient aucune observation. Si vous souhaitez omettre toutes les modalités de responsable, cette analyse est tout à fait appropriée. Toutefois, si des responsables doivent être ajoutés, la modalité maximale doit être définie sur 7 et les valeurs manquantes doivent être codées avec des valeurs supérieures à 7 ou inférieures à 1.

Pour les variables considérées comme nominales ou ordinales, la plage des modalités n'a aucune incidence sur les résultats. Pour les valeurs nominales, seule l'étiquette, et non la valeur qui lui est associée, est déterminante. Pour les variables ordinales, l'ordre des modalités est conservé dans les quantifications. Les valeurs de modalité proprement dites ne sont pas importantes. Tous les systèmes de codage aboutissant au même classement des modalités auront des résultats identiques. Par exemple, les trois premiers systèmes de codage du tableau sont fonctionnellement équivalents si la variable *travail* est analysée à un niveau ordinal. L'ordre des modalités est identique dans ces systèmes. En revanche, le système de codage D inverse les deuxième et troisième modalités, et génère des résultats différents de ceux des autres systèmes.

Bien que de nombreux systèmes de codage de variable soient fonctionnellement équivalents, on leur préfère l'utilisation d'autres systèmes présentant de légères différences entre les codes, car ces derniers influent sur le nombre de résultats générés par une procédure. Toutes les modalités codées dotées de valeurs comprises entre 1 et la valeur maximale définie par l'utilisateur sont valides. Si l'une de ces modalités est vide, les quantifications correspondantes seront manquantes par défaut ou nulles, selon la procédure utilisée. Bien qu'aucune de ces affectations n'ait d'incidence sur les analyses, des résultats sont créés pour ces modalités. Par conséquent, pour le système de codage B, la variable *travail* possède quatre modalités recevant des valeurs manquantes par défaut. Pour le système de codage C, on distingue également quatre modalités recevant des indicateurs manquants par défaut. En revanche, pour le système de codage A, il n'existe aucune quantification manquante par défaut. Utiliser des entiers consécutifs en tant que codes pour les variables traitées comme des variables nominales ou ordinales génère beaucoup moins de résultats sans affecter pour autant les autres résultats.

Les systèmes de codage des variables considérées comme numériques sont plus restreints que l'observation ordinale. Pour ces variables, les différences entre les modalités consécutives sont significatives. Le tableau suivant répertorie trois systèmes de codage pour la variable *âge*.

Table 1-3
Autres systèmes de codage d'âge

Modalité	A	B	C
20	20	1	1
22	22	3	2
25	25	6	3
27	27	8	4

Tout recodage des variables numériques doit conserver les différences entre les modalités. Une méthode garantissant leur conservation consiste à utiliser les valeurs d'origine. Toutefois, nombreuses sont les modalités qui risquent d'avoir au final des indicateurs manquants par défaut. Par exemple, le système de codage A emploie les valeurs observées d'origine. Pour les procédures de modalité, à l'exception de l'Analyse des correspondances, la valeur maximale de modalité est égale à 27, et la valeur minimale de modalité est définie sur 1. Les 19 premières modalités sont vides et reçoivent des indicateurs manquants par défaut. Le nombre de résultats peut devenir

rapidement conséquent si la modalité maximale est nettement supérieure à 1 et qu'il existe de nombreuses modalités vides comprises entre 1 et la valeur maximale.

Pour réduire le nombre de résultats, vous pouvez procéder à un recodage. Néanmoins, pour les variables numériques, vous ne devez pas utiliser la fonction de recodage automatique. Le codage appliqué aux entiers consécutifs génère des différences de 1 entre toutes les modalités consécutives et, par conséquent, l'ensemble des quantifications est espacé de la même manière. Les caractéristiques métriques jugées primordiales lorsqu'une variable est considérée comme numérique sont supprimées par l'application d'un recodage aux entiers consécutifs. Par exemple, le système de codage C du tableau correspond au recodage automatique de la variable *âge*. La différence entre les modalités 22 et 25 passe de trois à un. Les quantifications reflètent ce changement.

Un autre système de recodage conservant les différences entre les modalités consiste à déduire de chaque modalité la plus petite valeur de la modalité et à ajouter 1 à chaque différence. Le système B constitue l'aboutissement de cette transformation. La plus petite valeur de modalité, 20, a été déduite de chaque modalité, et 1 a été ajouté à chaque résultat. Les codes transformés possèdent une valeur minimale, 1, et l'ensemble des différences est identique aux données d'origine. La valeur maximale de modalité est désormais égale à 8. En outre, les quantifications nulles précédant la première quantification non nulle sont toutes supprimées. Toutefois, les quantifications non nulles qui correspondent à chaque modalité issue du système B sont identiques aux quantifications du système A.

Utilisation de la procédure la plus adaptée à votre application

Les méthodes contenues dans quatre de ces procédures (analyse des correspondances, analyse de correspondance multiple, analyse en composantes principales qualitatives et analyse de corrélation canonique non linéaire) font partie du processus général d'analyse des données multivariées, appelé **réduction des dimensions**. En termes plus précis, les relations entre les variables sont représentées dans plusieurs dimensions —deux ou trois— aussi souvent que possible. Vous pouvez ainsi décrire les structures ou les motifs des relations qu'il serait trop difficile de comprendre dans leur richesse et leur complexité originales. Dans les applications d'étude de marché, ces méthodes peuvent représenter un type de **configuration perceptuelle**. Ces procédures présentent un avantage majeur : elles adaptent les données à différents niveaux de codage optimal.

La régression nominale décrit la relation entre une variable de réponse qualitative et une combinaison de variables indépendantes qualitatives. L'influence de chaque variable indépendante sur la variable de réponse est signalée par la pondération de régression correspondante. Comme dans les autres procédures, les données peuvent être analysées avec plusieurs niveaux de codage optimal.

Le positionnement et le dépliage multidimensionnels décrivent les relations entre les objets dans un espace de petite dimension à l'aide des proximités entre les objets.

Voici quelques règles applicables à chaque procédure :

- Utilisez la régression nominale pour prévoir les valeurs d'une variable dépendante qualitative issue d'une combinaison de variables indépendantes qualitatives.
- Utilisez l'analyse en composantes principales qualitatives pour représenter les motifs de variation d'un ensemble de variables de niveaux de codage optimal mixtes.

- Utilisez l'analyse de corrélation canonique non linéaire pour évaluer l'importance de la corrélation de plusieurs ensembles de variables de niveaux de codage optimal mixtes.
- Utilisez l'analyse des correspondances pour analyser les tableaux de contingence à deux entrées ou les données pouvant être fournies en tant que tableau à deux entrées, comme les données de préférence de marque ou de choix sociométrique.
- Utilisez l'analyse de correspondance multiple pour analyser une matrice de données multivariées qualitatives si vous souhaitez simplement que les variables soient analysées au niveau nominal.
- Utilisez le positionnement multidimensionnel pour analyser des données de proximité. L'objectif est de trouver une représentation à moindres carrés d'un seul ensemble d'objets dans un espace de petite dimension.
- Utilisez le dépliage multidimensionnel pour analyser des données de proximité. L'objectif est de trouver une représentation à moindres carrés de deux ensembles d'objets dans un espace de petite dimension.

Régression nominale

La régression nominale convient le mieux si votre analyse a pour but de prévoir une variable (de réponse) dépendante issue d'un ensemble de variables indépendantes. Comme pour toutes les procédures de codage optimal, des valeurs d'échelle sont attribuées à chaque modalité de chaque variable, afin que ces valeurs soient optimales par rapport à la régression. La solution d'une régression nominale optimise la corrélation carrée entre la réponse transformée et la combinaison pondérée de variables explicatives transformées.

Relation avec les autres procédures de modalité. La régression nominale avec codage optimal est comparable à l'analyse de corrélation canonique avec codage optimal utilisant deux ensembles, dont l'un contient uniquement la variable dépendante. Dans la dernière méthode, la similitude des ensembles est calculée par comparaison de chaque ensemble à une variable inconnue située entre tous les ensembles. Dans la régression nominale, la similitude de la réponse transformée et de la combinaison linéaire de variables explicatives transformées est évaluée directement.

Relation avec les méthodes standard. Dans la régression linéaire standard, les variables qualitatives peuvent être soit recodées en tant que variables indicatrices, soit traitées de la même manière que les variables de niveau d'intervalle. Dans la première approche, le modèle inclut une constante et une pente différentes pour chaque combinaison de niveaux des variables qualitatives. Un grand nombre de paramètres à interpréter est ainsi généré. Dans la seconde approche, un seul paramètre est estimé pour chaque variable. Toutefois, la nature arbitraire des codages de modalité rend toute généralisation impossible.

Si une partie des variables n'est pas continue, d'autres types d'analyse sont disponibles. Si la réponse est continue et les variables explicatives qualitatives, l'analyse des variances est généralement utilisée. Si la réponse est qualitative et les variables explicatives continues, la régression logistique ou l'analyse discriminante peut convenir. Si la réponse et les variables explicatives sont qualitatives, les modèles log-linéaires sont généralement utilisés.

La régression avec codage optimal fournit trois niveaux de codage pour chaque variable. Les combinaisons de ces niveaux peuvent représenter des relations non linéaires très diverses auxquelles une méthode « standard » n'est pas du tout adaptée. Par conséquent, le codage optimal s'avère une solution beaucoup plus souple que les approches standard un peu plus complexes.

En outre, les transformations non linéaires des variables explicatives réduisent habituellement les dépendances des uns par rapport aux autres. Si vous comparez les valeurs propres de la matrice de corrélation des variables explicatives avec celles de la matrice de corrélation des variables explicatives codées de façon optimale, ces dernières sont généralement moins variables que les autres. En d'autres termes, dans la régression nominale, le codage optimal réduit les valeurs propres supérieures de la matrice de corrélation des variables explicatives et incrémente les valeurs propres inférieures.

Analyse en composantes principales qualitatives

L'utilisation de l'analyse en composantes principales qualitatives convient le mieux pour représenter les motifs de variation d'un ensemble de variables de niveaux de codage optimal mixtes. Cette méthode tente de réduire le nombre de dimensions d'un ensemble de variables et de représenter cette variation dans la mesure du possible. Des valeurs d'échelle sont attribuées à chaque modalité des variables afin que ces valeurs soient optimales par rapport à la solution en composantes principales. Les objets utilisés pour l'analyse reçoivent les coordonnées des composantes basées sur les données quantifiées. Les diagrammes de coordonnées des composantes révèlent les motifs figurant parmi les objets de l'analyse, ainsi que les objets inhabituels contenus dans les données. La solution d'une analyse des composantes principales qualitatives optimise les corrélations de coordonnées des objets avec chaque variable quantifiée pour le nombre de composantes (dimensions) indiqué.

Une fonction importante des composantes principales qualitatives consiste à vérifier les données de préférence, où les répondants classent ou évaluent un nombre d'éléments par rapport à la préférence. Dans la configuration habituelle des données IBM® SPSS® Statistics, les lignes correspondent aux individus, les colonnes, aux mesures des éléments et les scores figurant sur les lignes, aux scores de préférence (sur une échelle de 0 à 10, par exemple), ce qui rend les données dépendantes des lignes. Pour les données de préférence, vous pouvez considérer les individus comme des variables. Grâce à la procédure de transposition, vous pouvez transposer ces données. Les indicateurs sont les variables et toutes les variables sont déclarées ordinales. Si vous le souhaitez, vous pouvez utiliser plus de variables que d'objets pour la procédure CATPCA.

Relation avec les autres procédures de modalité. Si toutes les variables sont déclarées nominales multiples, l'analyse en composantes principales qualitatives génère une analyse équivalant à une analyse de correspondance multiple exécutée sur les mêmes variables. Par conséquent, l'analyse en composantes principales qualitatives peut être considérée comme un type d'analyse de correspondance multiple dans lequel certaines variables sont déclarées ordinales ou numériques.

Relation avec les méthodes standard. Si toutes les variables sont codées au niveau numérique, l'analyse en composantes principales qualitatives équivaut à l'analyse en composantes principales standard.

Plus généralement, l'analyse en composantes principales qualitatives représente un autre moyen de calculer les corrélations entre les échelles non numériques, et de leur appliquer une analyse factorielle ou en composantes principales standard. Toute utilisation simpliste du coefficient de

corrélation de Pearson habituel comme mesure d'association de données ordinales peut avoir une incidence significative sur l'estimation des corrélations.

Analyse de corrélation canonique non linéaire

L'Analyse de corrélation canonique non linéaire est une procédure très générale comportant de nombreuses tâches. Ce type d'analyse a pour but d'analyser les relations entre plusieurs ensembles de variables, au lieu des variables proprement dites, comme dans l'analyse en composantes principales. Par exemple, vous pouvez utiliser deux ensembles de variables : l'un peut inclure des éléments d'ordre démographique concernant un groupe de répondants, alors que l'autre peut contenir les réponses à un ensemble d'éléments d'attitude. Les niveaux de codage de l'analyse peuvent représenter une combinaison de niveaux nominal, ordinal et numérique. L'analyse de corrélation canonique avec codage optimal détermine la similitude entre les ensembles en comparant simultanément les variables canoniques de chaque ensemble à un groupe de coordonnées de compromis associé aux objets.

Relation avec les autres procédures de modalité. Si plusieurs ensembles de variables contiennent chacun une seule variable, l'analyse de corrélation canonique avec codage optimal équivaut à l'analyse en composantes principales avec codage optimal. Si toutes les variables d'une analyse de type « une variable par ensemble » sont nominales multiples, l'analyse de corrélation canonique avec codage optimal équivaut à l'analyse de correspondance multiple. Dans le cas de deux ensembles de variables, dont l'un comprend une seule variable, l'analyse de corrélation canonique avec codage optimal équivaut à la régression nominale avec codage optimal.

Relation avec les méthodes standard. L'analyse de corrélation canonique standard est une méthode statistique qui recherche une combinaison linéaire d'un premier ensemble de variables et celle d'un second ensemble de variables corrélées de façon optimale. Du fait de ces combinaisons linéaires, l'analyse de corrélation canonique peut rechercher les ensembles indépendants de combinaisons linéaires suivants, appelés variables canoniques. Le nombre maximal d'ensembles doit être égal au nombre de variables contenues dans le plus petit ensemble.

Si deux ensembles de variables sont utilisés dans l'analyse et toutes les variables définies comme étant numériques, l'analyse de corrélation canonique avec codage optimal équivaut à une analyse de corrélation canonique standard. Bien que IBM® SPSS® Statistics ne propose aucune procédure d'analyse de corrélation canonique, vous pouvez obtenir une bonne partie des statistiques concernées par le biais de l'analyse multivariée des variances.

L'analyse de corrélation canonique avec codage optimal fournit de nombreuses fonctions. Si vous utilisez deux ensembles de variables et que l'un d'eux contient une variable nominale déclarée nominale simple, les résultats de l'analyse de corrélation canonique avec codage optimal peuvent être interprétés d'une manière similaire à ceux de l'analyse de régression. Si vous considérez que cette variable est nominale multiple, l'analyse avec codage multiple constitue une alternative à l'analyse discriminante. Regrouper les variables dans plus de deux ensembles vous permet d'analyser les données de différentes manières.

Analyse des correspondances

L'analyse des correspondances a pour but de créer des diagrammes doubles pour les tableaux de correspondances. Dans un tableau de correspondances, les variables de ligne et de colonne sont supposées représenter les modalités non classées. Par conséquent, le niveau de codage optimal nominal est systématiquement utilisé. Seules les données nominales sont recherchées dans ces deux types de variable. Il s'agit en réalité de tenir compte du fait que certains objets se trouvent dans la même modalité, alors que ce n'est pas le cas pour d'autres. Aucune hypothèse n'est avancée concernant la distance ou l'ordre entre les modalités de la même variable.

L'analyse des correspondances peut notamment servir à analyser les tableaux de contingence à deux entrées. Si un tableau possède *r* lignes actives et *c* colonnes actives, le nombre de dimensions de la solution d'analyse des correspondances correspond au nombre minimal de *r* moins 1 ou de *c* moins 1, selon la valeur la plus faible. En d'autres termes, vous pouvez parfaitement représenter les modalités de ligne ou de colonne d'un tableau de contingence dans un espace de dimensions. En pratique, vous pouvez néanmoins représenter les modalités de ligne et de colonne d'un tableau à deux entrées dans un espace comportant peu de dimensions, plus précisément deux, pour la simple raison que la compréhension des diagrammes bidimensionnels est bien plus facile que celle des représentations spatiales multidimensionnelles.

Lorsqu'un nombre de dimensions inférieur au nombre maximal de dimensions possibles est utilisé, les statistiques créées lors de l'analyse décrivent la manière dont les modalités de ligne et de colonne sont reproduites dans la représentation comportant peu de dimensions. Si la qualité de la représentation de la solution bidimensionnelle est satisfaisante, vous pouvez vérifier les diagrammes des points de ligne et de colonne pour déterminer les modalités similaires de la variable de ligne et de la variable de colonne, et les modalités de ligne et de colonne similaires les unes aux autres.

Relation avec les autres procédures de modalité. L'analyse simple des correspondances se limite aux tableaux à deux entrées. Si plusieurs variables vous intéressent, vous pouvez en combiner certaines pour créer des variables d'interaction. Par exemple, pour les variables *région*, *travail* et *âge*, vous pouvez combiner *région* et *travail* afin de créer une variable *rétrav* possédant les 12 modalités répertoriées dans le tableau suivant. Cette variable crée un tableau à deux entrées avec la variable *âge* (12 lignes, 4 colonnes), qui peut faire l'objet d'une analyse de correspondances.

Table 1-4
Combinaisons des variables région et travail

Code de modalité	Définition de modalité	Code de modalité	Définition de modalité
1	Nord, stagiaire	7	Est, stagiaire
2	Nord, représentant	8	Est, représentant
3	Nord, responsable	9	Est, responsable
4	Sud, stagiaire	10	Ouest, stagiaire
5	Sud, représentant	11	Ouest, représentant
6	Sud, responsable	12	Ouest, responsable

Cette approche présente un défaut, à savoir que toute paire de variables peut être combinée. Nous pouvons combiner *travail* et *âge*, et ainsi obtenir une autre variable de 12 modalités. Nous pouvons également combiner *région* et *âge*, ce qui entraîne la création d'une variable de 16 modalités. Chacune de ces variables d'interaction génère un tableau à deux entrées avec l'autre variable. Les analyses des correspondances de ces trois tableaux donnent des résultats différents,

même si chaque résultat est valide. En outre, dans le cas de quatre variables au moins, vous pouvez créer des tableaux à deux entrées comparant une variable d'interaction avec une autre. Le nombre de tableaux possibles à analyser peut devenir très important, même pour quelques variables seulement. Vous pouvez combiner l'un de ces tableaux pour l'analyse ou les analyser tous. Vous pouvez également utiliser la procédure d'analyse de correspondance multiple pour vérifier toutes les variables à la fois sans avoir à créer de variables d'interaction.

Relation avec les méthodes standard. En outre, la procédure de tableau croisé permet d'analyser les tableaux de contingence, avec l'indépendance comme valeur commune aux différentes analyses. Toutefois, même dans les petits tableaux, déterminer l'origine d'un départ à partir de la valeur d'indépendance peut s'avérer complexe. L'analyse des correspondances est utile car elle analyse ces motifs pour les tableaux à deux entrées, quelle que soit leur taille. En cas d'association entre les variables de ligne et de colonne (c'est-à-dire si la valeur Khi-deux est significative), l'analyse des correspondances peut contribuer à révéler la nature de la relation.

Analyse de correspondance multiple

L'Analyse de correspondance multiple tente de créer une solution dans laquelle les objets faisant partie de la même modalité sont représentés proches les uns des autres, et les objets de modalités différentes, éloignés les uns des autres. Chaque objet se trouve aussi près que possible des points de modalité qui s'appliquent. Ainsi, les modalités divisent les objets en sous-groupes homogènes. Les variables sont considérées comme homogènes lorsqu'elles classent les objets des mêmes modalités dans les mêmes sous-groupes.

Pour une solution en une dimension, l'analyse de correspondance multiple attribue des valeurs d'échelle optimale (quantifications de modalité) à chaque modalité de chaque variable si bien que, dans l'ensemble, les modalités ont en moyenne une étendue maximale. Pour une solution en deux dimensions, l'analyse de correspondance multiple recherche un second ensemble de quantifications des modalités de chaque variable non lié au premier ensemble, en réessayant d'optimiser l'étendue, etc. Les modalités recevant autant de coordonnées qu'il existe de dimensions, les variables de l'analyse sont censées être nominales multiples au niveau de codage optimal.

L'analyse de correspondance multiple affecte également des coordonnées aux objets de l'analyse, afin que les quantifications de modalité représentent les moyennes, ou barycentres, des coordonnées des objets de la modalité.

Relation avec les autres procédures de modalité. L'analyse de correspondance multiple est également appelée analyse d'homogénéité ou double codage. Elle fournit des résultats, certes comparables mais pas identiques, à ceux de l'analyse des correspondances lorsque seules deux variables sont utilisées. L'analyse des correspondances génère des résultats uniques récapitulant l'ajustement et la qualité de la représentation de la solution, y compris les informations de stabilité. Par conséquent, dans le cas de deux variables, il vaut mieux généralement préférer l'analyse des correspondances à l'analyse de correspondance multiple. Ces deux procédures présentent une autre différence : le point de départ de l'analyse de correspondance multiple est une matrice de données, dans laquelle les lignes sont des objets et les colonnes sont des variables. Quant au point de départ de l'analyse des correspondances, il peut être la même matrice de données, une matrice de proximité générale ou un tableau de contingence joint, qui est une matrice récapitulative où les lignes et les colonnes représentent des modalités de variables. L'analyse de correspondance

multiple peut également être assimilée à l'analyse en composantes principales de données codées au niveau nominal multiple.

Relation avec les méthodes standard. L'analyse de correspondance multiple peut être considérée comme étant l'analyse d'un tableau de contingence à entrées multiples. Un tableau de contingence à entrées multiples peut également être analysé avec la procédure de tableaux croisés, mais celle-ci fournit des statistiques récapitulatives distinctes pour chaque modalité de chaque variable de contrôle. Grâce à l'analyse de correspondance multiple, il est généralement possible de récapituler la relation entre toutes les variables à l'aide d'un diagramme bidimensionnel. Un mode d'utilisation avancé de ce type d'analyse consiste à remplacer les valeurs de modalité d'origine par les valeurs d'échelle optimale de la première dimension, puis à effectuer une analyse multivariée secondaire. Puisque l'analyse de correspondance multiple remplace les étiquettes de modalité par des valeurs d'échelle numérique, de nombreuses procédures nécessitant des données numériques peuvent être appliquées lorsqu'elle est terminée. Par exemple, la procédure d'analyse factorielle crée une première composante principale équivalant à la première dimension de l'analyse de correspondance multiple. Les coordonnées des composantes de la première dimension sont identiques à celles des objets et les corrélations entre composantes, aux mesures de discrimination. Néanmoins, la deuxième dimension de l'analyse de correspondance multiple est différente de celle de l'analyse factorielle.

Positionnement multidimensionnel

Le positionnement multidimensionnel convient le mieux si votre analyse a pour but de rechercher une structure dans un ensemble de mesures de distance entre un ensemble d'objets ou d'observations unique. Pour cela, il affecte les observations à des positions particulières dans un espace conceptuel de petite dimension de telle sorte que les distances entre les points dans l'espace correspondent le mieux possible aux dissimilarités données. Le résultat est une représentation à moindres carrés des objets dans cet espace de petite dimension, qui vous aidera, dans certains cas, à mieux comprendre vos données.

Relation avec les autres procédures de modalité. Lorsque vous utilisez des données multivariées à partir desquelles vous créez des distances et que vous analysez ensuite avec le positionnement multidimensionnel, les résultats s'avèrent similaires à ceux de l'analyse des données utilisant une analyse des composantes principales qualitatives, impliquant la standardisation principale des objets. Ce type d'analyse en composantes principales est également appelé analyse des coordonnées principales.

Relation avec les méthodes standard. La procédure de positionnement multidimensionnel qualitatif (PROXSCAL) apporte des améliorations à la procédure de codage disponible dans l'option Statistiques de base (ALSCAL). PROXSCAL fournit un algorithme accéléré pour certains modèles et vous permet d'appliquer des restrictions à l'espace commun. En outre, PROXSCAL tente de minimiser le stress brut normalisé plutôt que le stress S (également appelé **pression**). En général, on dénote une certaine préférence pour le stress brut normalisé, car cette mesure est basée sur les distances, alors que le stress S est basé sur leur carré.

Dépliage multidimensionnel

Le Dépliage multidimensionnel convient mieux si votre analyse a pour but de rechercher une structure dans un ensemble de mesures de distance entre deux ensembles d'objets (appelés objets de ligne et de colonne). Pour cela, il affecte les observations à des positions particulières dans un espace conceptuel de petite dimension de telle sorte que les distances entre les points dans l'espace correspondent le mieux possible aux dissimilarités données. Le résultat est une représentation à moindres carrés des objets de ligne et de colonne dans cet espace de petite dimension, qui vous aidera, dans certains cas, à mieux comprendre vos données.

Relation avec les autres procédures de modalité. Si vos données sont constituées de distances entre un ensemble unique d'objets (une matrice carrée, symétrique), utilisez Positionnement multidimensionnel.

Relation avec les méthodes standard. La procédure de dépliage multidimensionnel des modalités (PREFSCAL) apporte des améliorations à la fonctionnalité de dépliage disponible dans l'option Statistiques de base (avec ALSCAL). PREFSCAL vous permet d'instaurer des restrictions sur l'espace commun. En outre, PREFSCAL tente de minimiser une mesure de stress pénalisée, l'aidant ainsi à éviter de dégénérer des solutions (auxquels les algorithmes précédents sont enclins).

Ratio d'aspect des diagrammes de codage optimal

Le ratio d'aspect des diagrammes de codage optimal est isotrope. Dans un diagramme bidimensionnel, la distance représentant une unité de la dimension 1 est égale à celle représentant une unité de la dimension 2. Si, dans ce type de diagramme, vous modifiez l'étendue d'une dimension, le système modifie la taille de l'autre dimension pour que les distances physiques restent égales. Il est impossible de remplacer un ratio d'aspect isotrope pour les procédures de codage optimal.

Lectures recommandées

Reportez-vous aux documents suivants pour obtenir des informations générales sur les méthodes de codage optimal.

Barlow, R. E., D. J. Bartholomew, D. J. Bremner, et H. D. Brunk. 1972. *Statistical inference under order restrictions*. New York: John Wiley and Sons.

Benzécri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. Dans : *Methodologies of Pattern Recognition,* S. Watanabe, éd. New York: Academic Press.

Bishop, Y. M., S. E. Feinberg, et P. W. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. Cambridge, Massachusetts: MIT Press.

De Leeuw, J. 1984. The Gifi system of nonlinear multivariate analysis. Dans : *Data Analysis and Informatics III,* E. Diday, et al., éd..

De Leeuw, J. 1990. Multivariate analysis with optimal scaling. Dans : *Progress in Multivariate Analysis,* S. Das Gupta, et J. Sethuraman, éds. Calcutta: Indian Statistical Institute.

De Leeuw, J., et J. Van Rijckevorsel. 1980. HOMALS and PRINCALS—Some generalizations of principal components analysis. Dans : *Data Analysis and Informatics,* E. Diday, et al., éd. Amsterdam: North-Holland.

De Leeuw, J., F. W. Young, et Y. Takane. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .

Gifi, A. 1990. *Nonlinear multivariate analysis*. Chichester: John Wiley and Sons.

Heiser, W. J., et J. J. Meulman. 1995. Nonlinear methods for the analysis of homogeneity and heterogeneity. Dans : *Recent Advances in Descriptive Multivariate Analysis,* W. J. Krzanowski, éd. Oxford: Oxford University Press.

Israëls, A. 1987. *Eigenvalue techniques for qualitative data*. Leiden: DSWO Press.

Krzanowski, W. J., et F. H. C. Marriott. 1994. *Multivariate analysis: Part I, distributions, ordination and inference*. Londres: Edward Arnold.

Lebart, L., A. Morineau, et K. M. Warwick. 1984. *Multivariate descriptive statistical analysis*. New York: John Wiley and Sons.

Max, J. 1960. Quantizing for minimum distortion. *Proceedings IEEE (Information Theory)*, 6, .

Meulman, J. J. 1986. *A distance approach to nonlinear multivariate analysis*. Leiden: DSWO Press.

Meulman, J. J. 1992. The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika*, 57, .

Nishisato, S. 1980. *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S. 1994. *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Rao, C. R. 1973. *Linear statistical inference and its applications*, 2nd éd. New York: John Wiley and Sons.

Rao, C. R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. Dans : *Multivariate Analysis, Vol. 5,* P. R. Krishnaiah, éd. Amsterdam: North-Holland.

Roskam, E. E. 1968. *Metric analysis of ordinal data in psychology*. Voorschoten: VAM.

Shepard, R. N. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, .

Wolter, K. M. 1985. *Introduction to variance estimation*. Berlin: Springer-Verlag.

Young, F. W. 1981. Quantitative analysis of qualitative data. *Psychometrika*, 46, .

Chapitre
2

Régression nominale (CATREG)

La **régression nominale** quantifie les données qualitatives en affectant des valeurs numériques aux modalités ; une équation de régression linéaire optimale est ainsi créée pour les variables transformées. La régression nominale est également appelée CATREG, acronyme de *cat*egorical *reg*ression.

L'analyse de la régression linéaire standard implique la réduction des différences de sommes des carrés entre une variable de réponse (dépendante) et une combinaison pondérée des prédicteurs (variables indépendantes). Les variables sont habituellement quantitatives, les données nominales étant recodées en variables binaires ou de contraste. En conséquence, les variables qualitatives servent à séparer les groupes d'observations et cette technique estime des séries de paramètres distinctes pour chaque groupe. Les coefficients estimés reflètent le mode d'affectation de la réponse due aux modifications des prédicteurs. Il est possible de prévoir la réponse pour n'importe quelle combinaison de valeurs de variables indépendantes.

Une autre approche consiste à effectuer la régression de la réponse sur les valeurs des variables indépendantes nominales proprement dites. Dans ce cas, un seul coefficient est estimé pour chaque variable. Toutefois, pour les variables qualitatives, les valeurs des modalités sont arbitraires. Le codage des modalités selon plusieurs méthodes produit différents coefficients, ce qui complique les comparaisons d'analyses portant sur les mêmes variables.

CATREG constitue une extension de l'approche standard en codant simultanément les variables qualitatives, ordinales et numériques. Cette procédure quantifie les variables qualitatives afin que les valeurs affectées reflètent les caractéristiques des modalités d'origine. La procédure traite les variables qualitatives quantifiées de la même façon que les variables numériques. L'utilisation de transformations non linéaires permet d'analyser les variables à différents niveaux afin de déterminer le modèle correspondant au meilleur ajustement possible.

Exemple : La régression nominale peut être utilisée pour décrire dans quelle mesure la satisfaction professionnelle dépend de la modalité d'emploi, de la région et de la durée du transport. Vous pourriez ainsi déterminer que les plus hauts niveaux de satisfaction professionnelle correspondent aux postes de direction et aux temps de transport les plus faibles. Vous avez ainsi la possibilité d'utiliser l'équation de régression résultante pour prévoir la satisfaction professionnelle relative à n'importe quelle combinaison de ces trois variables indépendantes.

Diagrammes et statistiques : Fréquences, coefficients de régression, tableau ANOVA, historique des itérations, valeurs affectées aux modalités, corrélations entre variables indépendantes non transformées, corrélations entre variables indépendantes transformées, les diagrammes de résidus et de transformation.

Données. CATREG traite les variables indicatrices de modalités. Les indicateurs de modalités doivent être des nombres entiers positifs. Vous pouvez utiliser la boîte de dialogue Discrétisation pour convertir les variables fractionnées et les variables chaîne en nombres entiers positifs.

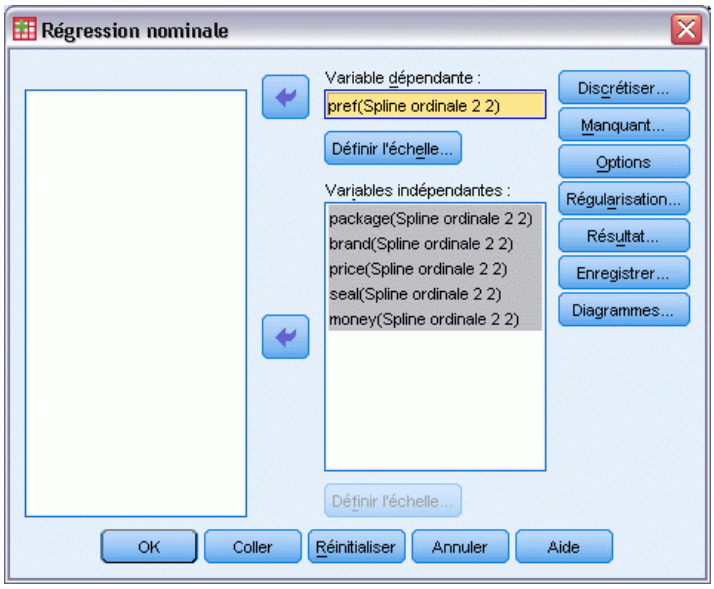
Hypothèses : Une seule variable de réponse est autorisée, mais le nombre maximal de variables explicatives est de 200. Les données doivent comporter au moins trois observations valides, le nombre d'observations valides ne devant pas dépasser le nombre de variables indépendantes plus un.

Procédures apparentées : La procédure CATREG équivaut à la procédure d'analyse de corrélation canonique nominale avec codage optimal (OVERALS) avec deux groupes, dont l'un ne comporte qu'une seule variable. Le codage de toutes les variables au niveau numérique correspond à l'analyse de régression multiple standard.

Pour obtenir une régression nominale

▶ A partir des menus, sélectionnez :
Analyse > Régression > Codage optimal (CATREG)...

Figure 2-1
Boîte de dialogue Régression nominale



▶ Sélectionnez la variable dépendante, ainsi que la ou les variables indépendantes.

▶ Cliquez sur OK.

Sinon, modifiez le niveau de codage de chaque variable.

Définir une échelle dans la régression nominale

Vous pouvez définir le niveau de codage optimal des variables dépendantes et indépendantes. Par défaut, elles sont codées comme des splines monotones de second degré (ordinales) avec deux points critiques intérieurs. En outre, vous pouvez également définir la pondération pour les variables d'analyse.

Figure 2-2
Boîte de dialogue Définir l'échelle

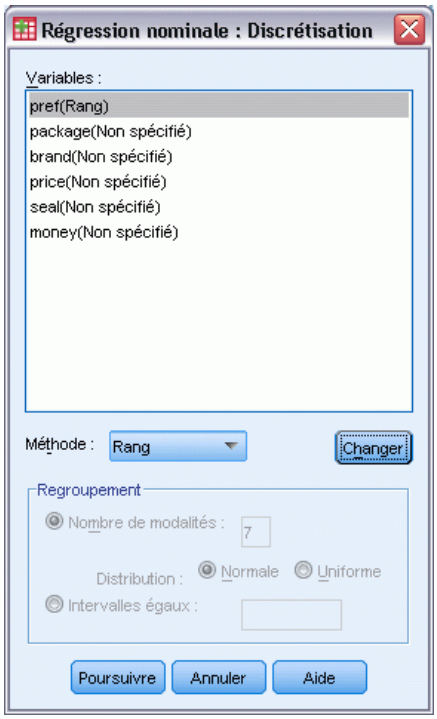
Niveau du codage optimal : Vous pouvez également sélectionner le niveau de codage pour la quantification de chaque variable.

- **Spline ordinale :** L'ordre des modalités de la variable observée est conservé dans la variable codée de façon optimale. Les points des modalités se trouvent sur une ligne droite (vecteur) passant par l'origine. La transformation résultante est un modèle polynomial monotone lissé du degré choisi. Ses différents éléments dépendent du nombre de noeuds intérieurs défini par l'utilisateur ainsi que du positionnement de ces derniers, déterminé par la procédure.
- **Spline nominale :** La seule information de la variable observée qui est conservée dans la variable codée de façon optimale est le groupe des objets dans les modalités. L'ordre des modalités de la variable observée n'est pas conservé. Les points des modalités se trouvent sur une ligne droite (vecteur) passant par l'origine. La transformation résultante est un modèle polynomial lissé, peut-être non monotone, du degré choisi. Ses différents éléments dépendent du nombre de noeuds intérieurs défini par l'utilisateur ainsi que du positionnement de ces derniers, déterminé par la procédure.
- **Ordinal.** L'ordre des modalités de la variable observée est conservé dans la variable codée de façon optimale. Les points des modalités se trouvent sur une ligne droite (vecteur) passant par l'origine. La transformation du résultat convient mieux que la transformation ordinale spline, mais s'avère moins lissée.
- **Nominal.** La seule information de la variable observée qui est conservée dans la variable codée de façon optimale est le groupe des objets dans les modalités. L'ordre des modalités de la variable observée n'est pas conservé. Les points des modalités se trouvent sur une ligne droite (vecteur) passant par l'origine. La transformation du résultat convient mieux que la transformation nominale spline mais s'avère moins lissée.
- **Numérique.** Les modalités sont considérées comme triées et espacées régulièrement (niveau d'intervalle). L'ordre des modalités ainsi que les distances égales entre les nombres de modalités de la variable sont conservées dans la variable codée de façon optimale. Les points des modalités se trouvent sur une ligne droite (vecteur) passant par l'origine. Lorsque toutes les variables sont au niveau numérique, l'analyse est analogue à celle en composantes principales standard.

Régression nominale : Discrétisation

La boîte de dialogue Discrétisation vous permet de choisir une méthode de recodage des variables. Les valeurs fractionnées sont regroupées en sept modalités (ou en nombre de valeurs distinctes de variables si le nombre est inférieur à sept) avec une distribution normale approximative, à moins qu'une autre configuration ne soit spécifiée. Les variables chaîne sont toujours converties en nombres entiers positifs en affectant des indicateurs de modalités selon l'ordre croissant alphanumérique. La discrétisation des variables chaîne s'applique à ces nombres entiers. Par défaut, d'autres variables sont laissées inutilisées. Les variables discrétisées sont ensuite utilisées dans l'analyse.

Figure 2-3
Discrétisation



Méthode. Choisissez entre Regroupement, Rang et Multiplier.

- **Regroupement :** Recodez en un nombre spécifié de modalités ou par intervalle.
- **Classement.** La variable est discrétisée via le classement des observations.
- **Multiplier :** Les valeurs courantes de la variable sont standardisées, multipliées par 10 et arrondies, et possèdent une constante ajoutée de sorte que la valeur discrétisée la plus faible soit égale à 1.

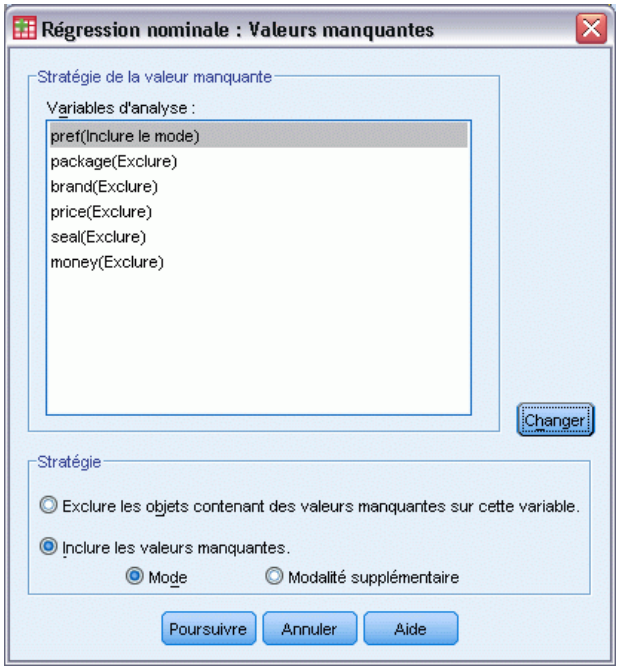
Regroupement : Les options suivantes sont disponibles lorsque vous discrétisez des variables par groupe :

- **Nombre de modalités :** Indiquez un nombre de modalités et définissez si les valeurs de la variable doivent faire l'objet d'une distribution approximativement gaussienne ou uniforme entre ces modalités.
- **Intervalles égaux :** Les variables sont recodées en modalités définies par ces intervalles de taille égale. N'oubliez pas de spécifier la longueur des intervalles.

Régression nominale : Valeurs manquantes

La boîte de dialogue Valeurs manquantes vous permet de choisir la stratégie de gestion des valeurs manquantes pour les variables de l'analyse et supplémentaires.

Figure 2-4
Boîte de dialogue Valeurs manquantes



Stratégie : Vous pouvez exclure des objets contenant des valeurs manquantes (suppression par liste) ou inclure des valeurs manquantes (traitement actif).

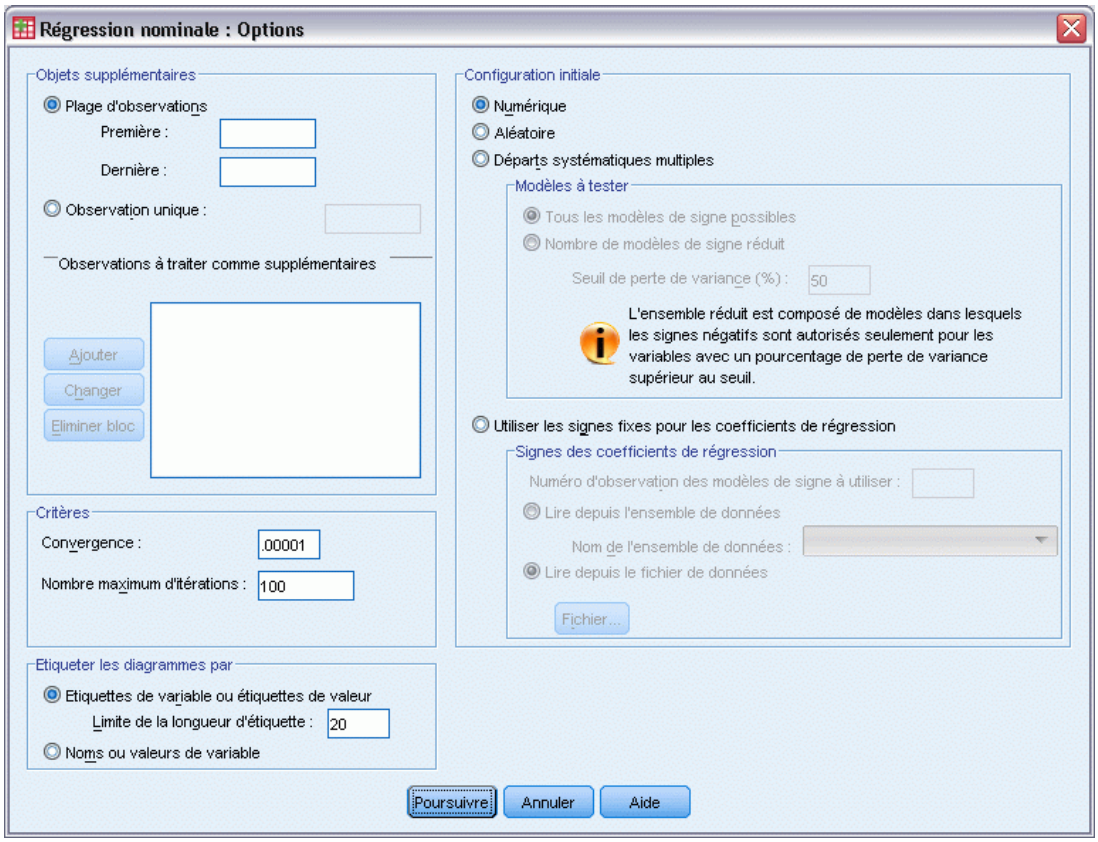
- **Exclure les objets contenant des valeurs manquantes sur cette variable.** Les objets contenant des valeurs manquantes dans la variable sélectionnée sont retirés de l'analyse. Cette option n'est pas disponible pour les variables supplémentaires.
- **Imputer les valeurs manquantes.** Des valeurs sont prises en compte pour les objets contenant des valeurs manquantes sur la variable sélectionnée. Vous pouvez choisir la méthode d'imputation. Sélectionnez Mode pour remplacer les valeurs manquantes par la modalité la plus fréquente. S'il existe plusieurs modes, utilisez celui dont l'indicateur de modalités est le plus petit. Sélectionnez Modalité supplémentaire pour remplacer les valeurs manquantes par la valeur affectée à une modalité supplémentaire. Cela suppose que les objets contenant

une valeur manquante pour cette variable sont considérés comme appartenant à la même modalité (supplémentaire).

Régression nominale : Options

La boîte de dialogue Options permet de sélectionner le style de configuration initiale, de spécifier les critères d'itération et de convergence, de sélectionner les objets supplémentaires et de définir l'étiquetage des diagrammes.

Figure 2-5
Options



Objets supplémentaires : Cette option permet de définir les objets à traiter comme objets supplémentaires. Entrez simplement le numéro d'un objet supplémentaire (ou spécifiez un intervalle d'observations), puis cliquez sur Ajouter. Vous ne pouvez pas pondérer des objets supplémentaires (les pondérations indiquées sont ignorées).

Configuration initiale : Si aucune variable n'est considérée comme nominale, sélectionnez la configuration Numérique. Si une variable au moins est considérée comme nominale, choisissez la configuration Aléatoire.

Si au moins une variable a un niveau d'échelle ordinal ou Spline ordinal, l'algorithme habituel pour les modèles peut également générer une solution moins optimale. Choisir les Départs multiples systématiques avec tous les types de signes possibles permettra toujours de trouver la

solution optimale, mais la durée d'exécution requise augmente rapidement en même temps que le nombre de variables ordinales et Spline ordinales dans l'ensemble de données. Vous pouvez réduire le nombre de types de test en spécifiant un pourcentage de perte de seuil de variance, pour lequel plus le seuil est élevé, plus le nombre de types de signes exclus augmente. Cette option ne permet pas de garantir l'obtention de la solution optimale, mais elle réduit le risque d'obtenir une solution moins optimale. De plus, si la solution optimale n'est pas trouvée, il y a moins de chances que la solution moins optimale soit très différente de la solution optimale. Lorque des départs multiples systématiques sont demandés, les signes des coefficients de régression pour chaque départ sont écrits dans un fichier de données IBM® SPSS® Statistics externe ou dans un ensemble de données de la session en cours. Pour plus d'informations, reportez-vous à la section Régression nominale : Enregistrement sur p. 25.

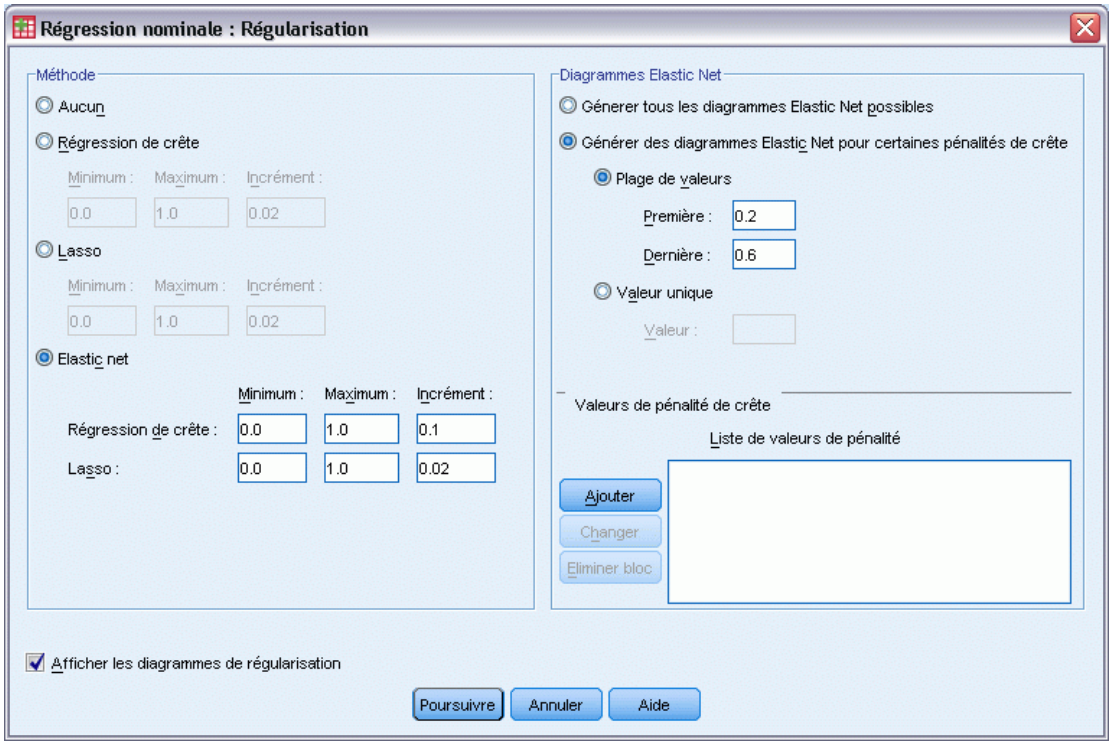
Les résultats d'une exécution précédente avec départs multiples systématiques vous permettent d' Utiliser des signes fixes pour les coefficients de régression. Les signes (indiqués par 1 et −1) doivent se trouver dans une ligne de l'ensemble de données ou du fichier spécifiés. Le chiffre de départ à valeur entière est le numéro d'observation de la ligne de ce fichier qui contient les signes à utiliser.

Critères. Vous pouvez spécifier le nombre maximal d'itérations que la régression peut prendre en charge dans ses calculs. Vous avez également la possibilité de sélectionner une valeur de critère de convergence. La régression interrompt son itération dès que la différence du total ajusté entre les deux dernières itérations est inférieur à la valeur de la convergence, ou dès que le nombre maximal d'itérations est atteint.

Etiqueter les diagrammes par : Vous permet de préciser si les étiquettes de variable et de valeurs ou les noms ou valeurs de variables sont utilisés dans les diagrammes. Vous pouvez également spécifier une longueur maximale pour les étiquettes.

Régularisation de régression nominale

Figure 2-6
Boîte de dialogue Régularisation



Méthode. Les méthodes de régularisation peuvent améliorer l'erreur de prédiction du modèle en réduisant la variabilité des estimations du coefficient de régression à l'aide d'une réduction des estimations tendant vers 0. Le Lasso et Elastic Net réduiront certaines estimations de coefficient à 0 exactement, permettant ainsi une forme de sélection de variables. Lorsqu'une méthode de régularisation est demandée, le modèle et les coefficients régularisés pour chaque valeur de coefficient de pénalité sont écrits dans un fichier de données IBM® SPSS® Statistics externe ou un ensemble de données de la session en cours. Pour plus d'informations, reportez-vous à la section Régression nominale : Enregistrement sur p. 25.

- **Régression de crête.** La régression de crête réduit les coefficients en introduisant un terme de pénalité égal à la somme des coefficients au carré multipliée par un **coefficient de pénalité**. Ce coefficient peut être compris entre 0 (aucune pénalité) et 1 ; cette procédure recherchera la "meilleure" valeur de pénalité si vous spécifiez un intervalle et un incrément.
- **Lasso.** Le terme de pénalité de Lasso est basé sur la somme des coefficients absolus et la spécification d'un coefficient de pénalité est semblable à celle d'une régression pseudo-orthogonale. Néanmoins, le Lasso nécessite beaucoup plus de calculs.
- **Elastic net.** Elastic Net regroupe simplement les pénalités de régression Lasso et de crête et effectuera une recherche dans la grille des valeurs spécifiées pour trouver les "meilleurs" coefficients de pénalité de régression Lasso et de crête. Pour une paire de pénalités de régression Lasso et de crête donnée, Elastic Net ne nécessite pas plus de calculs que le Lasso.

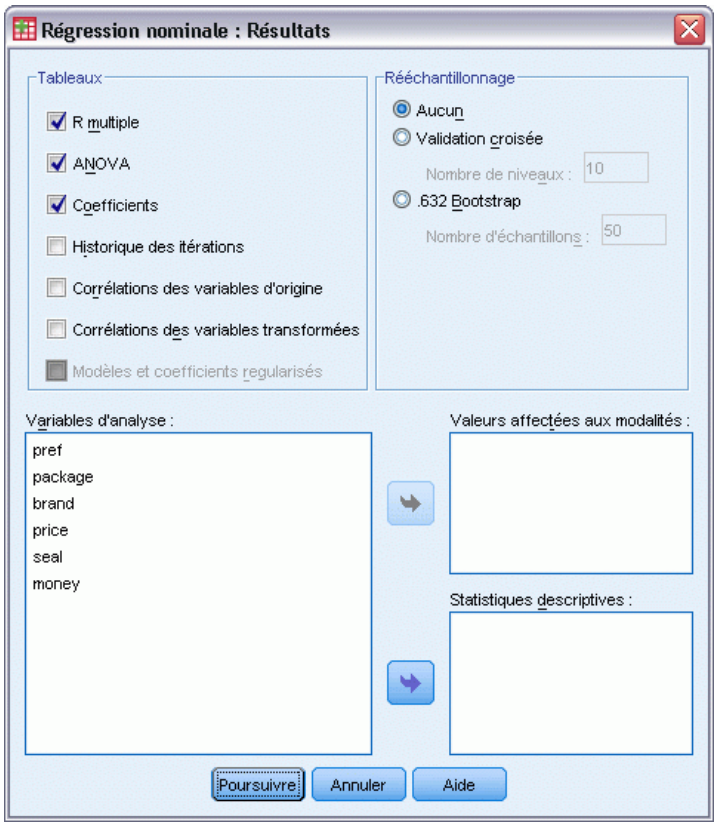
Afficher les diagrammes de régularisation. Il s'agit de diagrammes comparant les coefficients de régression et la pénalité de régularisation. Pendant que ce diagramme recherche un intervalle de valeurs pour le "meilleur" coefficient de pénalité, il affiche les modifications des coefficients de régression dans cet intervalle.

Diagrammes Elastic Net. Pour la méthode Elastic Net, des diagrammes de régularisation séparés sont générés par les valeurs de la pénalité de régression de crête. La fonction Tous les diagrammes possibles utilise chaque valeur de l'intervalle déterminé par les valeurs minimum et maximum de pénalité de régression de crête spécifiées. La fonction Pour certaines pénalités de crête permet de spécifier un sous-ensemble des valeurs dans l'intervalle déterminé par le minimum et le maximum. Entrez simplement le chiffre d'une valeur de pénalité (ou spécifiez un intervalle de valeurs), puis cliquez sur Ajouter.

Régression nominale : Résultat

La boîte de dialogue Résultats permet de sélectionner les statistiques à afficher dans le résultat.

Figure 2-7
Résultat



Tableaux. Génère des tableaux pour :

- **R multiple.** Comprend R^2 et R^2 ajusté, R^2 ajusté prend en compte le codage optimal.

- **ANOVA :** Cette option présente les sommes des carrés de régression et des résidus, le moyenne des carrés ainsi qu'un test-*F*. Deux tableaux ANOVA sont affichés : L'un avec des degrés de liberté pour la régression égaux au nombre de variables explicatives, et l'autre avec les degrés de liberté pour la régression prenant en compte le codage optimal.
- **Coefficients.** Cette option propose trois tableaux. Le tableau Coefficients : il comporte des bêtas, l'erreur standard des bêtas, des valeurs *t* et la signification ; le tableau Coefficients : Codage optimal qui contient l'erreur standard des bêtas prenant en compte les degrés de liberté du codage optimal ; le tableau des corrélations simples et partielles, qui comporte les mesures d'importance relative de Pratt pour les variables indépendantes transformées, ainsi que la tolérance avant et après transformation.
- **Historique d'itération.** Pour chaque itération, y comprises les valeurs de départ de l'algorithme, le *R* multiple et l'erreur de régression apparaissent. L'augmentation dans le *R* multiple est répertoriée en commençant à partir de la première itération.
- **Corrélations des variables d'origine :** Une matrice affichant les corrélations entre les variables sans transformation apparaît.
- **Corrélations des variables transformées :** Une matrice affichant les corrélations entre les variables transformées apparaît.
- **Modèles et coefficients régularisés.** Affiche les valeurs de pénalité, le R-deux et les coefficients de régression pour chaque modèle régularisé. Si une méthode de rééchantillonnage est spécifiée ou si des objets supplémentaires (observations de test) sont spécifiés, l'erreur de prévision ou la MSE de test sont également affichées.

Rééchantillonnage. Les méthodes de rééchantillonnage offrent une estimation de l'erreur de prédiction du modèle.

- **Validation croisée.** La validation croisée divise l'échantillon en plusieurs sous-échantillons ou niveaux. Les modèles de régression nominale sont générés en excluant à tour de rôle les données de chaque sous-échantillon. Le premier modèle est basé sur toutes les observations exceptées celles du premier sous-échantillon, le deuxième modèle est basé sur toutes les observations exceptées celles du deuxième sous-échantillon, etc. Pour chaque modèle, l'erreur de prédiction est estimée en appliquant le modèle au sous-échantillon exclu lors de sa génération.
- **Bootstrap .632** Avec le bootstrap, les observations sont extraites aléatoirement à partir des données avec remplacement. Ce processus se répète autant de fois que nécessaire pour obtenir un nombre d'échantillons du bootstrap. Un modèle est adapté à chaque échantillon du bootstrap et l'erreur de prédiction de chaque modèle est estimée par ce modèle et est ensuite appliquée aux observations ne se trouvant pas dans l'échantillon du bootstrap.

Valeurs affectées aux modalités : Les tableaux des valeurs transformées des variables sélectionnées apparaissent.

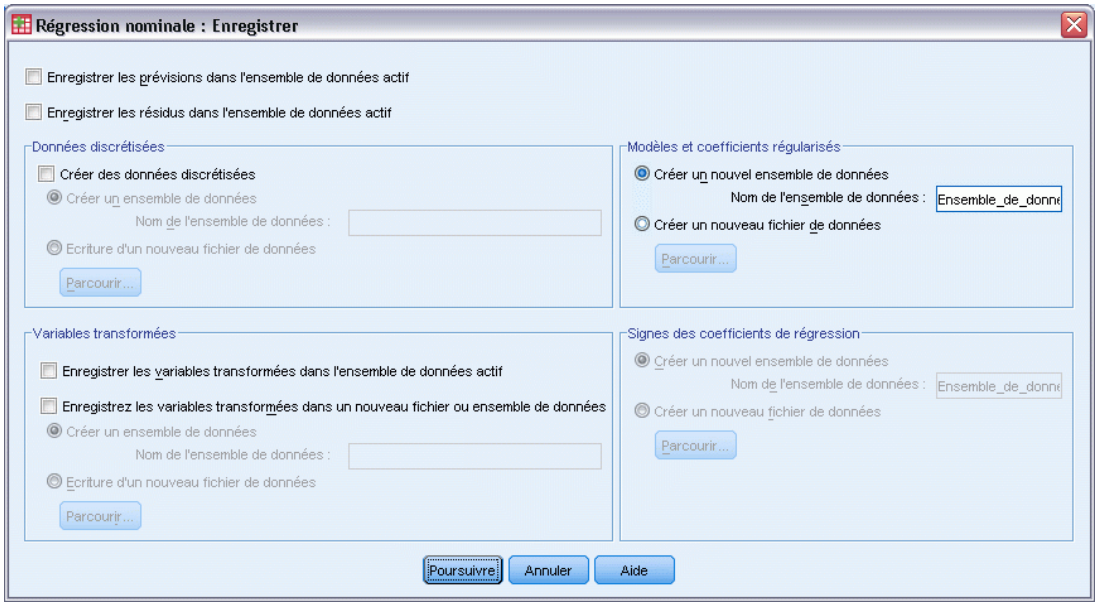
Statistiques descriptives : Les tableaux affichant les fréquences, les valeurs manquantes et les modes des variables sélectionnées apparaissent.

Régression nominale : Enregistrement

La boîte de dialogue Enregistrer vous permet d'enregistrer des prévisions, des résidus et des valeurs transformées dans l'ensemble de données actif et/ou d'enregistrer les données discrétisées, les valeurs transformées, les modèles et coefficients régularisés ainsi que les signes des coefficients de régression dans un fichier de données externe IBM® SPSS® Statistics ou un ensemble de données de la session en cours.

- Les ensembles de données sont disponibles lors de la session en cours mais ne sont pas disponibles lors des sessions suivantes, sauf si vous les enregistrez clairement comme fichiers de données. Les noms des ensembles de données doivent être conformes aux règles de dénomination de variables.
- Les noms de fichiers ou les noms de l'ensemble de données doivent être différents pour chaque type de données enregistrées.

Figure 2-8
Enregistrer



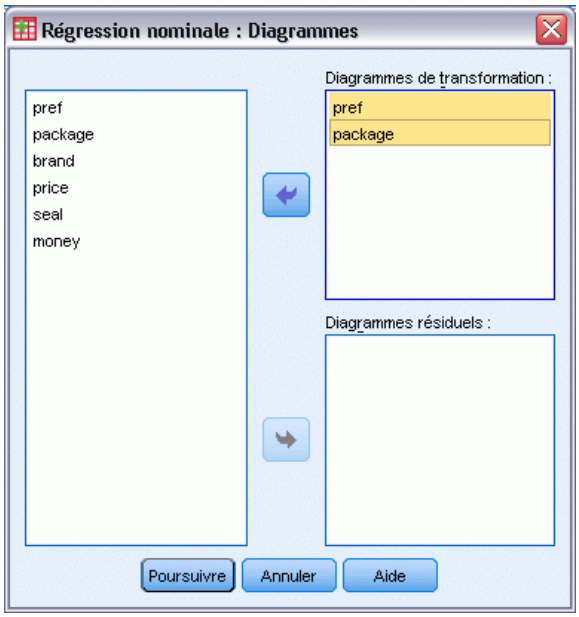
Les modèles et coefficients régularisés sont enregistrés à chaque fois qu'une méthode de régularisation est sélectionnée dans la boîte de dialogue Régularisation. Par défaut, la procédure crée un nouvel ensemble de données avec un nom unique, mais vous pouvez spécifier le nom de votre choix ou écrire dans un fichier externe.

Les signes des coefficients de régression sont enregistrés à chaque fois que des départs multiples systématiques sont utilisés comme configuration initiale dans la boîte de dialogue Options . Par défaut, la procédure crée un nouvel ensemble de données avec un nom unique, mais vous pouvez spécifier le nom de votre choix ou écrire dans un fichier externe.

Régression nominale des diagrammes de transformation

La boîte de dialogue Diagrammes vous permet de définir les variables qui produiront des diagrammes de résidus et de transformation.

Figure 2-9
Boîte de dialogue Diagrammes



Diagrammes de transformation : Pour chacune de ces variables, les valeurs affectées aux modalités sont représentées par rapport aux valeurs des modalités d'origine. Les modalités vides apparaissent sur l'axe horizontal mais n'affectent pas les calculs. Ces modalités sont identifiées par des interruptions dans la courbe reliant les valeurs affectées.

Diagrammes de résidus : Pour chacune de ces variables, les résidus (calculés pour la variable dépendante à partir de toutes les variables explicatives exceptée la variable explicative en question) sont appliqués aux indicateurs de modalités et aux valeurs affectées aux modalités optimales multipliées par bêta par rapport aux indicateurs de modalités.

Fonctionnalités supplémentaires de la commande CATREG

Vous pouvez personnaliser la régression nominale en collant vos sélections dans une fenêtre de syntaxe et en modifiant la syntaxe de commande `CATREG`. Le langage de syntaxe de commande vous permet aussi de :

- Spécifier les noms de racine des variables transformées lorsque vous les enregistrez dans l'ensemble de données actif (avec la sous-commande `SAVE`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
3

Analyse en composantes principales qualitatives (CATPCA)

Cette procédure quantifie simultanément des variables qualitatives en réduisant le nombre de dimensions des données. L'analyse en composantes principales qualitatives est également appelée CATPCA, acronyme de *CAT*egorical Principal Components Analysis.

Le but d'une telle analyse est de réduire un groupe original de variables en un groupe plus petit de composantes non corrélées représentant la plupart des informations rencontrées dans les variables d'origine. Cette technique est d'une grande utilité lorsqu'un grand nombre de variables empêche d'interpréter efficacement les relations entre les objets (sous-objets et unités). En réduisant le nombre de dimensions, vous pouvez interpréter plusieurs composantes et non plus un grand nombre de variables.

L'analyse en composantes principales standard comporte des relations linéaires entre les variables numériques. D'un autre côté, l'approche du codage optimal permet aux variables d'être codées à différents niveaux. Les variables qualitatives sont quantifiées de façon optimale par rapport au nombre de dimensions spécifié. En conséquence, des relations non linéaires entre les variables peuvent être spécifiées.

Exemple : L'analyse en composantes principales qualitatives peut être utilisée afin de représenter sur un diagramme les relations entre la modalité d'emploi, la région, le temps de transport (élevé, moyen ou faible), et la satisfaction professionnelle. Vous constatez peut-être que deux dimensions représentent une part importante de la variance. La première dimension peut séparer les modalités d'emploi par région, alors que la seconde sépare les modalités socioprofessionnelles en fonction du temps de transport. Notez également que la satisfaction professionnelle est liée au temps moyen de transport.

Diagrammes et statistiques : Effectifs, valeurs manquantes, niveau de codage optimal, mode, variance représentée par les coordonnées du barycentre, coordonnées vectorielles, total par variable et par dimension, corrélations entre composantes et variables initiales pour variables quantifiées par vecteur, valeurs affectées aux modalités et coordonnées, historique des itérations, corrélations des variables transformées et des valeurs propres de la matrice de corrélation, corrélations des variables d'origine et des valeurs propres de la matrice de corrélation, coordonnées des objets, diagrammes de modalités, diagrammes de modalités joints, diagrammes de transformation, diagrammes résiduels, diagrammes de représentation des barycentres projetés, diagrammes d'objets, diagrammes doubles, diagrammes triples et diagrammes des corrélations entre composantes et variables initiales.

Données : Les variables chaîne sont toujours converties en nombres entiers positifs par ordre croissant alphanumérique. Les valeurs manquantes définies par l'utilisateur, les valeurs manquantes par défaut et les valeurs inférieures à 1 sont considérées comme manquantes ; vous pouvez donc recoder ou ajouter une constante aux variables contenant des valeurs inférieures à 1 pour les définir comme non manquantes.

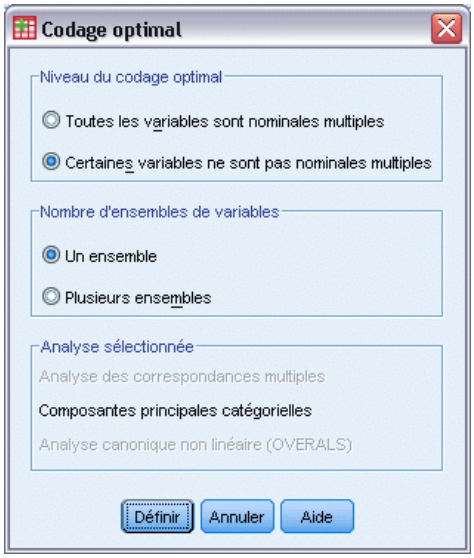
Hypothèses : Les données doivent contenir au moins trois observations valides. L'analyse repose sur des données sous forme de nombres entiers positifs. L'option de discrétisation classe automatiquement une variable fractionnée en regroupant ses valeurs en modalités avec une distribution "normale" et convertit automatiquement les valeurs des variables chaîne en nombre entiers positifs. Vous pouvez en outre, spécifier d'autres schémas de discrétisation.

Procédures apparentées : Le codage de toutes les variables au niveau numérique correspond à l'analyse en composantes principales standard. Les fonctionnalités de représentation alternée sont disponibles en utilisant les variables transformées dans une analyse en composantes principales linéaires standard. Si toutes les variables possèdent des niveaux de codage nominal multiple, l'analyse en composantes principales qualitatives est identique à l'analyse des correspondances. Si des groupes de variables sont intéressants, vous devez utiliser une analyse des corrélations canoniques nominales (non linéaires).

Obtenir une analyse en composantes principales catégorielles

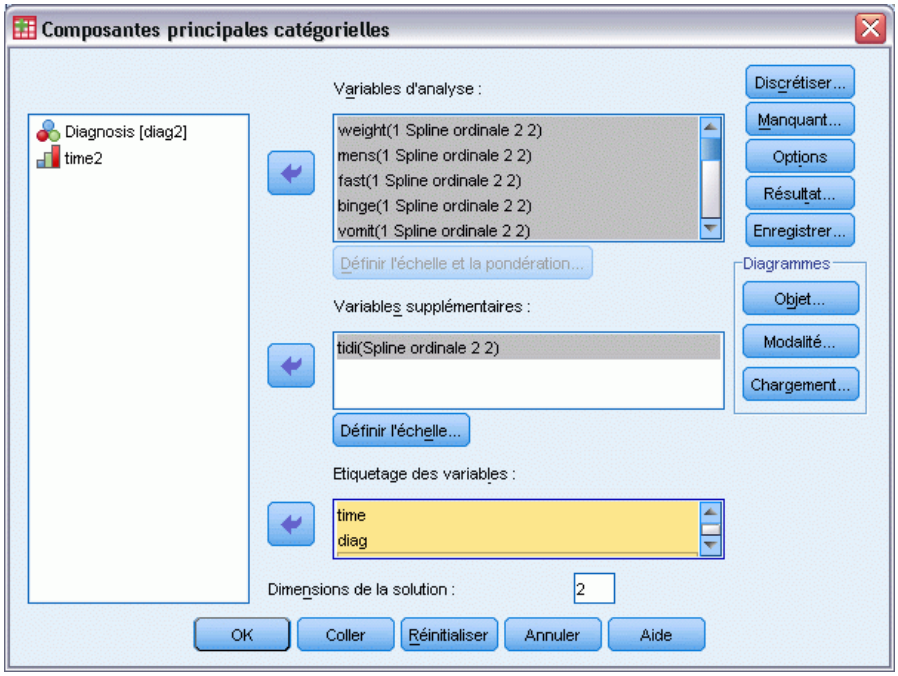
- A partir des menus, sélectionnez :
  Analyse > Réduction des dimensions > Codage optimal

Figure 3-1
Boîte de dialogue Niveau du codage optimal



- Sélectionnez Certaines variables non nominales multiples.
- Sélectionnez Un groupe.
- Cliquez sur Définir.

Figure 3-2
Boîte de dialogue Composantes principales qualitatives



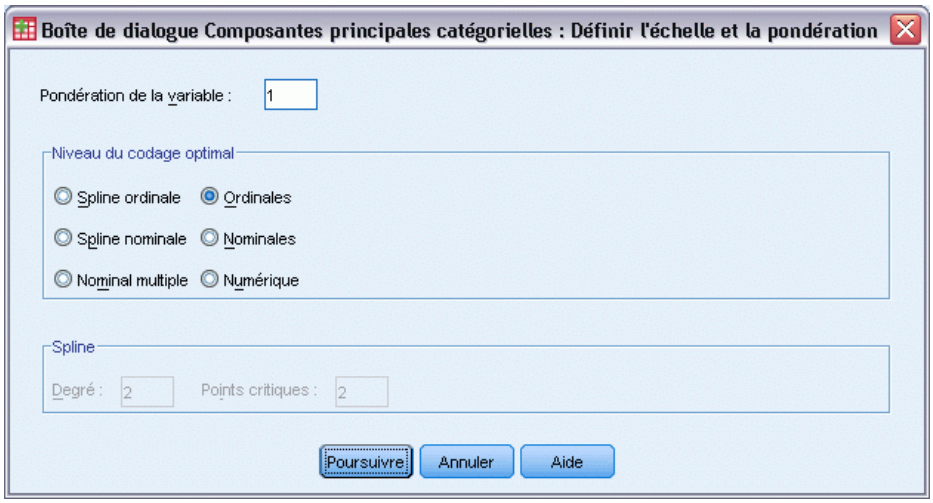
- Sélectionnez au moins deux variables d'analyse et spécifiez le nombre de dimensions de la solution.
- Cliquez sur OK.

  Vous pouvez peut-être spécifier des variables supplémentaires qui sont ajustées à la solution trouvée, ou des variables d'étiquettes pour les diagrammes.

Définir l'échelle et la pondération dans CATPCA

Vous pouvez définir le niveau de codage optimal des variables d'analyse et des variables supplémentaires. Par défaut, elles sont codées comme des splines monotones de second degré (ordinales) avec deux points critiques intérieurs. En outre, vous pouvez également définir la pondération pour les variables d'analyse.

Figure 3-3
Définir l'échelle et la pondération



Pondération de la variable : Vous pouvez choisir une pondération pour chaque variable. La valeur spécifiée doit être un nombre entier positif. La valeur par défaut est 1.

Niveau du codage optimal : Vous pouvez également sélectionner le niveau de codage à utiliser pour quantifier chaque variable.

- **Spline ordinale :** L'ordre des modalités de la variable observée est conservé dans la variable codée de façon optimale. Les points des modalités se trouvent sur une ligne droite (vecteur) passant par l'origine. La transformation résultante est un modèle polynomial monotone lissé du degré choisi. Ses différents éléments dépendent du nombre de noeuds intérieurs défini par l'utilisateur ainsi que du positionnement de ces derniers, déterminé par la procédure.
- **Spline nominale :** La seule information de la variable observée qui est conservée dans la variable codée de façon optimale est le groupe des objets dans les modalités. L'ordre des modalités de la variable observée n'est pas conservé. Les points des modalités se trouvent sur une ligne droite (vecteur) passant par l'origine. La transformation résultante est un modèle polynomial lissé, peut-être non monotone, du degré choisi. Ses différents éléments dépendent du nombre de noeuds intérieurs défini par l'utilisateur ainsi que du positionnement de ces derniers, déterminé par la procédure.
- **Nominal multiple :** La seule information de la variable observée qui est conservée dans la variable codée de façon optimale est le groupe des objets dans les modalités. L'ordre des modalités de la variable observée n'est pas conservé. Les points des modalités se trouvent sur les barycentres des objets dans les modalités particulières. L'option *Multiple* indique que divers groupes de valeurs affectées sont obtenus pour chaque dimension.
- **Ordinal :**L'ordre des modalités de la variable observée est conservé dans la variable codée de façon optimale. Les points des modalités se trouvent sur une ligne droite (vecteur) passant par l'origine. La transformation du résultat convient mieux que la transformation ordinale spline, mais s'avère moins lissée.
- **Nominal :**La seule information de la variable observée qui est conservée dans la variable codée de façon optimale est le groupe des objets dans les modalités. L'ordre des modalités de la variable observée n'est pas conservé. Les points des modalités se trouvent sur une ligne

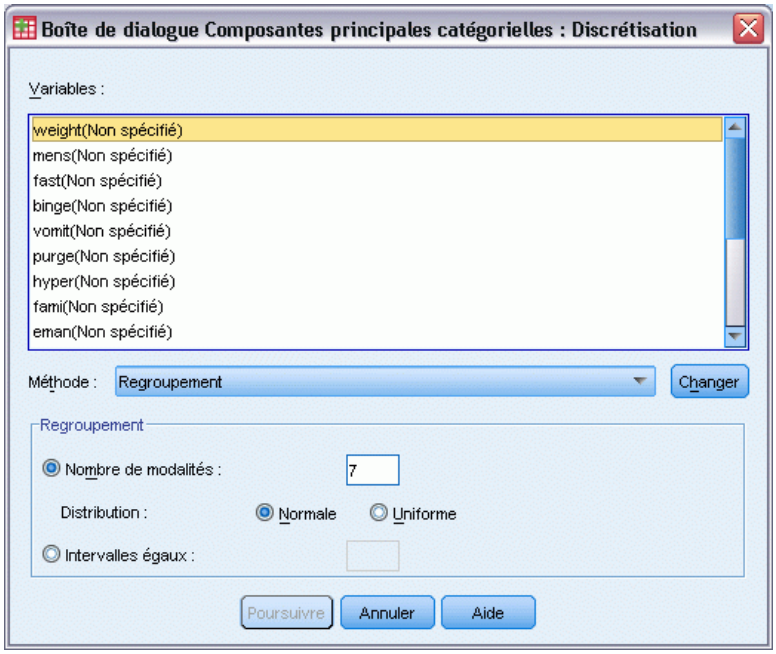
droite (vecteur) passant par l'origine. La transformation du résultat convient mieux que la transformation nominale spline mais s'avère moins lissée.

- **Numérique.** Les modalités sont considérées comme triées et espacées régulièrement (niveau d'intervalle). L'ordre des modalités ainsi que les distances égales entre les nombres de modalités de la variable sont conservées dans la variable codée de façon optimale. Les points des modalités se trouvent sur une ligne droite (vecteur) passant par l'origine. Lorsque toutes les variables sont au niveau numérique, l'analyse est analogue à celle en composantes principales standard.

Composantes principales qualitatives : Discrétisation

La boîte de dialogue Discrétisation vous permet de choisir une méthode de recodage des variables. Les valeurs fractionnées sont regroupées en sept modalités (ou en nombre de valeurs distinctes de variables si le nombre est inférieur à sept) avec une distribution normale approximative, à moins qu'une autre configuration ne soit spécifiée. Les variables chaîne sont toujours converties en nombres entiers positifs en affectant des indicateurs de modalités selon l'ordre croissant alphanumérique. La discrétisation des variables chaîne s'applique à ces nombres entiers. Par défaut, d'autres variables sont laissées inutilisées. Les variables discrétisées sont ensuite utilisées dans l'analyse.

Figure 3-4
Discrétisation



Méthode : Choisissez entre Regroupement, Rang et Multiplier.

- **Regroupement :** Recodez en un nombre spécifié de modalités ou par intervalle.

- **Rang :** La variable est discrétisée via le classement des observations.
- **Multiplier :** Les valeurs courantes de la variable sont standardisées, multipliées par 10 et arrondies, et possèdent une constante ajoutée de sorte que la valeur discrétisée la plus faible soit égale à 1.

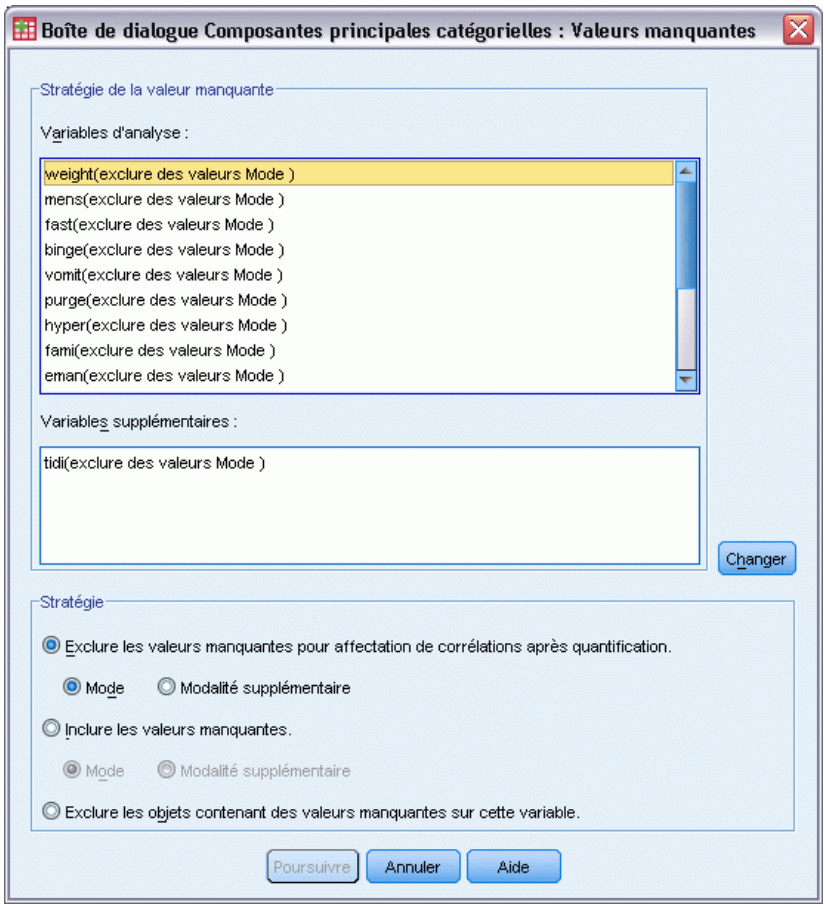
Regroupement : Les options suivantes sont disponibles lorsque vous discrétisez des variables par groupe :

- **Nombre de modalités :** Indiquez un nombre de modalités et définissez si les valeurs de la variable doivent faire l'objet d'une distribution approximativement gaussienne ou uniforme entre ces modalités.
- **Intervalles égaux :** Les variables sont recodées en modalités définies par ces intervalles de taille égale. N'oubliez pas de spécifier la longueur des intervalles.

Composantes principales qualitatives : Valeurs manquantes

La boîte de dialogue Valeurs manquantes vous permet de choisir la stratégie de gestion des valeurs manquantes pour les variables de l'analyse et supplémentaires.

Figure 3-5
Boîte de dialogue Valeurs manquantes



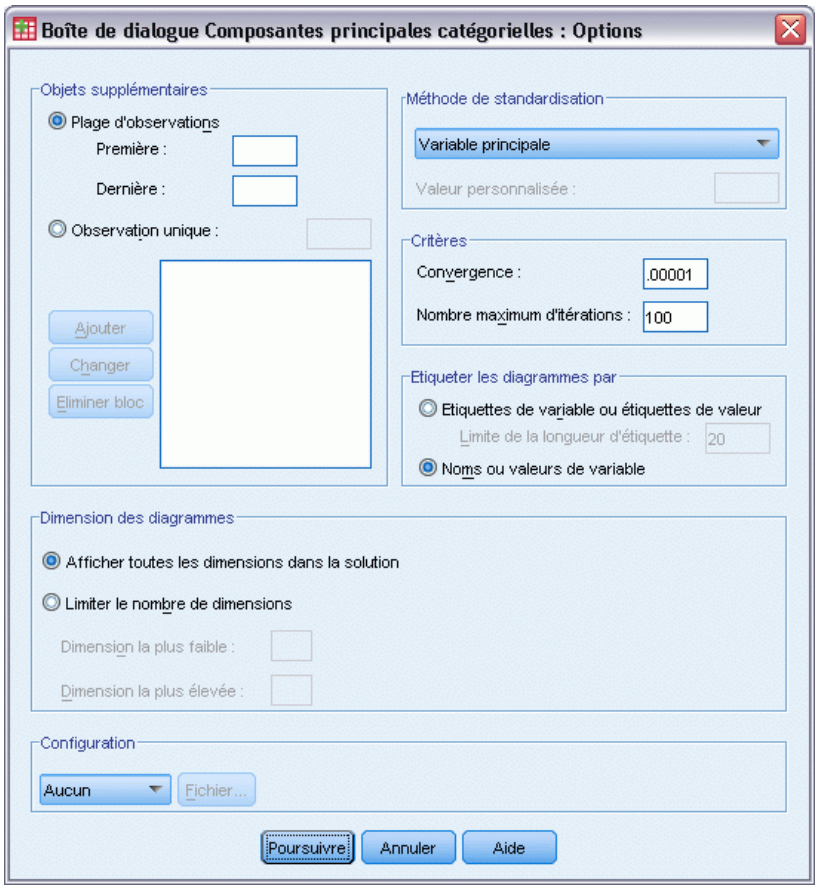
Stratégie : Choisissez d'exclure les valeurs manquantes (traitement passif), d'affecter des valeurs (traitement actif) ou d'exclure les objets contenant des valeurs manquantes (suppression des observations incomplètes).

- **Exclure les valeurs manquantes pour affectation de corrélations après quantification.** Les objets contenant des valeurs manquantes sur la variable sélectionnée ne contribuent pas à l'analyse de cette variable. Si un traitement passif est effectué sur toutes les variables, les objets dont les variables comportent des valeurs manquantes sont traités comme étant supplémentaires. Si les corrélations sont spécifiées dans la boîte de dialogue Résultat, les valeurs manquantes après analyse sont alors prises en compte avec la modalité la plus fréquente ou le mode de la variable pour les corrélations des variables d'origine. Pour corréler des variables codées de façon optimale, vous devez choisir une méthode d'imputation. Sélectionnez Mode pour remplacer les valeurs manquantes par le mode de la variable codée de façon optimale. Sélectionnez Modalité supplémentaire pour remplacer les valeurs manquantes par la valeur affectée à une modalité supplémentaire. Cela suppose que les objets contenant une valeur manquante pour cette variable sont considérés comme appartenant à la même modalité (supplémentaire).
- **Inclure les valeurs manquantes.** Des valeurs sont prises en compte pour les objets contenant des valeurs manquantes sur la variable sélectionnée. Vous pouvez choisir la méthode d'imputation : Sélectionnez Mode pour remplacer les valeurs manquantes par la modalité la plus fréquente. S'il existe plusieurs modes, utilisez celui dont l'indicateur de modalités est le plus petit. Sélectionnez Modalité supplémentaire pour remplacer les valeurs manquantes par la valeur affectée à une modalité supplémentaire. Cela suppose que les objets contenant une valeur manquante pour cette variable sont considérés comme appartenant à la même modalité (supplémentaire).
- **Exclure les objets contenant des valeurs manquantes sur cette variable.** Les objets contenant des valeurs manquantes dans la variable sélectionnée sont retirés de l'analyse. Cette option n'est pas disponible pour les variables supplémentaires.

Composantes principales qualitatives : Options

La boîte de dialogue Options vous permet de sélectionner la configuration initiale, de spécifier les itérations et les critères de convergence, de sélectionner une méthode de standardisation, de sélectionner une méthode d'étiquetage des diagrammes et, enfin, de spécifier des objets supplémentaires.

Figure 3-6
Options



Objets supplémentaires : Indiquez le numéro d'observation de l'objet, ou les premier et dernier numéros d'observation d'une plage d'objets que vous souhaitez définir comme objet supplémentaire, puis cliquez sur Ajouter. Poursuivez jusqu'à ce que vous ayez indiqué tous les objets supplémentaires. Si un objet est spécifié comme supplémentaire, alors les pondérations d'observation est ignorée pour cet objet.

Méthode de standardisation : Vous pouvez spécifier l'une des cinq options de standardisation des coordonnées des objets et des variables. Une seule méthode de standardisation peut être utilisée dans une analyse donnée.

- **Variable principale :** Cette option optimise l'association entre les variables. Les coordonnées des variables dans l'espace objet correspondent aux corrélations entre composants et variables initiales (corrélations comportant des composantes principales telles que des dimensions et des coordonnées d'objets). Cela est utile si vous êtes avant tout intéressé par les corrélations entre variables.
- **Objet principal :** Cette option optimise les distances entre les objets. Cela est utile si vous êtes avant tout intéressé par les différences ou similitudes entre objets.
- **Symétrique :** Utilisez cette option de standardisation si vous êtes avant tout intéressé par la relation entre les objets et les variables.

- **Indépendant :** Utilisez cette option de standardisation si vous souhaitez examiner les distances entre les objets ainsi que les corrélations entre variables séparément.
- **Personnalisée :** Vous pouvez spécifier toute valeur réelle comprise dans l'intervalle [–1, 1]. Une valeur de 1 correspond à la méthode Objet principal, une valeur de 0 correspond à la méthode Symétrique, et une valeur de –1 à la méthode Variable principale. En spécifiant une valeur comprise entre –1 et 1, la valeur propre peut comprendre à la fois les objets et les variables. Cette méthode est utile pour effectuer des diagrammes doubles ou triples.

Critères : Vous pouvez spécifier le nombre maximum d'itérations que la procédure peut prendre en charge dans ses calculs. Vous avez également la possibilité de sélectionner une valeur de critère de convergence. L'algorithme interrompt son itération dès que la différence du total ajusté entre les deux dernières itérations est inférieur à la valeur de la convergence ou dès que le nombre maximum d'itérations est atteint.

Etiqueter les diagrammes par : Vous permet de préciser si les étiquettes de variable et de valeurs ou les noms ou valeurs de variables sont utilisés dans les diagrammes. Vous pouvez également spécifier une longueur maximale pour les étiquettes.

Dimensions du diagramme. Permet de contrôler les dimensions contenues dans le résultat.

- **Afficher toutes les dimensions dans la solution.** Toutes les dimensions de la solution apparaissent dans une matrice de diagramme de dispersion.
- **Limiter le nombre de dimensions.** Les dimensions affichées sont limitées à des paires de dimensions représentées. Si vous restreignez ces dimensions, vous devez sélectionner la plus petite et la plus grande à tracer. La plus petite dimension peut être comprise entre 1 et le nombre de dimensions contenues dans la solution moins 1. En outre, elle est représentée par rapport aux dimensions plus grandes. La valeur de dimension la plus élevée peut être comprise entre 2 et le nombre de dimensions contenues dans la solution. Par ailleurs, elle indique la plus grande dimension à utiliser pour le traçage des paires de dimensions. Cette spécification s'applique à l'ensemble des représentations multidimensionnelles demandées.

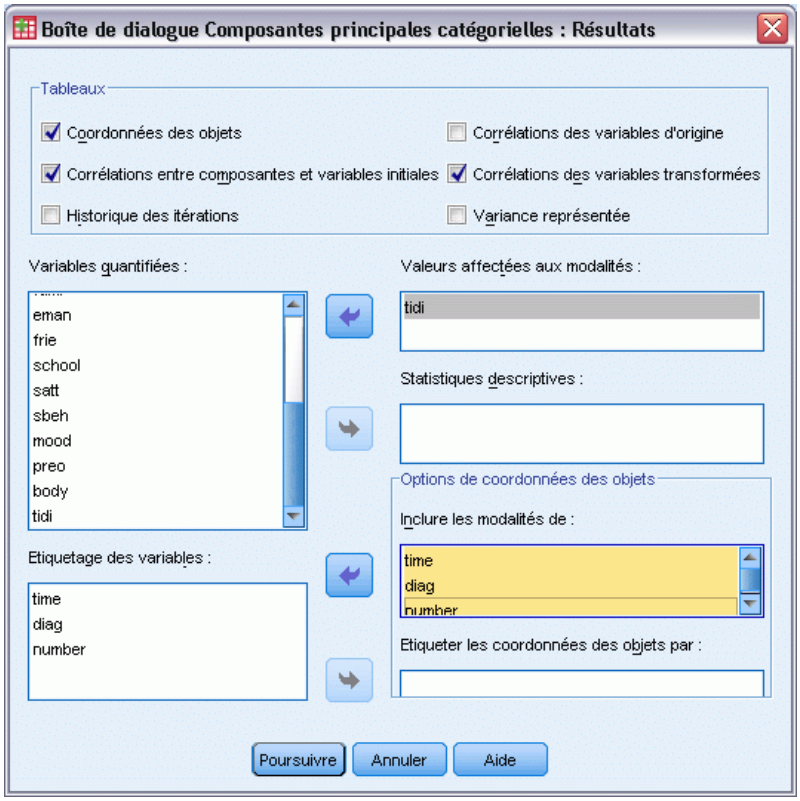
Configuration : Vous pouvez lire les données d'un fichier contenant les coordonnées de la configuration. La première variable du fichier doit contenir les coordonnées de la première dimension, la deuxième variable, celles de la deuxième dimension, et ainsi de suite.

- **Initiale :** La configuration du fichier spécifié sera utilisée comme point de départ de l'analyse.
- **Fixe :** La configuration du fichier spécifié sera utilisée pour ajuster les variables. Les variables ainsi ajustées doivent être sélectionnées comme des variables d'analyse, mais la configuration étant fixe, elles doivent être considérées comme des variables supplémentaires (il est donc inutile de les sélectionner comme telles).

Composantes principales qualitatives : Résultat

La boîte de dialogue Résultat vous permet de produire des tableaux affichant les coordonnées des objets, les corrélations entre composantes et variables initiales, un historique des itérations, les corrélations des variables d'origine et transformées, la variance représentée par variable et par dimension, les valeurs affectées aux modalités pour les variables sélectionnées et les statistiques descriptives pour les variables sélectionnées.

Figure 3-7
Résultat



Coordonnées des objets : Affiche les coordonnées des objets avec les options suivantes :

- **Inclure les modalités de :** Présente les indicateurs de modalités des variables d'analyse sélectionnées.
- **Etiqueter les objets du diagramme par :** Vous pouvez sélectionner l'une des variables spécifiées dans la liste de variables d'étiquetage pour étiqueter les objets.

Corrélations entre composantes et variables initiales : Affiche les corrélations entre composants et variables initiales pour toutes les variables n'ayant pas reçu de niveau de codage nominal multiple.

Historique des itérations : Pour chaque itération, la variance représentée, la perte et l'augmentation de la variance représentée sont affichées.

Corrélations des variables d'origine : Affiche la matrice de corrélation des variables d'origine ainsi que les valeurs propres de cette matrice.

Corrélations des variables transformées : Affiche la matrice de corrélation des variables transformées (codées de façon optimale) ainsi que les valeurs propres de cette matrice.

Variance expliquée par : Affiche le nombre de variances représentées par les coordonnées du barycentre, les coordonnées vectorielles et le total (coordonnées du barycentre et vectorielles combinées) par variable et par dimension.

Valeurs affectées aux modalités : Indique les valeurs affectées aux modalités et les coordonnées pour chaque dimension de la ou des variables sélectionnées.

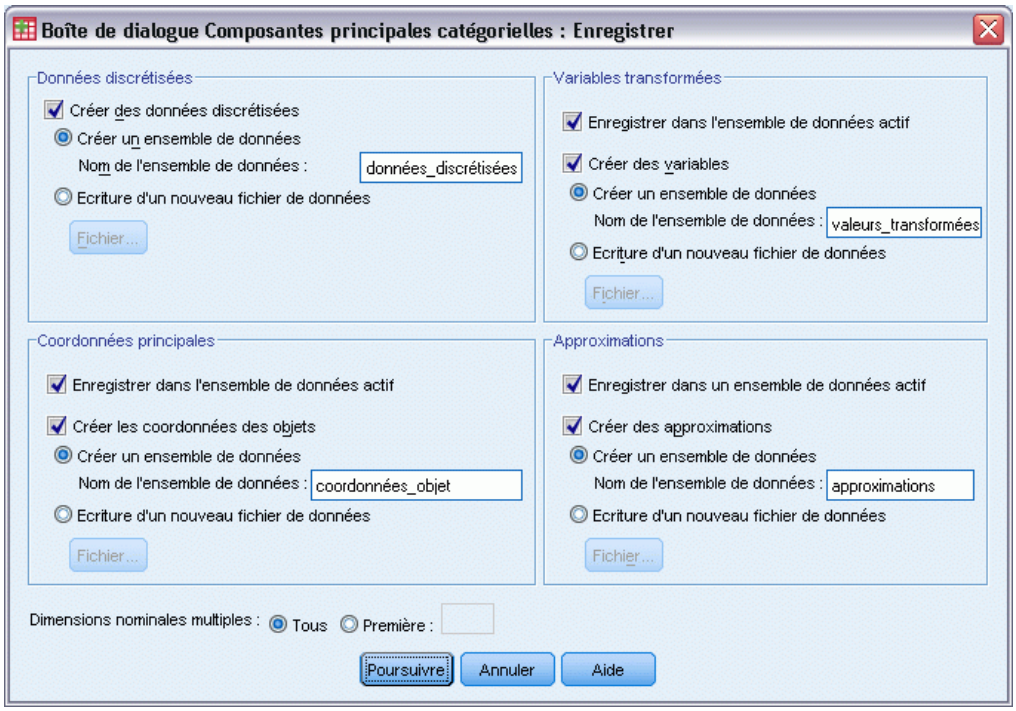
Statistiques descriptives : Affiche les effectifs, le nombre de valeurs manquantes et le mode de la ou des variables sélectionnées.

Composantes principales qualitatives : Enregistrer

La boîte de dialogue Enregistrer vous permet d'enregistrer les données discrétisées, les coordonnées des objets, les valeurs transformées et les approximations dans un fichier de données externe IBM® SPSS® Statistics ou un ensemble de données dans la session en cours. Vous pouvez également enregistrer les valeurs transformées, les coordonnées des objets et les approximations dans l'ensemble de données actif.

- Les ensembles de données sont disponibles lors de la session en cours mais ne sont pas disponibles lors des sessions suivantes, sauf si vous les enregistrez clairement comme fichiers de données. Les noms des ensembles de données doivent être conformes aux règles de dénomination de variables.
- Les noms de fichiers ou les noms de l'ensemble de données doivent être différents pour chaque type de données enregistrées.
- Si vous enregistrez les coordonnées des objets ou les valeurs transformées dans l'ensemble de données actif, vous pouvez indiquer le nombre des dimensions nominales multiples.

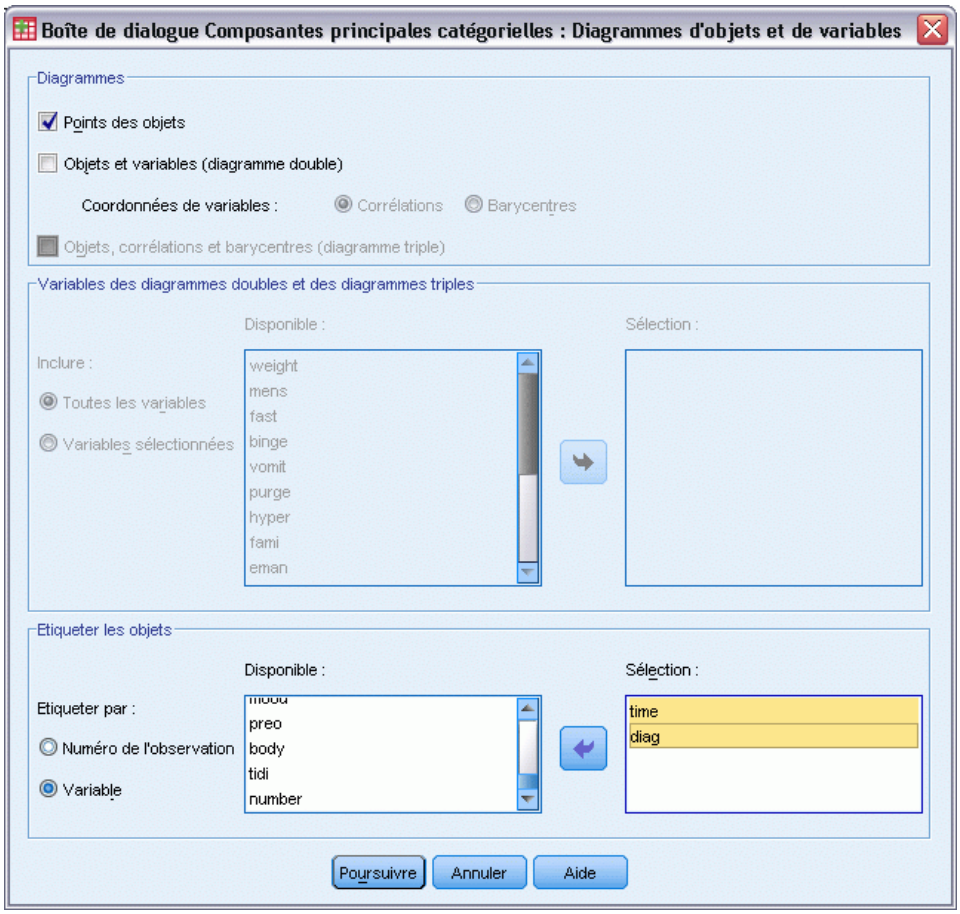
Figure 3-8
Enregistrer



Composantes principales qualitatives : Diagrammes d'objets et de variables

La boîte de dialogue Diagrammes d'objets et de variables vous permet de spécifier les types de diagrammes souhaités ainsi que les variables pour lesquels des diagrammes sont représentés.

Figure 3-9
Diagrammes d'objets et de variables



Points des objets. Un diagramme des points des objets s'affiche.

Objets et variables (biplot) : Les points des objets sont représentés avec les coordonnées de variables de votre choix : corrélations entre composants et variables initiales ou barycentres de variables.

Objets, corrélations et barycentres (triplot). Les points des objets sont représentés avec les barycentres des variables de niveau de codage nominal multiple et avec les corrélations entre composants et variables initiales des autres variables.

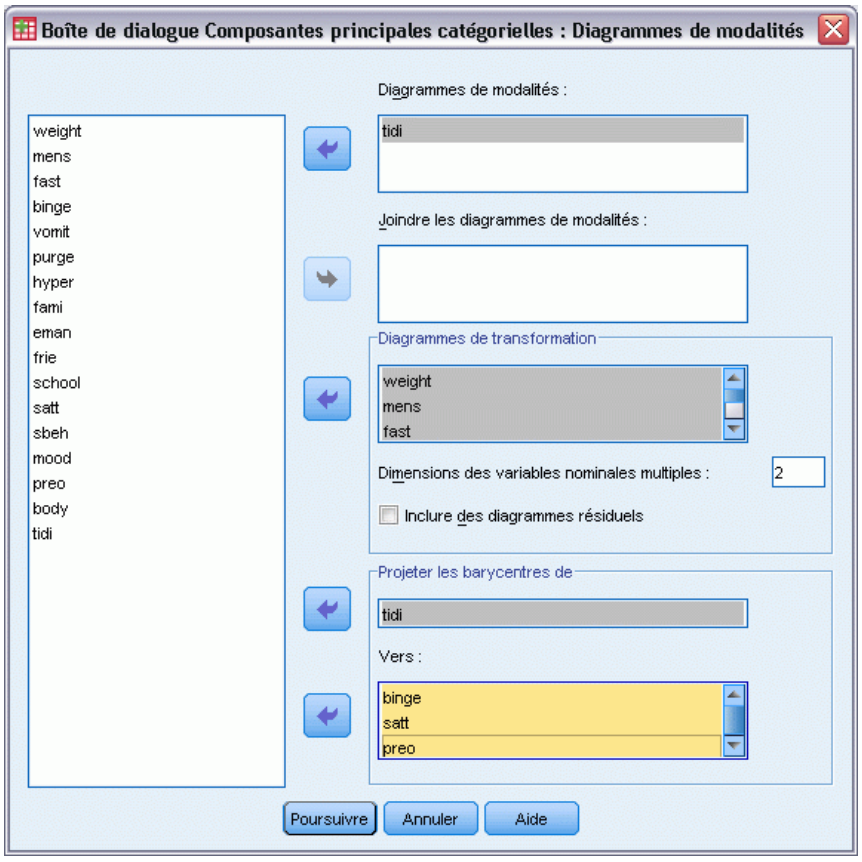
Variables des biplots et triplots : Vous pouvez choisir d'utiliser toutes les variables des diagrammes doubles et triples ou de sélectionner un sous-groupe.

Etiqueter objets : Vous pouvez choisir d'étiqueter des objets avec les modalités des variables sélectionnées (choisissez les valeurs des indicateurs de modalités ou les étiquettes de valeurs dans la boîte de dialogue Options) ou avec le nombre d'observations. Si vous avez sélectionné Variables, un seul diagramme est créé par variable.

Composantes principales qualitatives : Diagrammes de modalités

La boîte de dialogue Diagrammes de modalités vous permet de spécifier les types de diagrammes souhaités ainsi que les variables pour lesquelles des diagrammes seront représentés.

Figure 3-10
Boîte de dialogue Diagrammes de modalités



Diagrammes de modalités : Pour chaque variable sélectionnée, un diagramme des coordonnées du barycentre et vectorielles est représenté. Pour les variables contenant des niveaux de codage nominal multiple, les modalités figurent dans les barycentres des objets des modalités particulières. Pour les autres niveaux de codage, les modalités figurent dans un vecteur passant par l'origine.

Joindre les diagrammes de modalités : Il s'agit d'un diagramme simple représentant les coordonnées du barycentre et les coordonnées vectorielles de chaque variable sélectionnée.

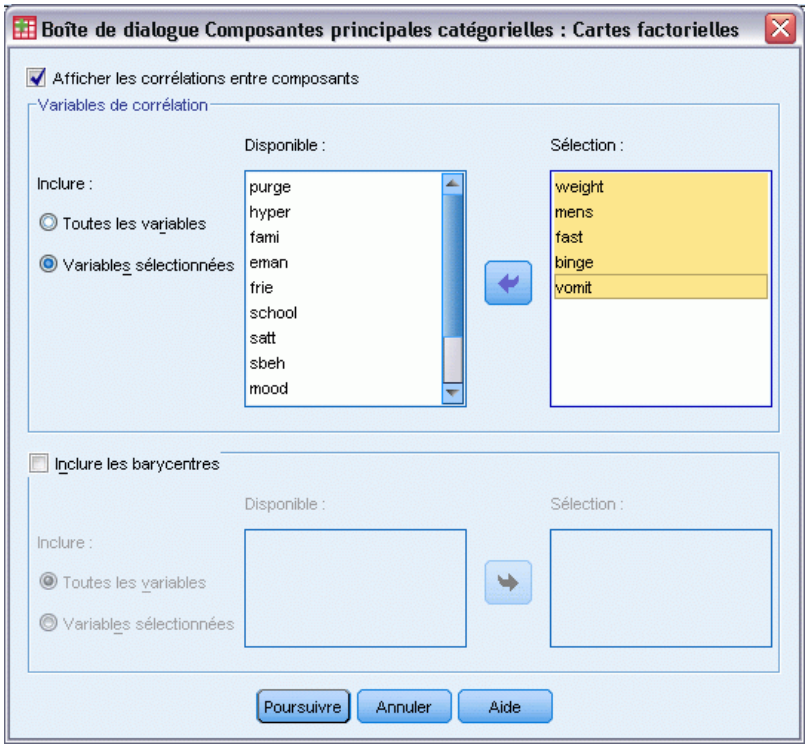
Diagrammes de transformation : Affiche un diagramme des valeurs affectées aux modalités optimales contre les indicateurs de modalités. Vous pouvez spécifier le nombre de dimensions souhaité pour les variables contenant des niveaux de codage nominal multiple. Un diagramme sera alors généré pour chaque dimension. Il vous est également possible de choisir d'afficher des diagrammes résiduels pour chaque variable sélectionnée.

Projeter les barycentres de : Vous pouvez choisir une variable et projeter ses barycentres sur les variables sélectionnées. Les variables comportant un niveau de codage nominal multiple ne peuvent pas être sélectionnées pour être projetées. Lorsque vous lancez ce diagramme, un tableau doté des coordonnées des barycentres projetés est également affiché.

Analyse des composantes principales qualitatives:Cartes factorielles

La boîte de dialogue Cartes factorielles permet de définir les variables à inclure dans le diagramme et d'indiquer si les barycentres seront également inclus.

Figure 3-11
Boîte de dialogue Cartes factorielles



Afficher les corrélations entre composants. Si cette option est sélectionnée, un diagramme des corrélations entre composants apparaît.

Variables de corrélation. Vous pouvez choisir d'utiliser toutes les variables d'un diagramme des corrélations entre composants ou de sélectionner un sous-groupe.

Inclure les barycentres. Les variables de niveau de codage nominal multiple ne possèdent pas de corrélation mais vous pouvez choisir d'inclure leurs barycentres dans le diagramme. Vous pouvez utiliser toutes les variables qualitatives multiples ou sélectionner un sous-groupe.

Fonctionnalités supplémentaires de la commande CATPCA

Vous pouvez personnaliser l'analyse des composantes principales qualitatives si vous collez vos sélections dans une fenêtre de syntaxe et modifiez la syntaxe de commande `CATPCA`. Le langage de syntaxe de commande vous permet aussi de :

- Spécifiez les noms de racine des variables transformées, les coordonnées des objets et les approximations lorsque vous les enregistrez dans l'ensemble de données actif (avec la sous-commande `SAVE`).
- Spécifier la longueur maximale pour les étiquettes de chaque diagramme séparément (avec la sous-commande `PLOT`).
- Spécifier une liste de variables distincte pour les diagrammes résiduels (avec la sous-commande `PLOT`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
4

Analyse canonique non linéaire (OVERALS)

L'analyse de corrélation canonique non linéaire correspond à l'analyse de corrélation canonique nominale avec codage optimal. Le but de cette procédure est de déterminer la similitude entre les groupes de variables qualitatives et les autres. Cette analyse est également connue sous l'acronyme OVERALS.

L'analyse de corrélation canonique standard est une extension de la régression multiple, dans laquelle le second groupe ne contient pas de variable de réponse unique, mais contient des variables de réponses multiples à la place. Elle sert à expliquer autant que possible la variance tirée des relations entre deux groupes de variables numériques dans un espace de petite dimension. Initialement, les variables de chaque groupe sont combinées de façon linéaire de sorte que les combinaisons comportent une corrélation maximale. Compte tenu de ces combinaisons, celles qui sont linéaires sont déterminées par celles qui ne le sont pas avec les combinaisons précédentes et par celles ayant la plus importante corrélation.

L'approche de codage optimal développe l'analyse standard de trois façons différentes. D'abord, OVERALS vous permet d'avoir plus de deux groupes de variables. Deuxièmement, les variables peuvent être codées soit de façon nominale, soit ordinale, soit numérique. En conséquence, des relations non linéaires entre les variables peuvent être analysées. Enfin, au lieu d'optimiser les corrélations entre les groupes de variable, ceux-ci sont comparés à un groupe de compromis inconnu défini par les coordonnées des objets.

Exemple : L'analyse de corrélation canonique nominale avec codage optimal peut être utilisée pour afficher graphiquement la relation entre un groupe de variables contenant une modalité d'emploi et les années d'étude, et un autre groupe de variables contenant la zone de résidence et le sexe. Il est possible que vous trouviez que les années d'étude et la zone de résidence établissent une différence plus importante que les autres variables. Mais, vous pouvez considérer que les années d'étude établissent une différence fondamentale sur la première dimension.

Diagrammes et statistiques : Effectifs, barycentres, historique des itérations, coordonnées des objets, valeurs affectées aux modalités, pondérations et corrélations entre composantes et variables initiales, ajustement unique et multiple, diagramme de coordonnées des objets, diagrammes de coordonnées des modalités, diagrammes de corrélations entre composantes et variables initiales, diagrammes de centres de classes et diagrammes de transformation.

Données : Utilisez des entiers pour coder les variables qualitatives (niveau de codage nominal ou ordinal). Pour réduire le nombre de résultats, utilisez des entiers consécutifs commençant par 1 pour coder les variables. Les variables codées à un niveau numérique ne doivent pas être recodées en entiers consécutifs. Pour réduire le nombre de résultats, pour chaque variable codée à un niveau numérique, soustrayez la plus petite valeur observée de chaque valeur et ajoutez-lui 1. Les valeurs fractionnelles sont tronquées après la décimale.

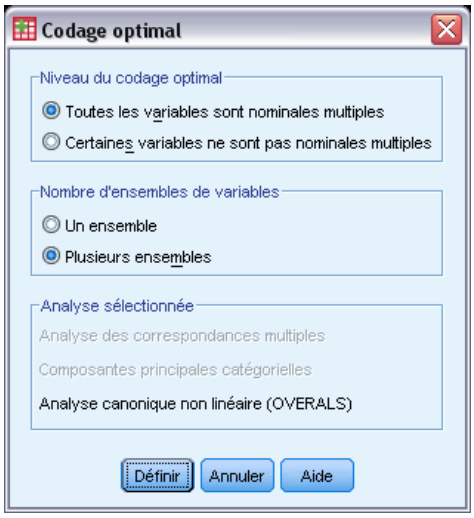
Hypothèses : Les variables peuvent être classées en deux groupes ou plus. Les variables dans l'analyse sont codées sous forme nominale multiple, nominale simple, ordinale ou numérique. Le nombre maximum de dimensions utilisées dans la procédure dépend du niveau de codage optimal des variables. Si toutes les variables sont indiquées comme étant ordinales, nominales simples ou numériques, le nombre maximum de dimensions est le plus petit des deux valeurs suivantes : le nombre d'observations moins 1 ou le nombre total des variables. Cependant, si seuls les deux groupes de variables sont définis, le nombre maximum de dimensions correspond au nombre de variables du plus petit groupe. Si plusieurs variables sont nominales multiples, le nombre maximum de dimensions correspond au nombre total de modalités nominales multiples plus le nombre de variables qualitatives non multiples et moins le nombre de variables qualitatives multiples. Par exemple, si l'analyse implique cinq variables et si l'une d'elles est nominale multiple avec quatre modalités, le nombre maximum de dimensions est (4 + 4 – 1) ou 7. Si vous spécifiez un nombre supérieur au maximum, la valeur maximale est alors utilisée.

Procédures apparentées : Si chaque groupe contient une variable, l'analyse de corrélation canonique non linéaire équivaut à l'analyse des composantes principales avec codage optimal. Si chacune de ces variables est nominale multiple, l'analyse correspond à l'analyse de correspondance multiple. Si deux groupes de variables sont impliqués et que l'un d'eux contient une seule variable, l'analyse correspond à une régression nominale avec codage optimal.

Obtenir une analyse de corrélation canonique non linéaire

▶ A partir des menus, sélectionnez :
Analyse > Réduction des dimensions > Codage optimal

Figure 4-1
Boîte de dialogue Niveau du codage optimal

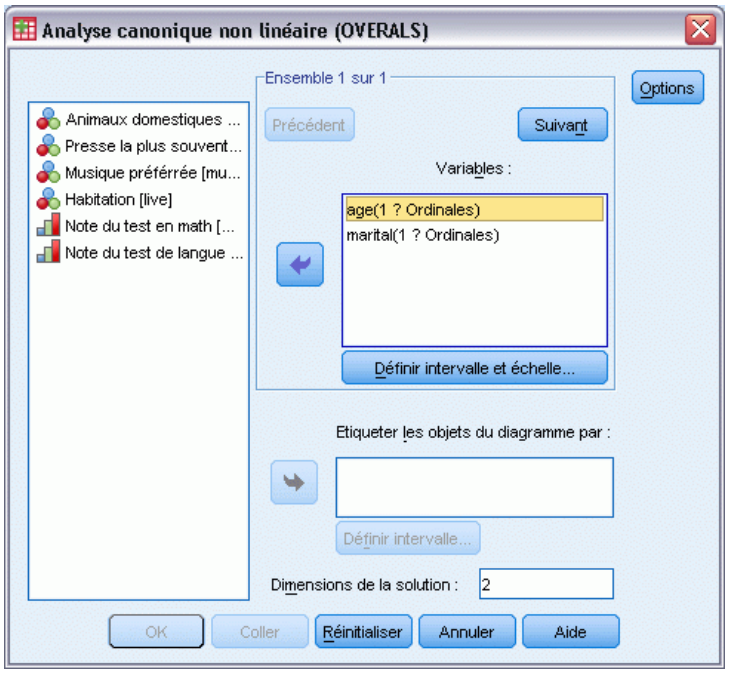


▶ Sélectionnez soit Toutes les variables qualitatives multiples, soit Certaines variables non nominales multiples.

▶ Sélectionnez Plusieurs groupes.

▶ Cliquez sur Définir.

Figure 4-2
Boîte de dialogue Analyse canonique non linéaire (OVERALS)



- Définissez au moins deux groupes de variables. Sélectionnez les variables que vous souhaitez inclure dans le premier groupe. Pour atteindre le dernier groupe, cliquez sur Suivant et sélectionnez les variables à inclure dans le second. Vous pouvez également, si vous le souhaitez, ajouter des groupes supplémentaires. Cliquez sur Précédent pour revenir au groupe de variables défini précédemment.

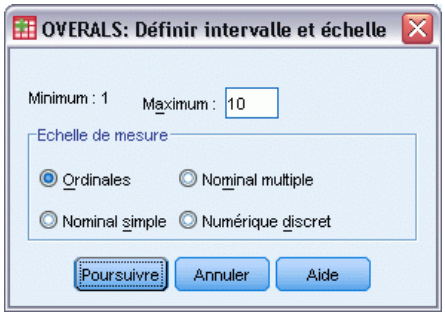
- Définissez la plage de valeurs et l'échelle de mesure (niveau de codage optimal) pour chaque variable sélectionnée.

- Cliquez sur OK.

- Eventuellement :
  - Sélectionner une ou plusieurs variables pour fournir les étiquettes de point aux diagrammes de coordonnées des objets. Chaque variable produit un diagramme séparé, avec les points étiquetés par ses valeurs. Vous devez définir une plage pour chacune de ces variables d'étiquettes de diagrammes. Lorsque vous utilisez la boîte de dialogue, une variable unique ne peut pas être utilisée à la fois dans l'analyse et sous forme de variable d'étiquette. Si vous souhaitez étiqueter un diagramme de coordonnées des objets avec une variable utilisée dans l'analyse, utilisez le sous-menu Calculer (disponible depuis le menu Transformer) pour créer une copie de cette variable. Utilisez la nouvelle variable pour étiqueter le diagramme. Il vous est également possible d'utiliser la syntaxe de commande.
  - Indiquez le nombre de dimensions souhaitées dans la solution. En général, choisissez autant de dimensions que nécessaires pour expliquer le maximum de la variation. Si l'analyse implique plusieurs dimensions, des diagrammes 3D des trois premières dimensions sont créés. D'autres dimensions peuvent également être affichées en éditant le diagramme.

Définir intervalle et échelle

Figure 4-3
Boîte de dialogue Définir intervalle et échelle



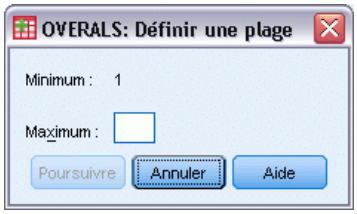
Vous devez définir une plage pour chaque variable. La valeur maximale indiquée doit être un nombre entier. Les valeurs des données fractionnelles sont tronquées dans l'analyse. Une valeur de modalité située en dehors de la plage spécifiée est ignorée dans l'analyse. Pour réduire le nombre de résultats, utilisez le sous-menu Recoder automatiquement (disponible depuis le menu Transformer) pour créer des modalités consécutives commençant par 1 pour des variables considérées comme nominales ou ordinales. Recoder en entiers consécutifs n'est pas recommandé pour les variables codées à un niveau numérique. Pour réduire le nombre de résultats pour les variables traitées comme numériques, pour chaque variable, soustrayez la valeur minimale de chaque valeur et ajoutez-lui 1.

Vous pouvez également sélectionner le codage à utiliser pour quantifier chaque variable.

- **Ordinal :** L'ordre des modalités de la variable observée est conservé dans la variable quantifiée.
- **Nominal simple :** Dans la variable quantifiée, les objets d'une même modalité reçoivent les mêmes coordonnées.
- **Nominal multiple :** Les quantifications peuvent différer pour chaque dimension.
- **Numérique discret :** Les modalités sont considérées comme triées et espacées régulièrement. Les différences entre le nombre des modalités et l'ordre de celles de la variable observée sont conservées dans la variable quantifiée.

Définir une plage

Figure 4-4
Boîte de dialogue Définir intervalle



Vous devez définir une plage pour chaque variable. La valeur maximale indiquée doit être un nombre entier. Les valeurs des données fractionnelles sont tronquées dans l'analyse. Une valeur de modalité située en dehors de la plage spécifiée est ignorée dans l'analyse. Pour réduire le

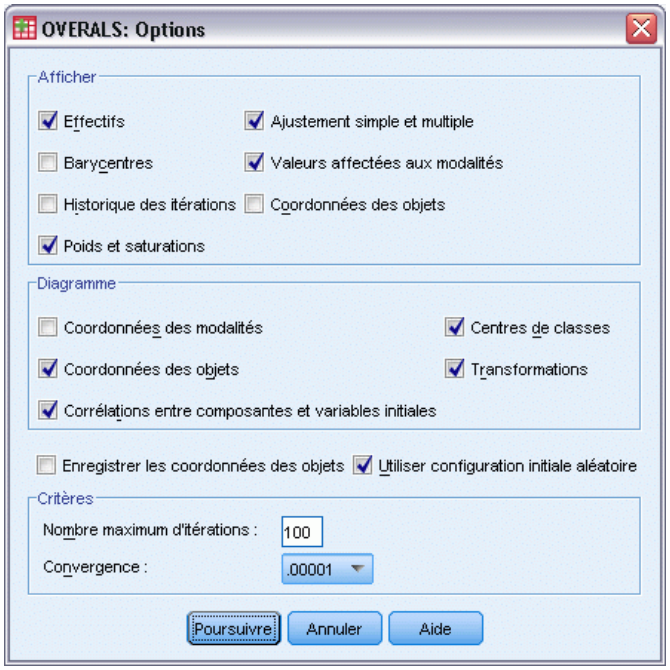
nombre de résultats, utilisez le sous-menu Recoder automatiquement (disponible depuis le menu Transformer) pour créer des modalités consécutives commençant par 1.

Vous devez également définir un intervalle pour chaque variable utilisée pour étiqueter les diagrammes de coordonnées des objets. Cependant, les étiquettes des modalités comportant des valeurs de données situées en dehors de la plage définie pour la variable apparaissent sur les diagrammes.

Analyse de corrélation canonique non linéaire – Options

La boîte de dialogue Options vous permet de sélectionner des statistiques et des diagrammes facultatifs, d'enregistrer les coordonnées des objets en tant que nouvelles variables dans l'ensemble de données actif, de spécifier les critères d'itérations et de convergence et d'indiquer une configuration initiale pour l'analyse.

Figure 4-5
Options



Afficher : Les statistiques disponibles incluent les effectifs marginaux, les barycentres, l'historique des itérations, les pondérations et corrélations entre composantes et variables initiales, les valeurs affectées aux modalités, les coordonnées des objets et l'ajustement unique et multiple.

- **Barycentres.** Quantifications des catégories, et moyennes projetées et réelles des coordonnées des objets (observations) inclus dans chaque ensemble pour ceux qui appartiennent à la même catégorie de la variable.
- **Poids et contributions (Corrélations entre composants et variables initiales).** Coefficients de régression dans chaque dimension pour chaque variable quantifiée d'un groupe. Les coordonnées des objets sont régressées sur les variables quantifiées et la projection de la

variable est quantifiée dans l'espace d'objet. Fournit une indication de la contribution que chaque variable apporte à la dimension dans chaque classe.

- **Ajustement simple et multiple.** Mesure la qualité de l'ajustement des coordonnées simple et multiple/quantifications de modalités par rapport aux objets.
- **Quantifications des modalités.** Affectation de coordonnées principales optimales aux modalités d'une variable.
- **Coordonnées des objets.** Quantification optimale affectée à un objet (observation) dans une dimension particulière.

Diagramme : Vous pouvez générer des diagrammes de coordonnées des modalités, de coordonnées des objets, de corrélations entre composantes et variables initiales, de centres de classes et de transformation.

Enregistrer les coordonnées des objets : Il est possible d'enregistrer les coordonnées des objets en tant que nouvelles variables dans l'ensemble de données actif. Ces coordonnées sont enregistrées en fonction du nombre de dimensions spécifiées dans la boîte de dialogue principale.

Utiliser configuration initiale aléatoire : Une configuration initiale aléatoire doit être utilisée si une partie ou la totalité des variables est nominale simple. Si cette case n'est pas cochée, une configuration initiale emboîtée est utilisée.

Critères : Vous pouvez spécifier le nombre maximum d'itérations que l'analyse canonique non linéaire peut prendre en charge dans ses calculs. Vous avez également la possibilité de sélectionner une valeur de critère de convergence. L'analyse interrompt son itération dès que la différence de l'ajustement total entre les deux dernières itérations est inférieure à la valeur de la convergence, ou dès que le nombre maximum d'itérations est atteint.

Fonctionnalités supplémentaires de la commande OVERALS

Vous pouvez personnaliser l'analyse canonique non linéaire en collant vos sélections dans une fenêtre de syntaxe et en modifiant la syntaxe de commande `OVERALS`. Le langage de syntaxe de commande vous permet aussi de :

- Spécifier les paires de dimensions à représenter, plutôt que représenter toutes les dimensions extraites (à l'aide du mot-clé `NDIM` de la sous-commande `PLOT`).
- Indiquer le nombre de caractères composant les étiquettes de valeurs utilisés pour étiqueter des points sur les diagrammes (avec la sous-commande `PLOT`).
- Désigner plus de cinq variables sous forme de variables d'étiquettes pour les diagrammes de coordonnées des objets (avec la sous-commande `PLOT`).
- Sélectionner les variables utilisées dans l'analyse en tant que variables d'étiquettes pour les diagrammes de coordonnées des objets (avec la sous-commande `PLOT`).
- Sélectionner les variables à fournir aux étiquettes de points pour le diagramme de coordonnées de quantification (avec la sous-commande `PLOT`).
- Indiquer le nombre d'observations à inclure dans l'analyse si vous ne souhaitez pas utiliser toutes les observations dans l'ensemble de données actif (avec la sous-commande `NOBSERVATIONS`).

- Spécifier les noms de racine des variables créées en enregistrant les coordonnées des objets (avec la sous-commande `SAVE`).
- Spécifier le nombre de dimensions à enregistrer, plutôt que de sauvegarder toutes les dimensions extraites (avec la sous-commande `SAVE`).
- Ecrire les valeurs affectées aux modalités dans un fichier de matrice (avec la sous-commande `MATRIX`).
- Produire des diagrammes à faible résolution pouvant être plus faciles à lire que des diagrammes à haute résolution (avec la sous-commande `SET`).
- Produire des diagrammes de barycentres et de transformations uniquement pour les variables spécifiées (avec la sous-commande `PLOT`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
5

Analyse des correspondances

L'une des fonctions de l'analyse des correspondances consiste à décrire les relations existant entre deux variables qualitatives dans un tableau de correspondances pour un espace comportant peu de dimensions, tout en décrivant simultanément les relations entre les modalités de chaque variable. Pour chacune des variables, les distances séparant les points des modalités d'un diagramme reflètent les relations existant entre ces modalités : plus les modalités sont similaires, plus elles sont proches les unes des autres. Les points de projection d'une variable du vecteur situés entre l'origine et l'un des points de modalité de l'autre variable décrivent les relations entre les deux variables.

Une analyse des tableaux de contingence implique fréquemment l'examen des profils des lignes et des colonnes ainsi qu'un test d'indépendance au moyen de la statistique Khi-deux. Toutefois, le nombre de profils peut s'avérer assez élevé et le test du Khi-deux n'indique pas la structure des dépendances. La procédure Tableaux croisés offre plusieurs mesures d'association et tests d'association mais ne permet pas de représenter graphiquement les relations existant entre les variables.

L'analyse factorielle constitue une technique standard de description des relations entre les variables d'un espace comportant peu de dimensions. Toutefois, l'analyse factorielle nécessite des données d'intervalle et le nombre d'observations doit être égal au nombre de variables multiplié par cinq. L'analyse des correspondances, en revanche, met en jeu des variables qualitatives et peut décrire les relations entre les modalités de chaque variable, ainsi que les relations entre les variables. En outre, l'analyse des correspondances permet d'analyser n'importe quel tableau de mesures de correspondances positives.

Exemple : L'analyse des correspondances peut être utilisée pour représenter graphiquement les relations existant entre la modalité socioprofessionnelle et le nombre de cigarettes consommées. Vous pourriez ainsi déterminer que la consommation de tabac diffère entre les jeunes cadres et les secrétaires, mais est similaire entre les secrétaires et les cadres supérieurs. Il vous serait également possible de déduire que les grands fumeurs sont principalement de jeunes cadres, alors que les fumeurs occasionnels sont généralement des secrétaires.

Diagrammes et statistiques : Mesures de correspondances, profils de lignes et de colonnes, valeurs singulières, scores de lignes et de colonnes, inertie, masse, statistiques de confiance des scores de lignes et de colonnes, statistiques de confiance des valeurs singulières, diagrammes de transformation, diagrammes de point de ligne, diagrammes de point de colonne et diagrammes doubles.

Données : Les variables qualitatives à analyser sont codées de façon nominale. Pour les données agrégées ou pour les mesures de correspondances autres que les effectifs, utilisez une variable de pondération présentant des valeurs de similarité positives. Pour les données de tableau, utilisez la syntaxe pour lire le tableau.

Hypothèses : Le nombre maximal de dimensions utilisé dans la procédure dépend du nombre de modalités de ligne et de colonne actives et du nombre de contraintes d'égalité. Si aucune contrainte d'égalité n'est appliquée et que toutes les modalités sont actives, le nombre de

dimensions maximal est inférieur de un au nombre de modalités de la variable présentant le plus petit nombre de modalités. Par exemple, si l'une des variables comporte cinq modalités et l'autre quatre, le nombre maximal de dimensions sera de trois. Les modalités supplémentaires ne sont pas actives. Par exemple, si une variable comporte cinq modalités, dont deux supplémentaires, et que l'autre variable possède quatre modalités, le nombre maximal de dimensions sera égal à deux. Tous les groupes de modalités faisant l'objet d'une contrainte d'égalité doivent être considérés comme une seule modalité. Ainsi, si une variable comporte cinq modalités, dont trois doivent être égales, vous devrez considérer cette variable comme ne possédant que trois modalités pour déterminer le nombre maximal de dimensions. Deux de ces modalités sont non contraintes, et la troisième correspond aux trois modalités contraintes. Si vous définissez un nombre de dimensions supérieur au nombre maximal autorisé, la valeur maximale sera appliquée par défaut.

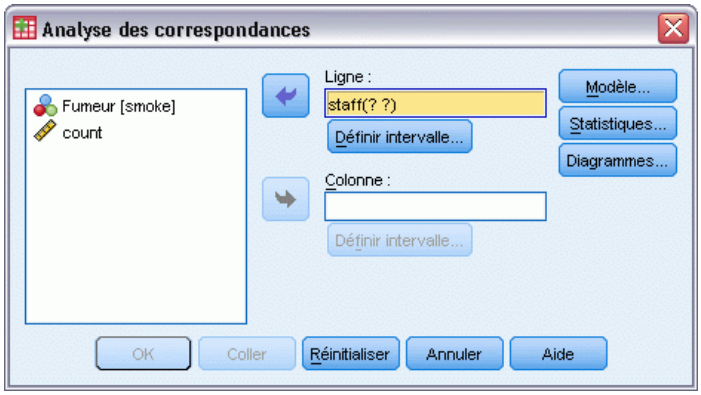
Procédures apparentées : Si vous travaillez avec plus de deux variables, procédez à une analyse de correspondance multiple. Si les variables doivent être codées de façon ordinale, utilisez l'analyse des composantes principales qualitatives.

Pour obtenir une analyse des correspondances

- A partir des menus, sélectionnez :

  Analyse > Réduction des dimensions > Analyse des correspondances...

Figure 5-1
Boîte de dialogue Analyse des correspondances

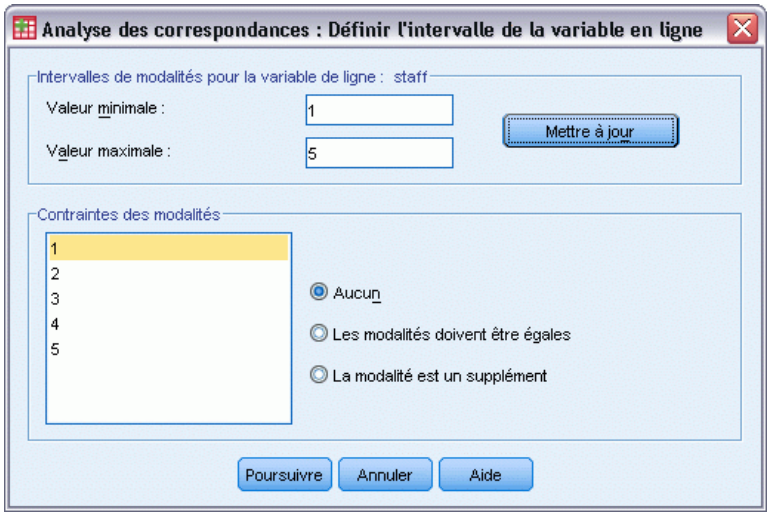


- Sélectionnez une variable de ligne.
- Sélectionnez une variable de colonne.
- Définir les plages des variables.
- Cliquez sur OK.

Définition de la plage de ligne dans l'analyse des correspondances

Vous devez définir une plage pour la variable en ligne. Les valeurs minimale et maximale spécifiées doivent être des nombres entiers. Les valeurs des données fractionnelles sont tronquées dans l'analyse. Une valeur de modalité située en dehors de la plage spécifiée est ignorée dans l'analyse.

Figure 5-2
Boîte de dialogue Définir l'intervalle de la variable en ligne



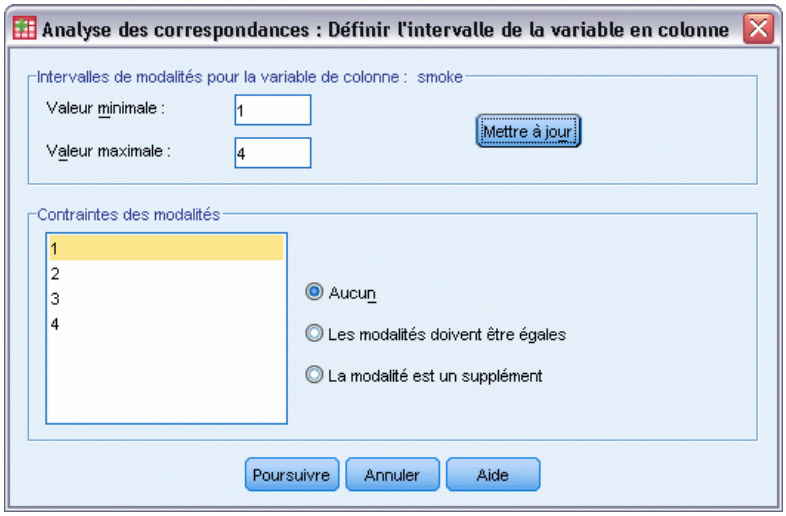
A l'origine, toutes les modalités sont non contraintes et actives. Vous pouvez par la suite contraindre certaines modalités de ligne à être égales à d'autres modalités de ligne, ou définir une modalité de ligne comme supplémentaire.

- **Les modalités doivent être égales :** Les modalités doivent présenter des scores identiques. Appliquez des contraintes d'égalité si l'ordre obtenu pour les modalités n'est pas souhaitable ou est contraire à l'intuition. Le nombre maximal de modalités de ligne pouvant faire l'objet d'une contrainte d'égalité correspond au nombre total de modalités de ligne actives moins 1. Pour imposer différentes contraintes d'égalité aux groupes de modalités, utilisez la syntaxe. Par exemple, utilisez la syntaxe pour contraindre les modalités 1 et 2 à être égales, puis pour appliquer la même contrainte aux modalités 3 et 4.
- **La modalité est un supplément :** Les modalités supplémentaires n'influencent pas l'analyse, mais sont représentées dans l'espace défini par les modalités actives. Les modalités supplémentaires ne jouent aucun rôle dans la définition des dimensions. Le nombre maximal de modalités de ligne supplémentaires correspond au nombre total de modalités de ligne moins 2.

Définition de la plage de colonne dans l'analyse des correspondances

Vous devez définir une plage pour la variable en colonne. Les valeurs minimale et maximale spécifiées doivent être des nombres entiers. Les valeurs des données fractionnelles sont tronquées dans l'analyse. Une valeur de modalité située en dehors de la plage spécifiée est ignorée dans l'analyse.

Figure 5-3
Boîte de dialogue Définir l'intervalle de la variable en colonne



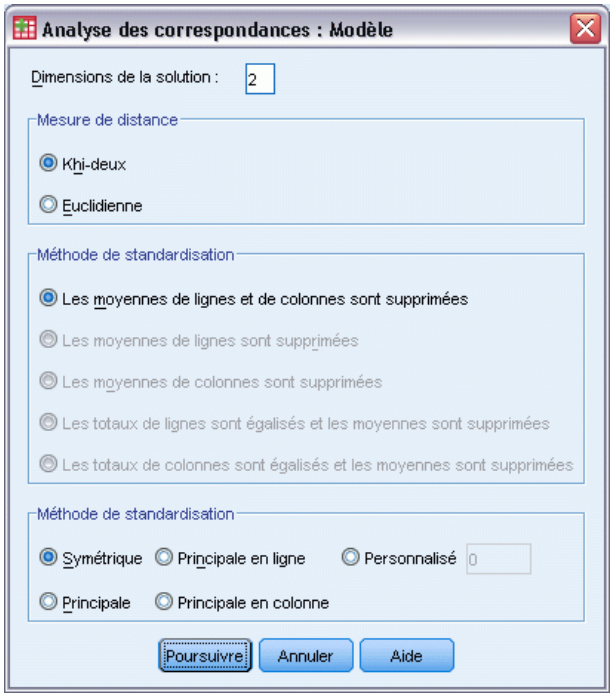
A l'origine, toutes les modalités sont non contraintes et actives. Vous pouvez par la suite contraindre certaines modalités de colonne à être égales à d'autres modalités de colonne ou définir une modalité de colonne comme supplémentaire.

- **Les modalités doivent être égales :** Les modalités doivent présenter des scores identiques. Appliquez des contraintes d'égalité si l'ordre obtenu pour les modalités n'est pas souhaitable ou est contraire à l'intuition. Le nombre maximal de modalités de colonne pouvant faire l'objet d'une contrainte d'égalité correspond au nombre total de modalités de colonne actives moins 1. Pour imposer différentes contraintes d'égalité aux groupes de modalités, utilisez la syntaxe. Par exemple, utilisez la syntaxe pour contraindre les modalités 1 et 2 à être égales, puis pour appliquer la même contrainte aux modalités 3 et 4.
- **La modalité est un supplément :** Les modalités supplémentaires n'influencent pas l'analyse, mais sont représentées dans l'espace défini par les modalités actives. Les modalités supplémentaires ne jouent aucun rôle dans la définition des dimensions. Le nombre maximal de modalités de colonne supplémentaires correspond au nombre total de modalités de colonne moins 2.

Modèle d'analyse des correspondances

La boîte de dialogue Modèle vous permet de définir le nombre de dimensions, la mesure de distance, la méthode de standardisation et la méthode de standardisation.

Figure 5-4
Boîte de dialogue Modèle



Dimensions de la solution : Spécifiez le nombre de dimensions. En général, choisissez autant de dimensions que nécessaires pour expliquer le maximum de la variation. Le nombre maximal de dimensions dépend du nombre de modalités actives utilisé dans l'analyse et des contraintes d'égalité. Le nombre maximal de dimensions est égal au plus petit d'entre ces deux nombres :

- Le nombre de modalités de ligne actives moins le nombre de modalités de ligne faisant l'objet d'une contrainte d'égalité, plus le nombre de groupes de modalités de ligne avec contrainte ;
- Le nombre de modalités de colonne actives moins le nombre de modalités de colonne faisant l'objet d'une contrainte d'égalité, plus le nombre de groupes de modalités de colonne avec contrainte.

Mesure de distance : Vous pouvez sélectionner la mesure de la distance entre les lignes et les colonnes du tableau des correspondances. Choisissez l'une des options suivantes :

- **Khi-deux :** Utilisez une distance de profil pondérée, la pondération correspondant à la masse des lignes ou des colonnes. Cette mesure est requise pour l'analyse des correspondances standard.
- **Euclidienne :** Utilisez la racine carrée de la somme des différences entre paires de lignes et paires de colonnes élevées au carré.

Méthode de standardisation : Choisissez l'une des options suivantes :

- **Moyennes de lignes et de colonnes éliminées :** Les lignes et les colonnes sont centrées. Cette méthode est requise pour l'analyse des correspondances standard.
- **Moyennes de lignes éliminées :** Seules les lignes sont centrées.
- **Moyennes de colonnes éliminées :** Seules les colonnes sont centrées.

- **Les Totaux de lignes sont égalisés et les moyennes éliminées :** Les marges des lignes sont égalisées avant que les lignes ne soient centrées.
- **Les totaux de colonnes sont égalisés et les moyennes éliminées :** Les marges des colonnes sont égalisées avant que les colonnes soient centrées.

Méthode de standardisation : Choisissez l'une des options suivantes :

- **Symétrique :** Pour chaque dimension, les scores des lignes représentent la moyenne pondérée des scores des colonnes, divisée par la valeur singulière correspondante ; les scores des colonnes représentent la moyenne pondérée des scores des lignes, divisée par la valeur singulière correspondante. Utilisez cette méthode si vous souhaitez examiner les différences ou les similitudes existant entre les modalités des deux variables.
- **Principale** : Les distances entre les points des lignes et des colonnes sont des approximations des distances du tableau des correspondances en fonction de la mesure de distance sélectionnée. Appliquez cette méthode si vous souhaitez examiner les différences existant entre les modalités de l'une ou des deux variables, plutôt que les différences entre ces deux variables.
- **Principale en ligne :** Les distances entre les points des lignes sont des approximations des distances du tableau des correspondances en fonction de la mesure de distance sélectionnée. Les scores des lignes correspondent à la moyenne pondérée des scores des colonnes. Utilisez cette méthode si vous souhaitez examiner les différences ou les similitudes existant entre les modalités de la variable de ligne.
- **Principale en colonne :** Les distances entre les points des colonnes sont des approximations des distances du tableau des correspondances en fonction de la mesure de distance sélectionnée. Les scores des colonnes correspondent à la moyenne pondérée des scores des lignes. Utilisez cette méthode si vous souhaitez examiner les différences ou les similitudes existant entre les modalités de la variable en colonne.
- **Personnalisée :** Vous devez définir une valeur comprise entre –1 et 1. La valeur –1 correspond à la méthode de standardisation principale en colonne. La valeur 1 correspond à la méthode de standardisation principale en ligne. La valeur 0 correspond à la méthode de standardisation symétrique. Toutes les autres valeurs dispersent l'inertie sur les scores des lignes et des colonnes à différents degrés. Cette méthode s'avère utile pour la création de diagrammes doubles adaptés à vos besoins.

Statistiques de l'analyse des correspondances

La boîte de dialogue Statistiques vous permet de définir les résultats numériques que vous souhaitez obtenir.

Figure 5-5
Boîte de dialogue Statistiques

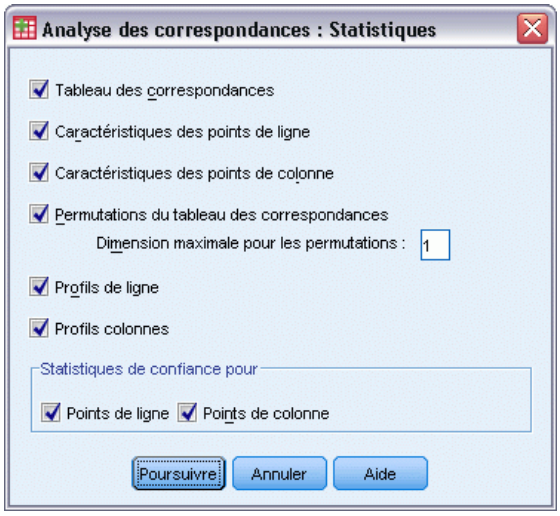


Tableau des correspondances : Tableau croisé des variables d'entrée incluant les totaux marginaux de ligne et de colonne.

Descriptives des points lignes : Pour chaque modalité de ligne, indique les scores, la masse, l'inertie, la contribution du point à l'inertie de la dimension ainsi que la contribution de la dimension à l'inertie du point.

Descriptives des points colonnes : Pour chaque modalité de colonne, indique les scores, la masse, l'inertie, la contribution du point à l'inertie de la dimension ainsi que la contribution de la dimension à l'inertie du point.

Profils lignes : Pour chaque modalité de ligne, indique la distribution entre les modalités de la variable en colonne.

Profils colonnes : Pour chaque modalité de colonne, indique la distribution entre les modalités de la variable en ligne.

Permutations du tableau des correspondances : Réorganisation du tableau des correspondances afin que les lignes et les colonnes apparaissent dans l'ordre croissant en fonction des scores de la première dimension. Une option vous permet de définir le nombre maximal de dimensions pour lequel vous souhaitez créer des tableaux permutés. Un tableau permuté sera alors généré pour chaque dimension comprise entre 1 et le nombre défini par vous.

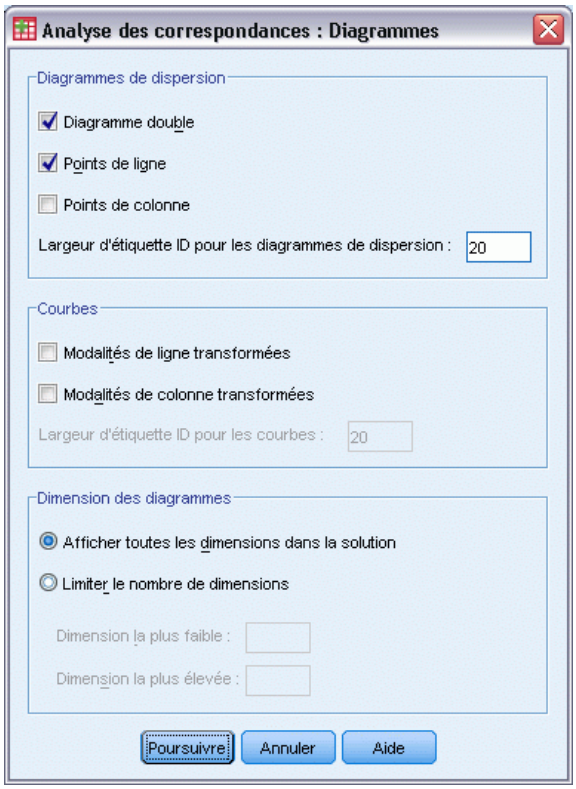
Statistiques de confiance pour points lignes : Ecart type et corrélations de tous les points de ligne non supplémentaires.

Statistiques de confiance pour points colonnes : Ecart type et corrélations de tous les points de colonne non supplémentaires.

Diagrammes de l'analyse des correspondances

La boîte de dialogue Diagrammes vous permet de définir les diagrammes que vous souhaitez créer.

Figure 5-6
Boîte de dialogue Diagrammes



Diagrammes de dispersion : Génère une matrice de tous les diagrammes présentant les dimensions par paire. Les diagrammes de dispersion disponibles sont les suivants :

- **Diagramme double :** Crée une matrice des diagrammes joints représentant les points des lignes et des colonnes. Si vous avez sélectionné la méthode de standardisation principale, l'option Diagramme double ne sera pas disponible.
- **Points lignes :** Crée une matrice des diagrammes représentant les points des lignes.
- **Points colonnes :** Crée une matrice des diagrammes représentant les points des colonnes.

Une option vous permet de définir le nombre de caractères composant les étiquettes de valeurs utilisées pour l'étiquetage des points. Cette valeur doit être un nombre entier positif inférieur ou égal à 20.

Courbes. Crée un diagramme pour chaque dimension de la variable sélectionnée. Les courbes disponibles sont les suivantes :

- **Modalités de lignes transformées :** Produit un diagramme représentant les valeurs des modalités de ligne d'origine par rapport aux scores des lignes qui leur correspondent.
- **Modalités de colonnes transformées :** Produit un diagramme représentant les valeurs des modalités de colonne d'origine par rapport aux scores des colonnes qui leur correspondent.

Une option vous permet de définir le nombre de caractères composant les étiquettes de valeurs utilisées pour l'étiquetage de l'axe des modalités. Cette valeur doit être un nombre entier positif inférieur ou égal à 20.

Dimensions du diagramme. Permet de contrôler les dimensions contenues dans le résultat.

- **Afficher toutes les dimensions dans la solution.** Toutes les dimensions de la solution apparaissent dans une matrice de diagramme de dispersion.
- **Limiter le nombre de dimensions.** Les dimensions affichées sont limitées à des paires de dimensions représentées. Si vous restreignez ces dimensions, vous devez sélectionner la plus petite et la plus grande à tracer. La plus petite dimension peut être comprise entre 1 et le nombre de dimensions contenues dans la solution moins 1. En outre, elle est représentée par rapport aux dimensions plus grandes. La valeur de dimension la plus élevée peut être comprise entre 2 et le nombre de dimensions contenues dans la solution. Par ailleurs, elle indique la plus grande dimension à utiliser pour le traçage des paires de dimensions. Cette spécification s'applique à l'ensemble des représentations multidimensionnelles demandées.

Fonctionnalités supplémentaires de la commande CORRESPONDENCE

Vous pouvez personnaliser votre analyse des correspondances en collant vos sélections dans une fenêtre de syntaxe, puis en modifiant la syntaxe de la commande `CORRESPONDENCE`. Le langage de syntaxe de commande vous permet aussi de :

- Indiquer les données des tableaux comme entrées au lieu d'utiliser les données d'observation (au moyen de la sous-commande `TABLE = ALL`).
- Spécifier le nombre de caractères composant les étiquettes de valeurs utilisées pour l'étiquetage des points de chaque type de matrice de diagramme de dispersion ou de diagramme double (au moyen de la sous-commande `PLOT`).
- Indiquer le nombre de caractères composant les étiquettes de valeurs utilisées pour l'étiquetage des points de chaque type de courbe (au moyen de la sous-commande `PLOT`).
- Créer une matrice des scores des lignes et des colonnes dans un fichier de données de matrice (avec la sous-commande `OUTFILE`).
- Créer une matrice des statistiques de confiance (variances et covariances) pour les valeurs singulières et les scores dans un fichier de données de matrice (avec la sous-commande `OUTFILE`).
- Appliquer une contrainte d'égalité à plusieurs groupes de modalités (au moyen de la sous-commande `EQUAL`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
6

Analyse de correspondance multiple

L'analyse de correspondance multiple quantifie les données (qualitatives) nominales en attribuant des valeurs numériques aux observations (objets) et aux modalités, pour que les objets faisant partie de la même modalité soient proches les uns des autres et ceux de différentes modalités, éloignés les uns des autres. Chaque objet se trouve aussi près que possible des points de modalité qui s'appliquent. Ainsi, les modalités divisent les objets en sous-groupes homogènes. Les variables sont considérées comme homogènes lorsqu'elles classent les objets des mêmes modalités dans les mêmes sous-groupes.

Exemple : L'analyse de correspondance multiple peut être utilisée pour afficher graphiquement la relation entre la modalité d'emploi, la classification des minorités et le sexe. Vous pouvez trouver que la classification par minorités et le sexe sont discriminant pour les personnes, mais que la modalité d'emploi ne l'est pas. Vous avez également la possibilité de constater que les modalités Latino et Afro-Américaines sont similaires les unes des autres.

Diagrammes et statistiques. Coordonnées des objets, mesures de discrimination, historique des itérations, corrélations des variables d'origine et des variables transformées, quantifications des modalités, statistiques descriptives, diagrammes de points des objets, diagrammes doubles, diagrammes de modalités, diagrammes de modalités joints, diagrammes de transformation et diagrammes de mesures de discrimination.

Données. Les variables chaîne sont toujours converties en nombres entiers positifs par ordre croissant alphanumérique. Les valeurs manquantes définies par l'utilisateur, les valeurs manquantes par défaut et les valeurs inférieures à 1 sont considérées comme manquantes ; vous pouvez donc recoder ou ajouter une constante aux variables contenant des valeurs inférieures à 1 pour les définir comme non manquantes.

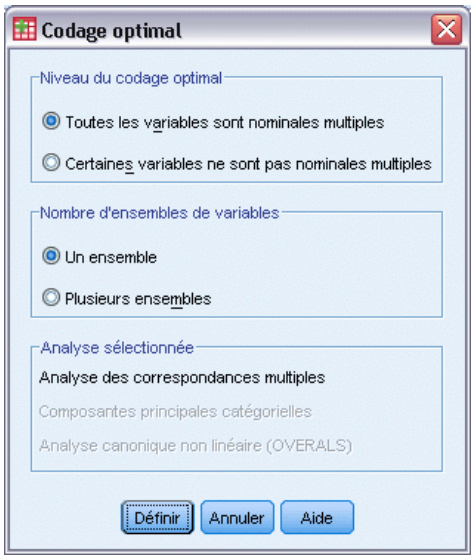
Hypothèses : Toutes les variables contiennent le niveau de codage nominal multiple. Les données doivent contenir au moins trois observations valides. L'analyse repose sur des données sous forme de nombres entiers positifs. L'option de discrétisation classe automatiquement une variable fractionnée en regroupant ses valeurs en modalités avec une distribution quasi normale et convertit automatiquement les valeurs des variables chaîne en nombre entiers positifs. Vous pouvez en outre, spécifier d'autres schémas de discrétisation.

Procédures apparentées : Pour deux variables, l'analyse de correspondance multiple est identique à l'analyse des correspondances. Si vous pensez que ces variables possèdent des propriétés ordinales ou numériques, vous devez utiliser l'analyse des composantes principales qualitatives. Si des groupes de variables sont intéressants, vous devez utiliser une analyse des corrélations canoniques (non linéaires).

Pour obtenir une analyse de correspondance multiple

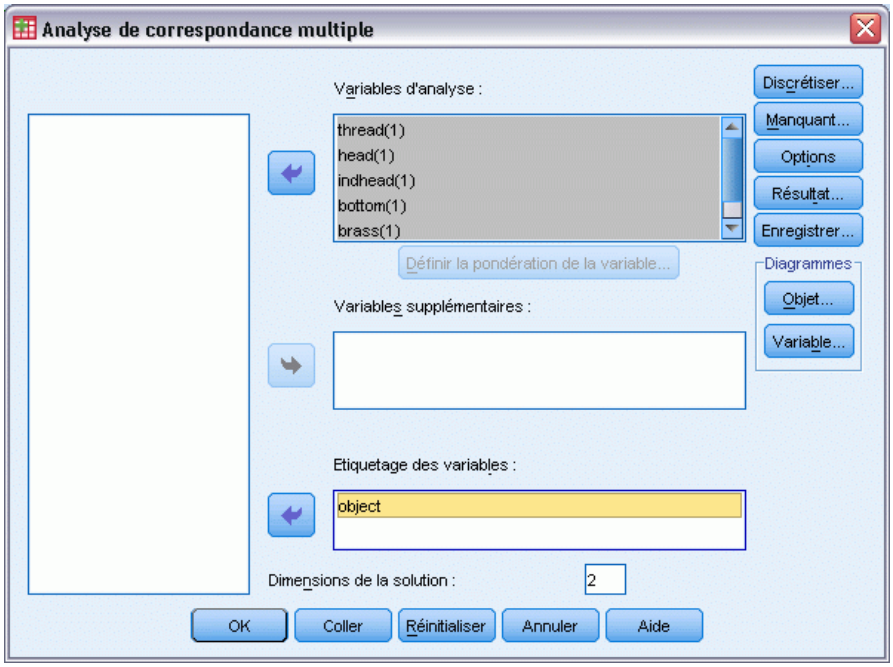
- A partir des menus, sélectionnez :
  Analyse > Réduction des dimensions > Codage optimal

Figure 6-1
Boîte de dialogue Niveau du codage optimal



- Sélectionnez Toutes les variables nominales multiples.
- Sélectionnez Un groupe.
- Cliquez sur Définir.

Figure 6-2
Boîte de dialogue Analyse des correspondances multiples



- Sélectionnez au moins deux variables d'analyse et spécifiez le nombre de dimensions de la solution.

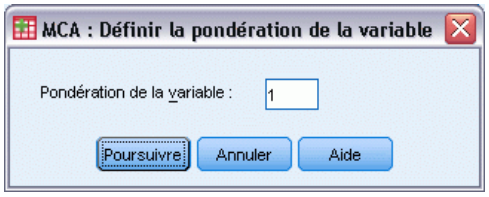
▶ Cliquez sur OK.

Vous pouvez peut-être spécifier des variables supplémentaires qui sont ajustées à la solution trouvée, ou des variables d'étiquettes pour les diagrammes.

Définition d'une pondération de variable dans une analyse de correspondance multiple

Vous pouvez définir la pondération pour les variables d'analyse.

Figure 6-3
Boîte de dialogue Définir la pondération de la variable

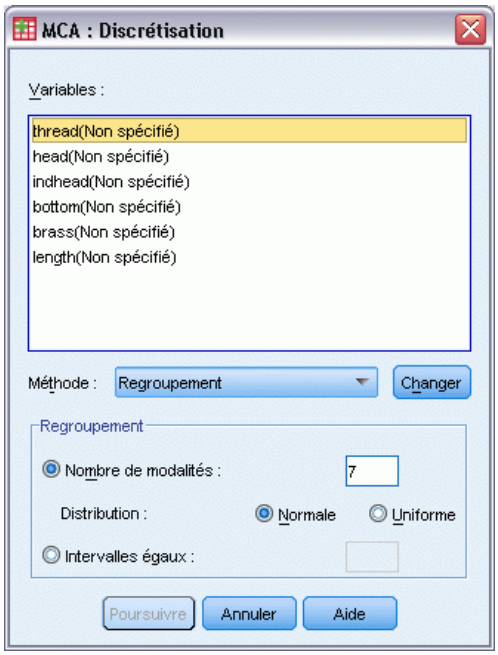


Pondération de la variable : Vous pouvez choisir une pondération pour chaque variable. La valeur spécifiée doit être un nombre entier positif. La valeur par défaut est 1.

Analyse des correspondances multiples : Discrétisation

La boîte de dialogue Discrétisation vous permet de choisir une méthode de recodage des variables. Les valeurs fractionnées sont regroupées en sept modalités (ou en nombre de valeurs distinctes de variables si le nombre est inférieur à sept) avec une distribution normale approximative, à moins qu'une autre configuration ne soit spécifiée. Les variables chaîne sont toujours converties en nombres entiers positifs en affectant des indicateurs de modalités selon l'ordre croissant alphanumérique. La discrétisation des variables chaîne s'applique à ces nombres entiers. Par défaut, d'autres variables sont laissées inutilisées. Les variables discrétisées sont ensuite utilisées dans l'analyse.

Figure 6-4
Discrétisation



Méthode : Choisissez entre Regroupement, Rang et Multiplier.

- **Regroupement :** Recodez en un nombre spécifié de modalités ou par intervalle.
- **Rang :** La variable est discrétisée via le classement des observations.
- **Multiplier :** Les valeurs courantes de la variable sont standardisées, multipliées par 10 et arrondies, et possèdent une constante ajoutée de sorte que la valeur discrétisée la plus faible soit égale à 1.

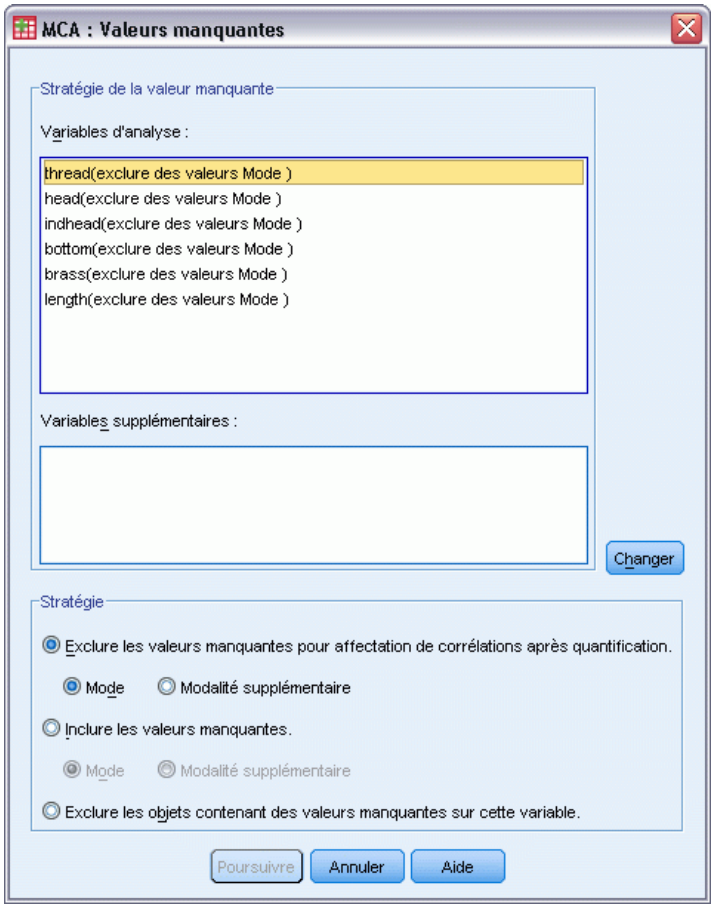
Regroupement : Les options suivantes sont disponibles lorsque vous discrétisez des variables par groupe :

- **Nombre de modalités :** Indiquez un nombre de modalités et définissez si les valeurs de la variable doivent faire l'objet d'une distribution approximativement gaussienne ou uniforme entre ces modalités.
- **Intervalles égaux :** Les variables sont recodées en modalités définies par ces intervalles de taille égale. N'oubliez pas de spécifier la longueur des intervalles.

Analyse des correspondances multiples : Valeurs manquantes

La boîte de dialogue Valeurs manquantes vous permet de choisir la stratégie de gestion des valeurs manquantes pour les variables de l'analyse et supplémentaires.

Figure 6-5
Boîte de dialogue Valeurs manquantes



Stratégie de la valeur manquante. Choisissez d'exclure les valeurs manquantes (traitement passif), d'affecter des valeurs (traitement actif) ou d'exclure les objets contenant des valeurs manquantes (suppression des observations incomplètes).

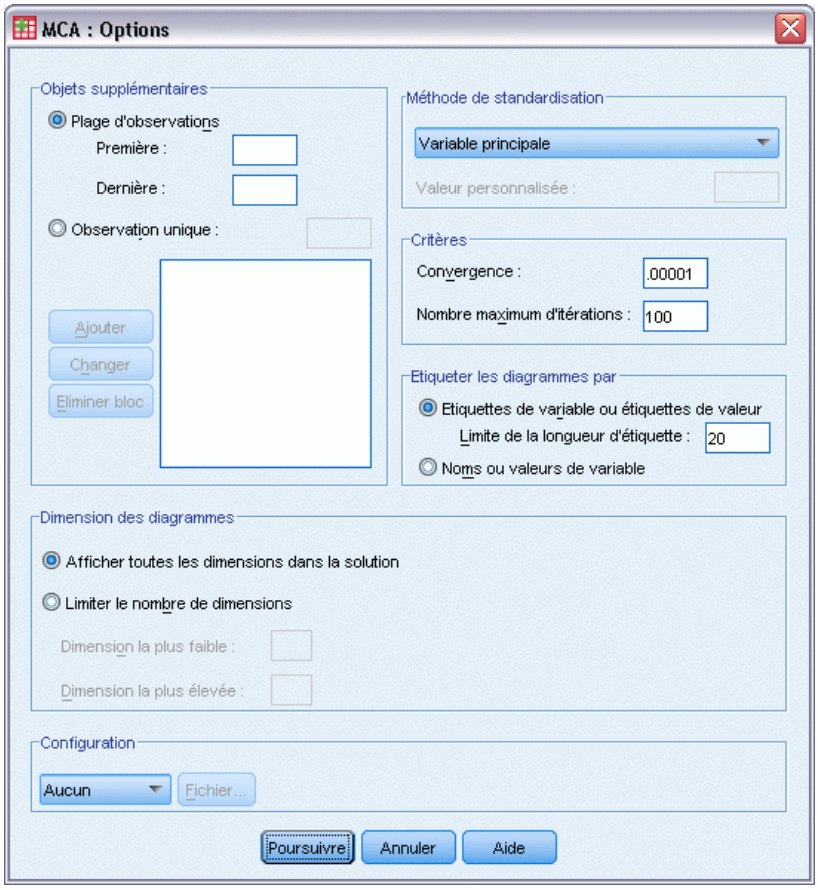
- **Exclure les valeurs manquantes pour affectation de corrélations après quantification.** Les objets contenant des valeurs manquantes sur la variable sélectionnée ne contribuent pas à l'analyse de cette variable. Si un traitement passif est effectué sur toutes les variables, les objets dont les variables comportent des valeurs manquantes sont traités comme étant supplémentaires. Si les corrélations sont spécifiées dans la boîte de dialogue Résultat, les valeurs manquantes après analyse sont alors prises en compte avec la modalité la plus fréquente ou le mode de la variable pour les corrélations des variables d'origine. Pour corréler des variables codées de façon optimale, vous devez choisir une méthode d'imputation. Sélectionnez Mode pour remplacer les valeurs manquantes par le mode de la variable codée de façon optimale. Sélectionnez Modalité supplémentaire pour remplacer les valeurs manquantes par la valeur affectée à une modalité supplémentaire. Cela suppose que les objets contenant une valeur manquante pour cette variable sont considérés comme appartenant à la même modalité (supplémentaire).

- **Inclure les valeurs manquantes.** Des valeurs sont prises en compte pour les objets contenant des valeurs manquantes sur la variable sélectionnée. Vous pouvez choisir la méthode de calcul : Sélectionnez Mode pour remplacer les valeurs manquantes par la modalité la plus fréquente. S'il existe plusieurs modes, utilisez celui dont l'indicateur de modalités est le plus petit. Sélectionnez Modalité supplémentaire pour remplacer les valeurs manquantes par la valeur affectée à une modalité supplémentaire. Cela suppose que les objets contenant une valeur manquante pour cette variable sont considérés comme appartenant à la même modalité (supplémentaire).
- **Exclure les objets contenant des valeurs manquantes sur cette variable.** Les objets contenant des valeurs manquantes dans la variable sélectionnée sont retirés de l'analyse. Cette option n'est pas disponible pour les variables supplémentaires.

Analyse des correspondances multiples : Options

La boîte de dialogue Options vous permet de sélectionner la configuration initiale, de spécifier les itérations et les critères de convergence, de sélectionner une méthode de standardisation, de sélectionner une méthode d'étiquetage des diagrammes et, enfin, de spécifier des objets supplémentaires.

Figure 6-6
Options



Objets supplémentaires : Indiquez le numéro d'observation de l'objet (ou les premier et dernier numéros d'observation d'une plage d'objets) que vous souhaitez définir comme objet supplémentaire, puis cliquez sur Ajouter. Poursuivez jusqu'à ce que vous ayez indiqué tous les objets supplémentaires. Si un objet est spécifié comme supplémentaire, alors les pondérations d'observation est ignorée pour cet objet.

Méthode de standardisation : Vous pouvez spécifier l'une des cinq options de standardisation des coordonnées des objets et des variables. Une seule méthode de standardisation peut être utilisée dans une analyse donnée.

- **Variable principale :** Cette option optimise l'association entre les variables. Les coordonnées des variables dans l'espace objet correspondent aux corrélations entre composants et variables initiales (corrélations comportant des composantes principales telles que des dimensions et des coordonnées d'objets). Cela est utile si vous êtes avant tout intéressé par les corrélations entre variables.
- **Objet principal :** Cette option optimise les distances entre les objets. Cela est utile si vous êtes avant tout intéressé par les différences ou similitudes entre objets.
- **Symétrique :** Utilisez cette option de standardisation si vous êtes avant tout intéressé par la relation entre les objets et les variables.
- **Indépendant :** Utilisez cette option de standardisation si vous souhaitez examiner les distances entre les objets ainsi que les corrélations entre variables séparément.
- **Personnalisée :** Vous pouvez spécifier toute valeur réelle comprise dans l'intervalle [–1, 1]. Une valeur de 1 correspond à la méthode Objet principal, une valeur de 0 correspond à la méthode Symétrique, et une valeur de –1 à la méthode Variable principale. En spécifiant une valeur comprise entre –1 et 1, la valeur propre peut comprendre à la fois les objets et les variables. Cette méthode est utile pour effectuer des diagrammes doubles ou triples.

Critères : Vous pouvez spécifier le nombre maximum d'itérations que la procédure peut prendre en charge dans ses calculs. Vous avez également la possibilité de sélectionner une valeur de critère de convergence. L'algorithme interrompt son itération dès que la différence du total ajusté entre les deux dernières itérations est inférieur à la valeur de la convergence ou dès que le nombre maximum d'itérations est atteint.

Etiqueter les diagrammes par : Vous permet de préciser si les étiquettes de variable et de valeurs ou les noms ou valeurs de variables sont utilisés dans les diagrammes. Vous pouvez également spécifier une longueur maximale pour les étiquettes.

Dimensions du diagramme. Permet de contrôler les dimensions contenues dans le résultat.

- **Afficher toutes les dimensions dans la solution.** Toutes les dimensions de la solution apparaissent dans une matrice de diagramme de dispersion.
- **Limiter le nombre de dimensions.** Les dimensions affichées sont limitées à des paires de dimensions représentées. Si vous restreignez ces dimensions, vous devez sélectionner la plus petite et la plus grande à tracer. La plus petite dimension peut être comprise entre 1 et le nombre de dimensions contenues dans la solution moins 1. En outre, elle est représentée par rapport aux dimensions plus grandes. La valeur de dimension la plus élevée peut être comprise entre 2 et le nombre de dimensions contenues dans la solution. Par ailleurs, elle indique la plus grande dimension à utiliser pour le traçage des paires de dimensions. Cette spécification s'applique à l'ensemble des représentations multidimensionnelles demandées.

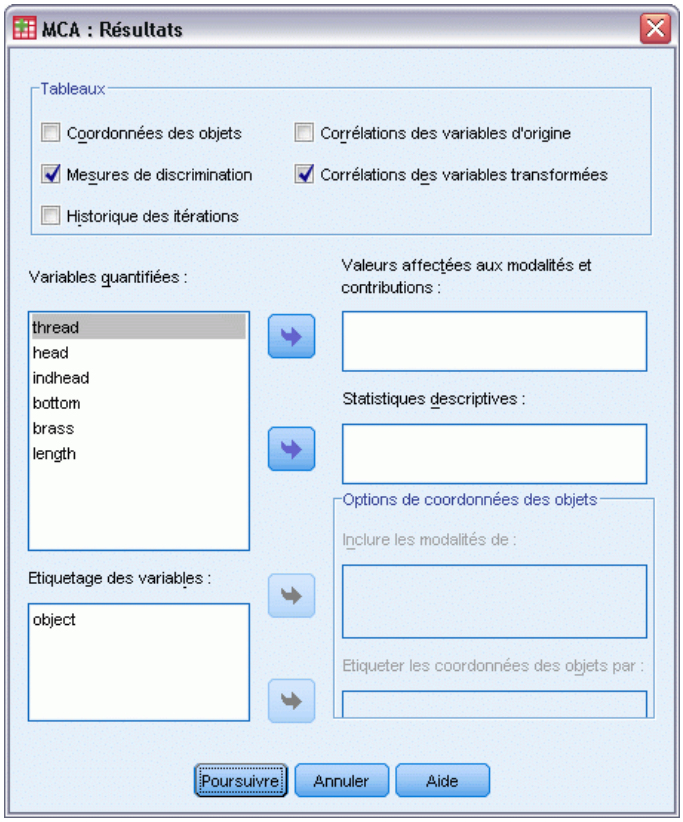
Configuration : Vous pouvez lire les données d'un fichier contenant les coordonnées de la configuration. La première variable du fichier doit contenir les coordonnées de la première dimension, la deuxième variable, celles de la deuxième dimension, et ainsi de suite.

- **Initiale :** La configuration du fichier spécifié sera utilisée comme point de départ de l'analyse.
- **Fixe :** La configuration du fichier spécifié sera utilisée pour ajuster les variables. Les variables ainsi ajustées doivent être sélectionnées comme des variables d'analyse, mais la configuration étant fixe, elles doivent être considérées comme des variables supplémentaires (il est donc inutile de les sélectionner comme telles).

Analyse des correspondances multiples : Résultats

La boîte de dialogue Résultat vous permet de créer des tableaux pour les coordonnées des objets, les mesures de discrimination, l'historique des itérations, les corrélations des variables d'origine et des variables transformées, ainsi que les quantifications des modalités et statistiques descriptives des variables sélectionnées.

Figure 6-7
Résultat



Coordonnées des objets : Affiche les coordonnées des objets, y compris la masse, l'inertie et les contributions, ainsi que les options suivantes :

- **Inclure les modalités de :** Présente les indicateurs de modalités des variables d'analyse sélectionnées.
- **Etiqueter les objets du diagramme par :** Vous pouvez sélectionner l'une des variables spécifiées dans la liste de variables d'étiquetage pour étiqueter les objets.

Mesures de discrimination. Affiche les mesures de discrimination par variable et par dimension.

Historique des itérations : Pour chaque itération, la variance représentée, la perte et l'augmentation de la variance représentée sont affichées.

Corrélations des variables d'origine : Affiche la matrice de corrélation des variables d'origine ainsi que les valeurs propres de cette matrice.

Corrélations des variables transformées : Affiche la matrice de corrélation des variables transformées (codées de façon optimale) ainsi que les valeurs propres de cette matrice.

Valeurs affectées aux modalités et contributions. Indique les valeurs affectées aux modalités (coordonnées), y compris la masse, l'inertie et les contributions pour chaque dimension de la ou des variables sélectionnées.

Remarque : les coordonnées et les contributions (dont la masse et l'inertie) sont affichées dans des strates distinctes des résultats du tableau pivotant, les coordonnées étant affichées par défaut. Pour afficher les contributions, double-cliquez sur le tableau et sélectionnez Contributions dans la liste déroulante Strate.

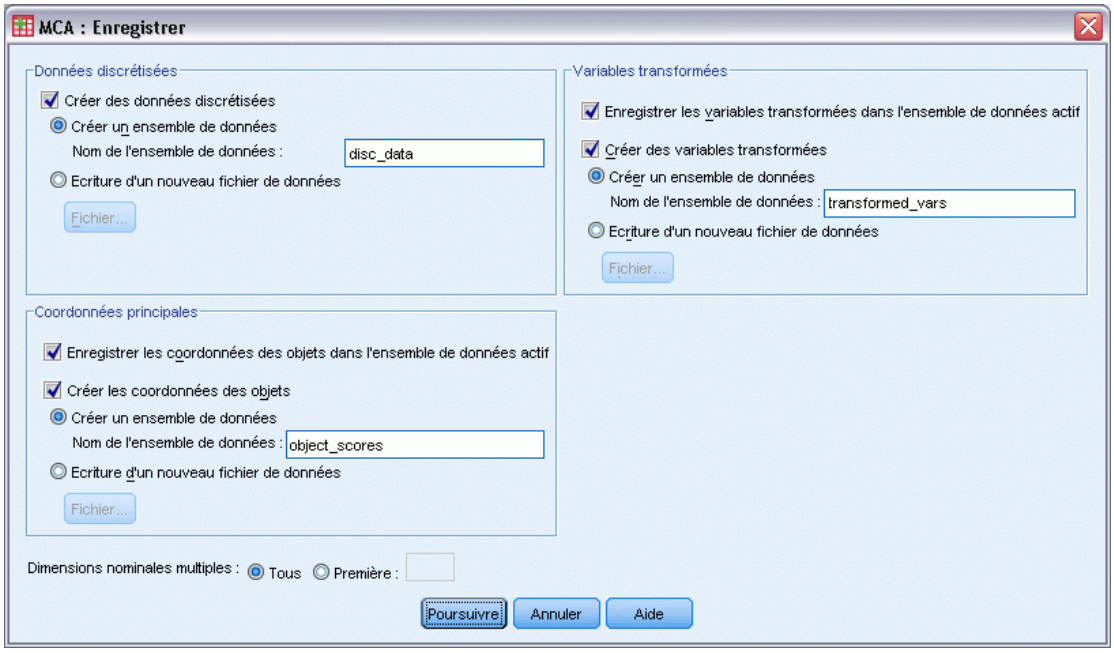
Statistiques descriptives : Affiche les effectifs, le nombre de valeurs manquantes et le mode de la ou des variables sélectionnées.

Analyse des correspondances multiples : Enregistrer

La boîte de dialogue Enregistrer vous permet d'enregistrer les données discrétisées, les coordonnées des objets et les valeurs transformées dans un fichier de données externe IBM® SPSS® Statistics ou un ensemble de données dans la session en cours. Vous pouvez également enregistrer les valeurs transformées et les coordonnées des objets dans l'ensemble de données actif.

- Les ensembles de données sont disponibles lors de la session en cours mais ne sont pas disponibles lors des sessions suivantes, sauf si vous les enregistrez clairement comme fichiers de données. Les noms des ensembles de données doivent être conformes aux règles de dénomination de variables.
- Les noms de fichiers ou les noms de l'ensemble de données doivent être différents pour chaque type de données enregistrées.
- Si vous enregistrez les coordonnées des objets ou les valeurs transformées dans l'ensemble de données actif, vous pouvez indiquer le nombre des dimensions nominales multiples.

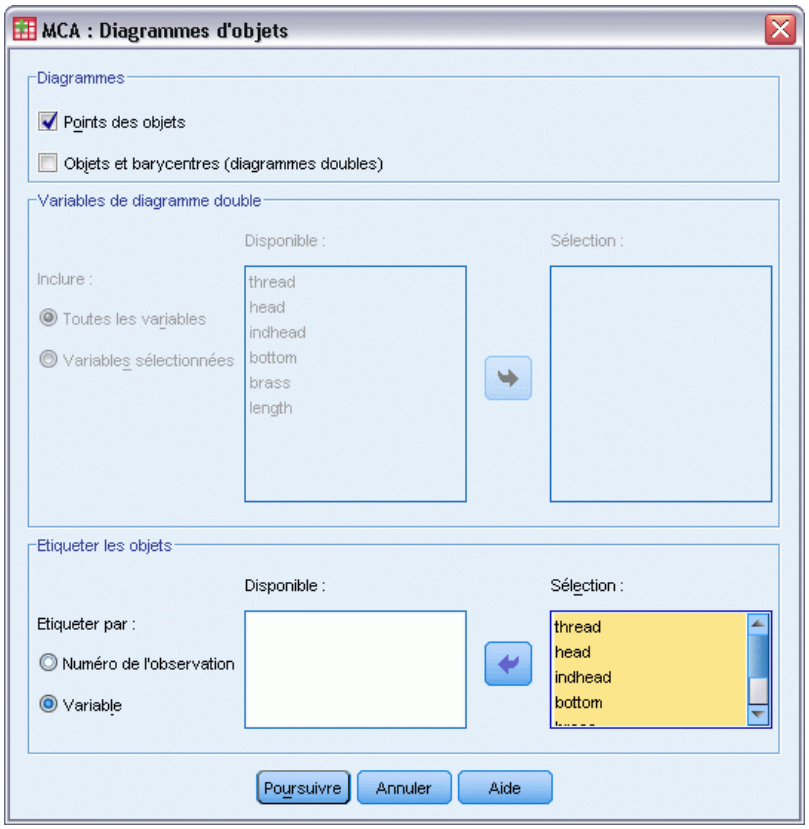
Figure 6-8
Enregistrer



Analyse des correspondances multiples : Diagrammes d'objets

La boîte de dialogue Diagrammes d'objets vous permet d'indiquer les types de diagrammes souhaités ainsi que les variables à représenter.

Figure 6-9
Boîte de dialogue Diagrammes d'objets



Points des objets. Un diagramme des points des objets s'affiche.

Objets et barycentres (diagrammes doubles) : Les points des objets sont représentés avec les barycentres de variable.

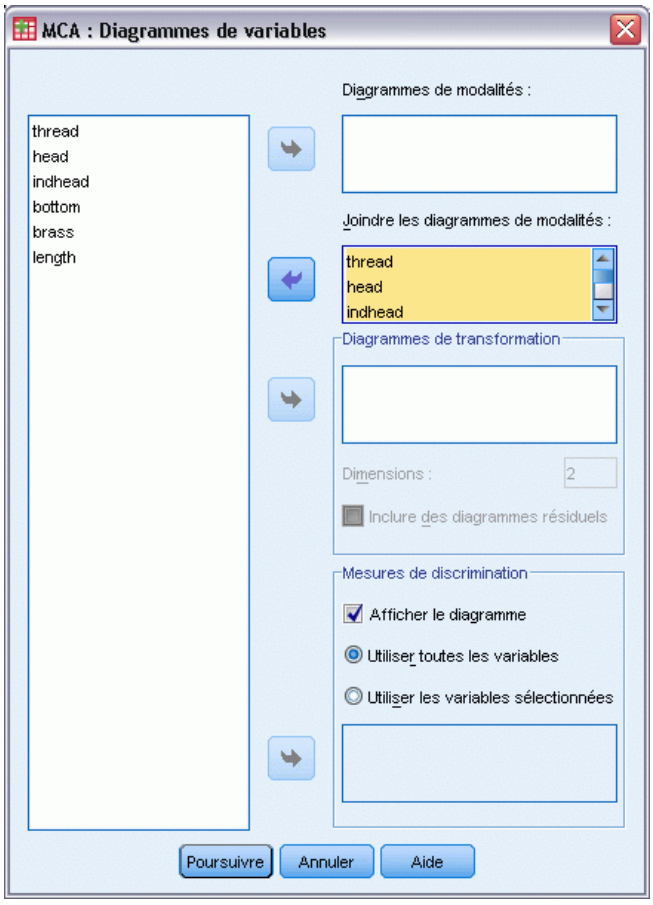
Variables de diagramme double. Vous pouvez choisir d'utiliser toutes les variables des diagrammes doubles ou de sélectionner un sous-groupe.

Etiqueter objets : Vous pouvez choisir d'étiqueter des objets avec les modalités des variables sélectionnées (choisissez les valeurs des indicateurs de modalités ou les étiquettes de valeurs dans la boîte de dialogue Options) ou avec le nombre d'observations. Si vous avez sélectionné Variables, un seul diagramme est créé par variable.

Analyse des correspondances multiples : Diagrammes de variables

La boîte de dialogue Diagrammes de variables vous permet d'indiquer les types de diagrammes souhaités ainsi que les variables à représenter.

Figure 6-10
Boîte de dialogue Diagrammes de variables



Diagrammes de modalités : Pour chaque variable sélectionnée, un diagramme des coordonnées du barycentre est représenté. Les modalités se trouvent dans les barycentres des objets des modalités concernées.

Joindre les diagrammes de modalités : Il s'agit d'un diagramme simple représentant les coordonnées du barycentre de chaque variable sélectionnée.

Diagrammes de transformation : Affiche un diagramme des valeurs affectées aux modalités optimales contre les indicateurs de modalités. Vous pouvez spécifier le nombre de dimensions souhaité. Un diagramme sera créé pour chaque dimension. Il vous est également possible de choisir d'afficher des diagrammes résiduels pour chaque variable sélectionnée.

Mesures de discrimination. Crée un diagramme des mesures de discrimination pour les variables sélectionnées.

Commande MULTIPLE CORRESPONDENCE - Caractéristiques additionnelles

Vous pouvez personnaliser votre analyse de correspondance multiple en collant vos sélections dans une fenêtre de syntaxe, puis en modifiant la syntaxe de la commande `MULTIPLE CORRESPONDENCE`. Le langage de syntaxe de commande vous permet aussi de :

- Spécifiez les noms de racine des variables transformées, les coordonnées des objets et les approximations lorsque vous les enregistrez dans l'ensemble de données actif (avec la sous-commande `SAVE`).
- Spécifier la longueur maximale pour les étiquettes de chaque diagramme séparément (avec la sous-commande `PLOT`).
- Spécifier une liste de variables distincte pour les diagrammes résiduels (avec la sous-commande `PLOT`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
7

Positionnement multidimensionnel (PROXSCAL)

Le positionnement multidimensionnel tente de déterminer la structure d'un groupe de mesures de proximité entre les objets. Ce procédé est effectué en affectant des observations à des positions particulières dans un espace conceptuel de petite dimension de telle sorte que les distances entre les points dans l'espace correspondent le mieux possible aux (dis)similarités données. Le résultat est une représentation à moindres carrés des objets dans cet espace de petite dimension, qui vous aidera, dans certains cas, à mieux comprendre vos données.

Exemple : Le positionnement multidimensionnel peut être très utile pour déterminer les relations perceptuelles. Par exemple, en considérant l'image de votre produit, vous pouvez mener une enquête en vue d'obtenir un fichier de données décrivant la similarité distinguée (ou proximité) de votre produit comparée à celle de vos concurrents. En utilisant ces variables de proximité et indépendantes (un prix, par exemple), vous pouvez essayer de déterminer quelles variables sont importantes suivant le mode d'affichage de ces produits et vous pouvez ajuster votre image en fonction.

Diagrammes et statistiques : Historique des itérations, mesures de stress, décomposition du stress, coordonnées de l'espace commun, distances des objets dans la configuration finale, pondérations des espaces individuels, espaces individuels, proximités transformées, variables indépendantes transformées, diagrammes de stress, diagrammes de dispersion des espaces communs, diagrammes de dispersion de pondération des espaces individuels, diagrammes de dispersion des espaces individuels, diagrammes de transformation, diagrammes résiduels de Shepard et diagrammes de transformation des variables explicatives.

Données : Les données peuvent être indiquées dans le formulaire des matrices de proximité ou des variables qui sont converties en matrice de proximité. Les matrices peuvent être formatées en colonnes ou entre les colonnes. Les proximités peuvent être traitées par niveaux de codage rapport, intervalle, ordinal ou spline.

Hypothèses : Trois variables au moins doivent être spécifiées. Le nombre de dimensions ne doit pas dépasser le nombre d'objets moins un. La réduction du nombre de dimensions est omise si elle est combinée aux départs aléatoires multiples. Si vous indiquez une source seulement, tous les modèles équivalent au modèle d'identité, puis l'analyse sélectionne par défaut le modèle d'identité.

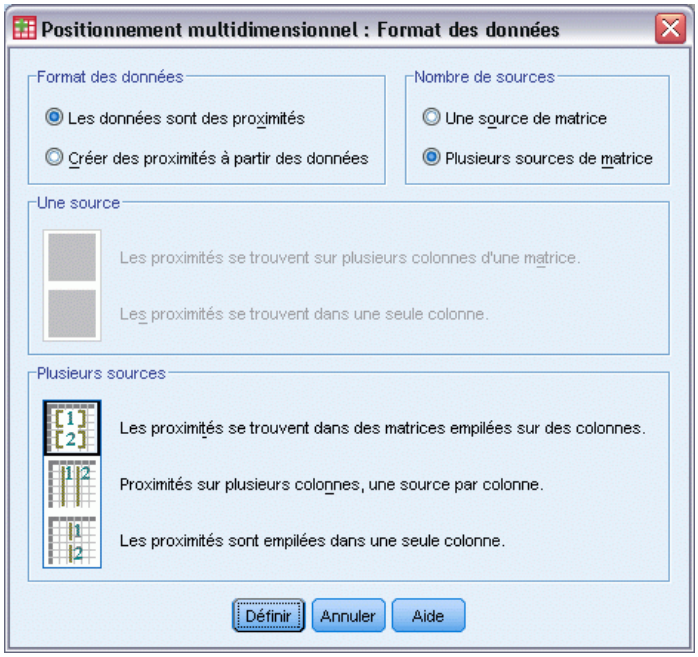
Procédures apparentées : Le codage de toutes les variables à un niveau numérique correspond au positionnement multidimensionnel standard.

Obtenir un positionnement multidimensionnel

▶ A partir des menus, sélectionnez :
Analyse > Echelle > Positionnement multidimensionnel (PROXSCAL)

Cette opération ouvre la boîte de dialogue Format des données.

Figure 7-1
Boîte de dialogue Format des données



▶ Spécifiez le format des données :

Format des données : Indiquez si vos données constituent des mesures de proximité ou si vous souhaitez créer des proximités à partir des données.

Nombre de sources : Si vos données sont des proximités, spécifiez si vous avez des sources uniques ou multiples de mesures de proximité.

Une source : S'il existe une source de proximité, spécifiez si votre fichier de données est formaté avec les proximités d'une matrice sur des colonnes ou sur une colonne unique avec deux variables séparées pour identifier les lignes et colonnes de chaque proximité.

- **Les proximités sont dans une matrices dans des colonnes.** La matrice de proximité s'étend à des colonnes dont le nombre est égal au nombre d'objets. Vous accédez ensuite à la boîte de dialogue Proximités sur plusieurs colonnes de matrices.
- **Les proximités sont dans une seule colonne.** Les matrices de proximité sont réduites dans une seule colonne, ou variable. Deux variables supplémentaires identifiant la ligne et la colonne de chaque cellule sont nécessaires. Vous accédez ensuite à la boîte de dialogue Proximités sur une seule colonne.

Plusieurs sources : S'il existe plusieurs sources de proximités, spécifiez si le fichier de données est formaté avec les proximités des matrices empilées sur plusieurs colonnes, sur des colonnes multiples contenant une source par colonne ou sur une colonne simple.

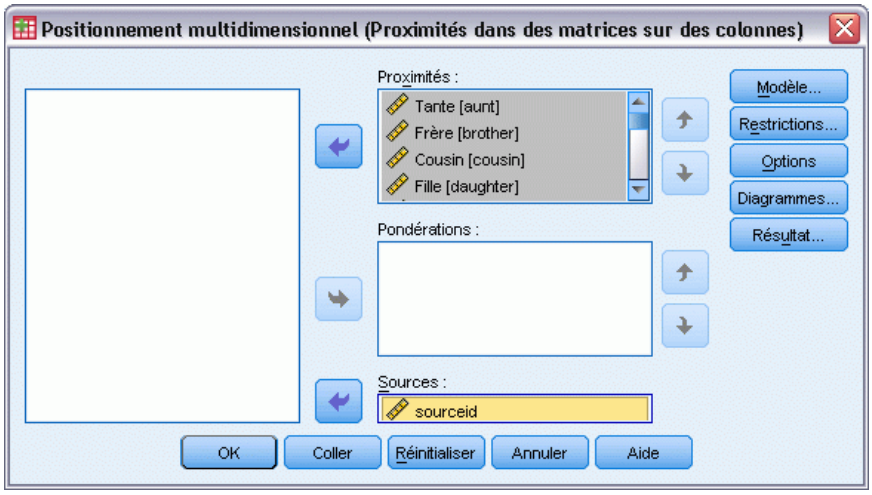
- **Les proximités sont dans des matrices empilées dans des colonnes.** Les matrices de proximité s'étalent sur un nombre de colonnes équivalent au nombre d'objets et sont empilées les unes sur les autres sur un nombre de lignes équivalent au produit du nombre d'objets et du nombre de sources. Vous accédez ensuite à la boîte de dialogue Proximités sur plusieurs colonnes de matrices.
- **Les proximités sont dans des colonnes, une source par colonne.** Les matrices de proximité sont réduites dans plusieurs colonnes, ou variables. Deux variables supplémentaires identifiant la ligne et la colonne de chaque cellule sont nécessaires. Vous accédez ensuite à la boîte de dialogue Proximités sur des colonnes.
- **Les proximités sont empilées dans une seule colonne.** Les matrices de proximité sont réduites dans une seule colonne, ou variable. Trois variables supplémentaires identifiant la ligne, la colonne et la source de chaque cellule, sont nécessaires. Vous accédez ensuite à la boîte de dialogue Proximités sur une seule colonne.

▶ Cliquez sur Définir.

Proximités dans des matrices sur plusieurs colonnes

Si vous sélectionnez les proximités dans un modèle de matrice des données pour une ou plusieurs sources dans la boîte de dialogue Format des données, la boîte de dialogue principale s'affiche comme ci-dessous :

Figure 7-2
Boîte de dialogue Proximités dans des matrices sur plusieurs colonnes



▶ Sélectionnez deux ou plusieurs variables de proximité. (Veuillez vous assurer que l'ordre des variables dans la liste correspond à l'ordre des colonnes des proximités.)

▶ Sélectionnez éventuellement un nombre de variables de pondération égal au nombre des variables de proximité. (Veuillez vous assurer que l'ordre des pondérations correspond à celui des proximités qu'elles pondèrent.)

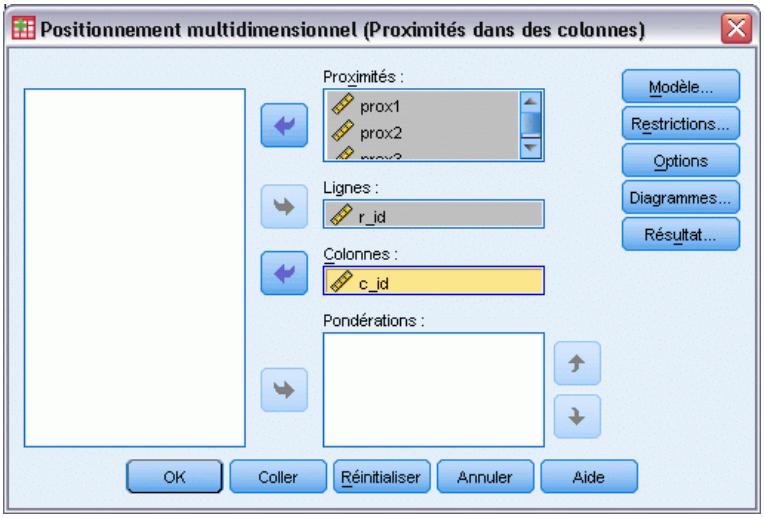
- S'il existe plusieurs sources, vous pouvez également sélectionner une variable de sources. (Le nombre d'observations dans chaque variable de proximité doit être égal au nombre de variables de proximité multiplié par le nombre de sources.)

  De plus, vous pouvez définir un modèle pour un positionnement multidimensionnel, placer les restrictions dans l'espace commun, définir les critères de convergence, spécifier la configuration initiale à utiliser et enfin, choisir des diagrammes et des résultats.

Proximités sur plusieurs colonnes

Si vous sélectionnez le modèle de colonnes multiples pour plusieurs sources dans la boîte de dialogue Format des données, la boîte de dialogue principale s'affiche comme ci-dessous :

Figure 7-3
Boîte de dialogue Proximités sur plusieurs colonnes



- Sélectionnez deux ou plusieurs variables. (Chaque variable est considérée comme étant une matrice de proximité provenant d'une source distincte.)

- Sélectionnez une variable de lignes pour définir les positions de lignes pour les proximités dans chaque variable de proximités.

- Sélectionnez une variable de colonnes pour définir les positions de colonnes pour les proximités dans la variable des proximités. (Les cellules de la matrice de proximité n'ayant pas de désignation lignes/colonnes sont considérées comme manquantes.)

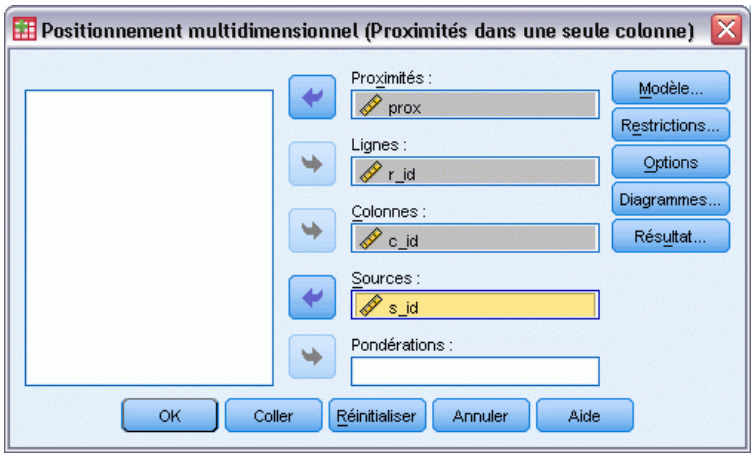
- Sélectionnez éventuellement un nombre de variables de pondération égal au nombre des variables de proximité.

  De plus, vous pouvez définir un modèle pour un positionnement multidimensionnel, placer les restrictions dans l'espace commun, définir les critères de convergence, spécifier la configuration initiale à utiliser et enfin, choisir des diagrammes et des résultats.

Proximités dans une colonne

Si vous sélectionnez le modèle de colonne unique pour une ou plusieurs sources dans la boîte de dialogue Format des données, la boîte de dialogue principale s'affiche comme ci-dessous :

Figure 7-4
Boîte de dialogue Proximités dans une colonne



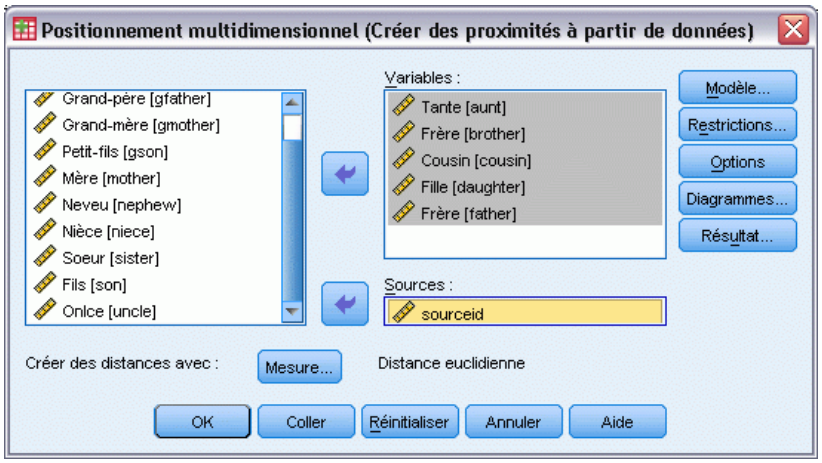
- Sélectionnez une variable de proximité. (On considère qu'il existe une ou plusieurs matrices des proximités.)
- Sélectionnez une variable de lignes pour définir les positions de lignes pour les proximités dans la variable des proximités.
- Sélectionnez une variable de colonnes pour définir les positions de colonnes pour les proximités dans la variable des proximités.
- S'il existe plusieurs sources, sélectionnez une variable de sources. (Pour chaque source, les cellules de la matrice de proximité n'ayant pas de désignation lignes/colonnes sont considérées comme manquantes.)
- Eventuellement, choisissez une variable de pondération.

De plus, vous pouvez définir un modèle pour un positionnement multidimensionnel, placer les restrictions dans l'espace commun, définir les critères de convergence, spécifier la configuration initiale à utiliser et enfin, choisir des diagrammes et des résultats.

Créer des proximités à partir des données

Si vous sélectionnez de créer des proximités dans la boîte de dialogue Format des données, la boîte de dialogue principale s'affiche comme ci-dessous :

Figure 7-5
Boîte de dialogue Créer des proximités à partir des données

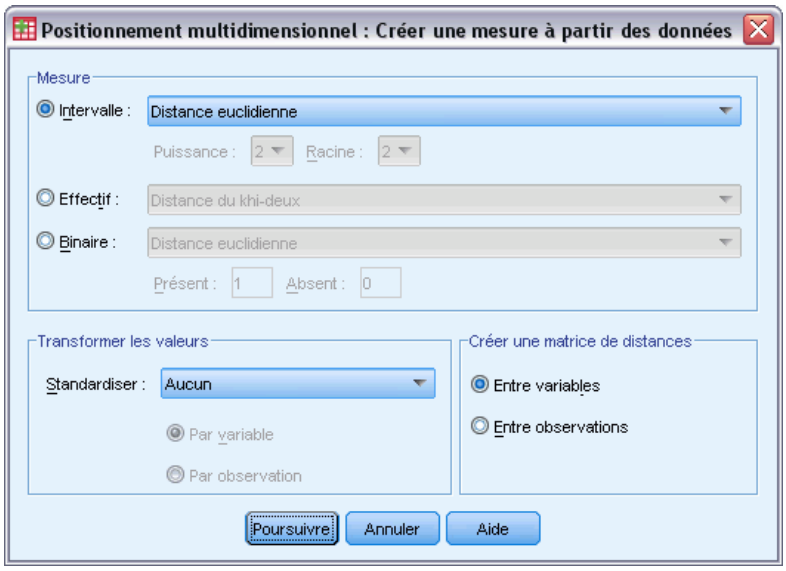


- Si vous créez des distances entre les variables (voir la boîte de dialogue Créer une mesure à partir des données), sélectionnez au moins trois variables. Ces variables seront utilisées pour créer la matrice de proximité (ou les matrices, s'il existe plusieurs sources). Si vous créez des distances entre les observations, seule une variable est requise.
- S'il existe plusieurs sources, sélectionnez une variable de sources.
- Choisissez éventuellement une mesure de création de proximités.

  De plus, vous pouvez définir un modèle pour un positionnement multidimensionnel, placer les restrictions dans l'espace commun, définir les critères de convergence, spécifier la configuration initiale à utiliser et enfin, choisir des diagrammes et des résultats.

Créer une mesure à partir des données

Figure 7-6
Boîte de dialogue Créer une mesure à partir des données



Le positionnement multidimensionnel utilise les données de dissimilarité pour créer une solution de codage. Si vos données sont multivariées (valeurs des variables mesurées), vous devez créer des données de dissimilarité afin de calculer une solution de positionnement multidimensionnel. Vous pouvez spécifier les détails de création de mesures de dissimilarité à partir de vos données.

Mesure : Vous permet de spécifier la mesure de dissimilarité adaptée à votre analyse. Sélectionnez une possibilité dans le groupe Mesure correspondant à votre type de données, puis sélectionnez l'une des mesures dans la liste déroulante correspondant à ce type de mesure. Les possibilités sont :

- **Intervalle :** Distance Euclidienne, Carré de la distance Euclidienne, Distance de Tchebycheff, Distance de Manhattan, Distance de Minkowski ou Autre.
- **Effectif** : Distance du Khi-deux ou Distance du phi-deux.
- **Binaire :** Distance Euclidienne, Carré de la distance Euclidienne, Ecart de taille, Différence de motif, Variance ou Lance et Williams.

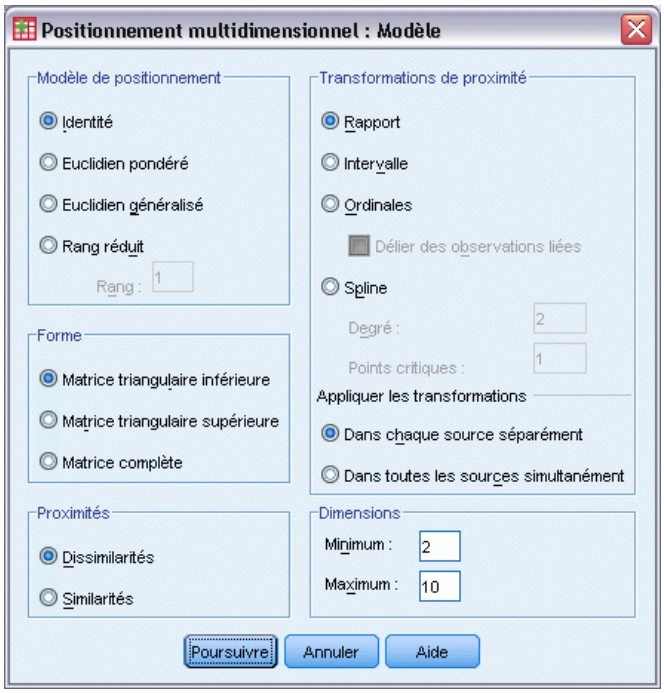
Créer une matrice de distances : Vous permet de choisir l'unité d'analyse. Les possibilités sont Par variables ou Par observations.

Transformer les valeurs : Dans certains cas, comme lorsque les variables sont mesurées selon des échelles très différentes, vous voudrez standardiser des valeurs avant de calculer les proximités (ne s'applique pas aux données binaires). Sélectionnez une méthode de standardisation dans la liste déroulante Standardiser (si la standardisation n'est pas nécessaire, sélectionnez Aucune).

Définir un modèle de positionnement multidimensionnel

La boîte de dialogue Modèle vous permet d'indiquer un modèle de positionnement, son nombre minimum et maximum de dimensions, la structure de la matrice de proximité, la transformation à utiliser sur les proximités, et de déterminer si les proximités sont transformées dans chaque source séparément ou sans condition sur la source.

Figure 7-7
Boîte de dialogue Modèle



Modèle de positionnement. Choisissez parmi les options suivantes :

- **Identité :** Toutes les sources ont la même configuration.
- **Euclidien pondéré :** Ce modèle est un modèle des différences individuelles. Chaque source comporte un espace individuel dans lequel chaque dimension de l'espace commun est pondérée de façon différentielle.
- **Euclidien généralisé :** Ce modèle est un modèle des différences individuelles. Chaque source comporte un espace individuel qui est égal à une rotation de l'espace commun, suivie d'une pondération différenciée des dimensions.
- **Rang réduit :** Il s'agit d'un modèle euclidien généralisé pour lequel vous pouvez spécifier le rang de l'espace individuel. Vous devez spécifier un rang supérieur ou égal à 1 et inférieur au nombre maximum de dimensions.

Forme : Spécifiez si les proximités doivent être extraites des parties triangulaires inférieure ou supérieure de la matrice de proximité. Vous pouvez indiquer que la totalité de la matrice est utilisée, auquel cas la somme pondérée des parties triangulaires supérieure et inférieure sera analysée. Dans tous les cas, la matrice complète doit être spécifiée, y compris la diagonale, même si les parties spécifiées seront les seules à être utilisées.

Proximités : Spécifiez si votre matrice de proximité contient des mesures de similarité ou de dissimilarité.

Transformations de proximité : Choisissez parmi les options suivantes :

- **Rapport :** Les proximités transformées sont proportionnelles aux proximités originales. Uniquement disponible pour les proximités à valeurs positives.
- **Intervalle :** Les proximités transformées sont proportionnelles aux proximités originales et à une constante. Cette constante fait en sorte que les proximités transformées soient positives.
- **Ordinal :** Les proximités transformées ont le même ordre que les originales. Vous spécifiez si les proximités liées doivent être gardées liées ou autorisées à ne plus l'être.
- **Spline.** Les proximités transformées représentent une transformation polynomiale non décroissante lissée des proximités originales. Vous spécifiez le degré de la fonction polynomiale ainsi que le nombre de points critiques.

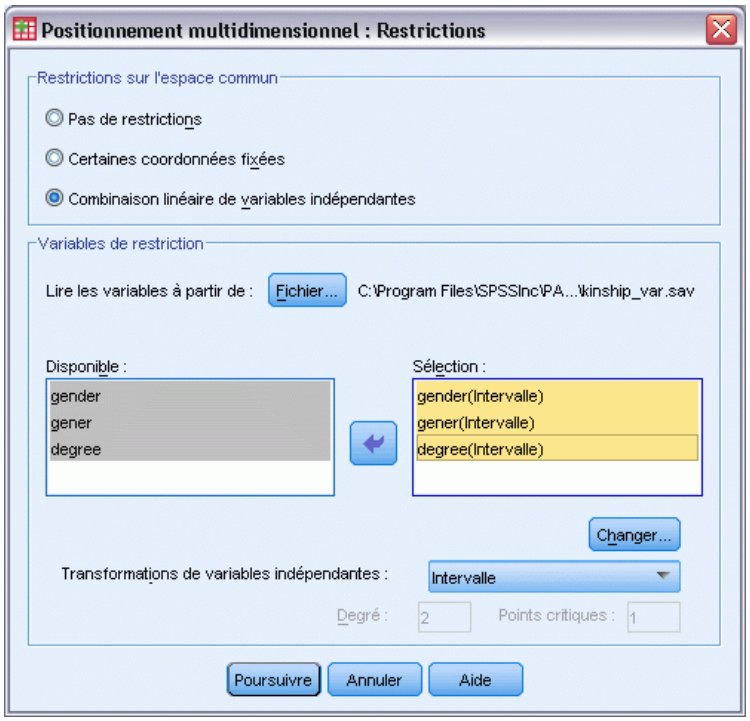
Appliquer les transformations : Spécifiez si seules les proximités de chaque source sont comparées entre elles ou si les comparaisons sont sans condition sur la source.

Dimensions : Par défaut, une solution est calculée dans deux dimensions (minimum =2, maximum =2). Vous choisissez un entier minimum et maximum depuis 1 jusqu'au nombre d'objets moins 1 (tant que le minimum reste inférieur ou égal au maximum.) La procédure calcule une solution des dimensions maximales, puis réduit le nombre de dimensions en matière d'étapes, jusqu'à ce que la plus petite soit atteinte.

Positionnement multidimensionnel : Restrictions

La boîte de dialogue Restrictions vous permet de placer les restrictions sur l'espace commun.

Figure 7-8
Boîte de dialogue Restrictions



Restrictions sur l'espace commun : Spécifiez le type de restriction souhaité.

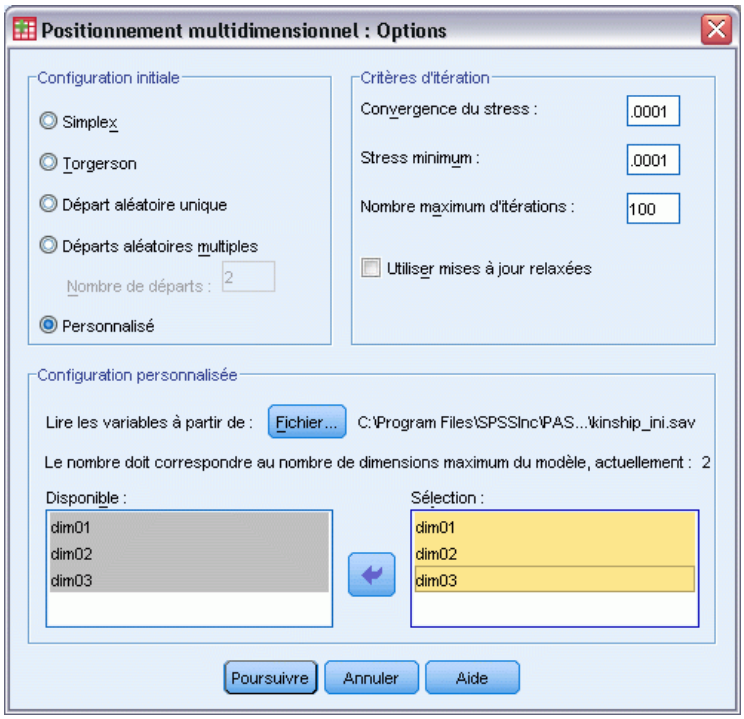
- **Pas de restrictions :** Aucune restriction n'est placée sur l'espace commun.
- **Certaines coordonnées fixées :** La première variable sélectionnée contient les coordonnées des objets sur la première dimension ; la seconde correspond aux coordonnées des objets sur la deuxième dimension, et ainsi de suite. Une valeur manquante indique qu'une coordonnée sur une dimension est libre. Le nombre de variables sélectionnées doit être égal au nombre maximum de dimensions requis.
- **Combinaison linéaire de variables indépendantes :** L'espace commun se réduit à une combinaison linéaire des variables sélectionnées.

Variables de restriction : Sélectionnez les variables qui définissent les restrictions sur l'espace commun. Si vous spécifiez une combinaison linéaire, vous spécifiez une transformation d'intervalle, nominale, ordinale ou spline pour des variables de restriction. Dans tous les cas, le nombre d'observations pour chaque variable doit être égal au nombre d'objets.

Positionnement multidimensionnel : Options

La boîte de dialogue Options vous permet de sélectionner le style de configuration initiale, de spécifier les critères d'itération et de convergence, et de sélectionner des mises à jour standard ou relaxées.

Figure 7-9
Options



Configuration initiale : Choisissez l'une des options suivantes :

- **Simplex .** Les objets sont placés à la même distance les uns des autres dans la dimension maximale. Une itération est prise pour améliorer cette configuration à haute dimension, suivie d'une réduction de dimension en vue d'obtenir une configuration initiale comportant le nombre maximum de dimensions spécifié dans la boîte de dialogue Modèle.
- **Torgerson :** Une solution de codage classique est utilisée comme configuration initiale.
- **Départ aléatoire unique :** Une configuration est choisie de façon aléatoire.
- **Départs aléatoires multiples :** Plusieurs configurations sont choisies de façon aléatoire, et celle ayant le stress brut le moins standardisé est utilisée comme configuration initiale.
- **Personnalisé :** Vous sélectionnez les variables contenant les coordonnées de votre configuration initiale. Le nombre de variables sélectionnées doit être égal au nombre de dimensions spécifié, la première variable correspondant aux coordonnées sur la dimension 1, la seconde correspondant aux coordonnées sur la dimension 2, etc. Le nombre d'observations dans chaque variable doit être égal au nombre d'objets.

Critères d'itération : Spécifiez les valeurs des critères d'itération.

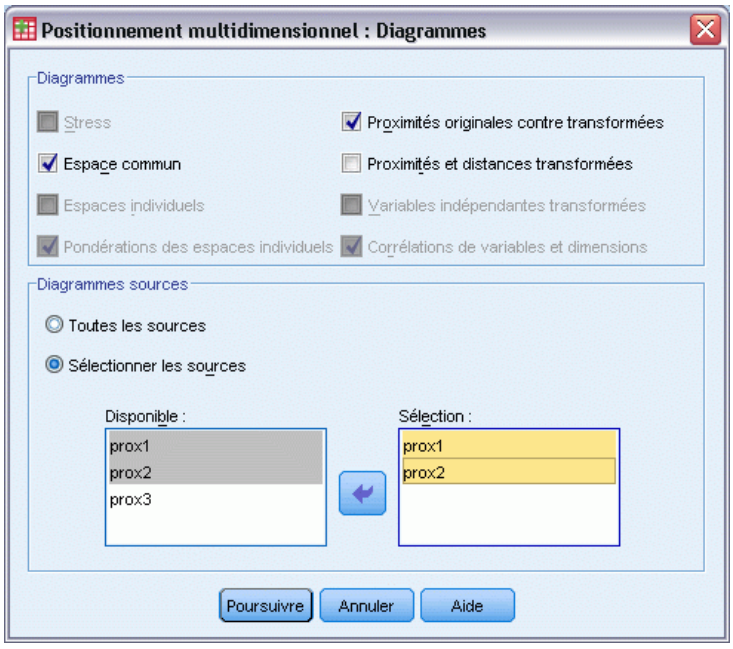
- **Convergence du stress :** L'algorithme interrompt son itération lorsque la différence des valeurs du stress brut standardisé consécutif est inférieure au nombre spécifié ici, lequel doit être compris entre 0,0 et 1,0.
- **Stress minimum :** L'algorithme s'interrompt lorsque le stress brut standardisé tombe en dessous du nombre spécifié ici, lequel doit être compris entre 0,0 et 1,0.

- **Nombre maximum d'itérations :** L'algorithme exécute le nombre d'itérations spécifiées ici, à moins que l'un des critères ci-dessus ne soit déjà satisfait.
- **Utiliser mises à jour relaxées :** Ces mises à jour accélèrent l'algorithme ; elles ne peuvent être utilisées ni avec les modèles autres que le modèle d'identité, ni avec des restrictions.

Positionnement multidimensionnel : Diagrammes, Version 1

La boîte de dialogue Diagrammes vous permet de spécifier quels diagrammes doivent être produits. Si vous avez le format de données Proximités sur plusieurs colonnes, la boîte de dialogue Diagrammes suivante s'affiche. Pour les diagrammes Pondérations des espaces individuels, Proximités originales contre transformées et Proximités contre distances transformées, il vous est possible de spécifier les sources pour lesquelles les diagrammes doivent être générés. La liste des sources disponibles constitue la liste des variables de proximité dans la boîte de dialogue principale.

Figure 7-10
Boîte de dialogue Diagrammes, version 1



Stress : Un diagramme est produit à partir du stress brut standardisé par opposition aux dimensions. Ce diagramme est uniquement généré si le nombre maximum de dimensions est supérieur au nombre minimum.

Espace commun : Une matrice de diagramme de dispersion des coordonnées de l'espace commun est affiché.

Espaces individuels : Pour chaque source, les coordonnées des espaces individuels sont affichées dans les matrices de diagramme de dispersion. Cela est uniquement possible si l'un des modèles de différences individuels est spécifié dans la boîte de dialogue Modèle.

Pondérations des espaces individuels : Un diagramme de dispersion est produit à partir des pondérations des espaces individuels. Cela est uniquement possible si l'un des modèles de différences individuels est spécifié dans la boîte de dialogue Modèle. Pour le modèle Euclidien pondéré, les pondérations sont imprimées dans des diagrammes dont une dimension sur chaque axe. Pour le modèle Euclidien généralisé, un diagramme est produit par dimension, indiquant à la fois la rotation et sa pondération. Le modèle Rang réduit génère le même diagramme que le modèle Euclidien généralisé, mais réduit le nombre de dimensions des espaces individuels.

Proximités originales contre transformées : Les diagrammes sont générés à partir des proximités originales par opposition aux proximités transformées.

Proximités et distances transformées. Les proximités et distances transformées sont représentées sous la forme d'un diagramme.

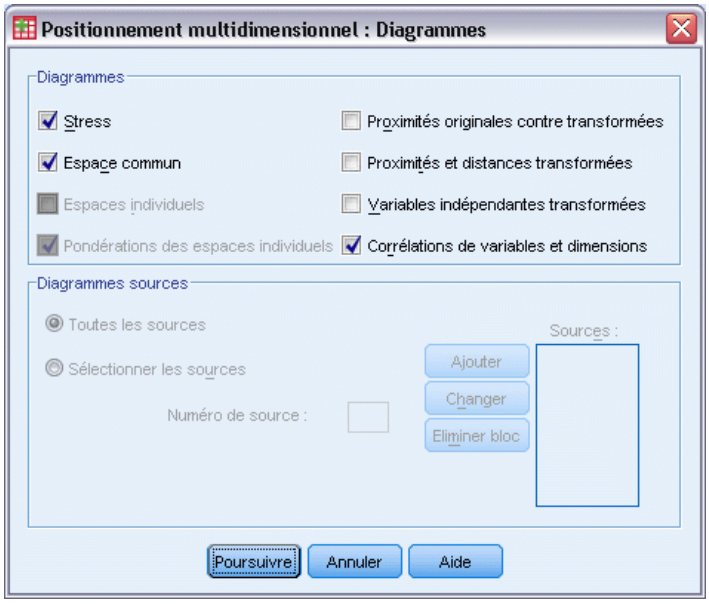
Variables indépendantes transformées : Les diagrammes de transformation sont produits pour les variables indépendantes.

Corrélations de variables et dimensions : Un diagramme de corrélation entre les variables indépendantes et les dimensions de l'espace commun s'affiche.

Positionnement multidimensionnel : Diagrammes, Version 2

La boîte de dialogue Diagrammes vous permet de spécifier quels diagrammes doivent être produits. Si votre format des données est autre que Proximités sur plusieurs colonnes, la boîte de dialogue Diagrammes suivante s'affiche. Pour les diagrammes Pondérations des espaces individuels, Proximités originales contre transformées et Proximités contre distances transformées, il vous est possible de spécifier les sources pour lesquelles les diagrammes doivent être générés. Les numéros de source entrés doivent être des valeurs de la variable de sources spécifiée dans la boîte de dialogue principale et être classés de 1 jusqu'au nombre de sources.

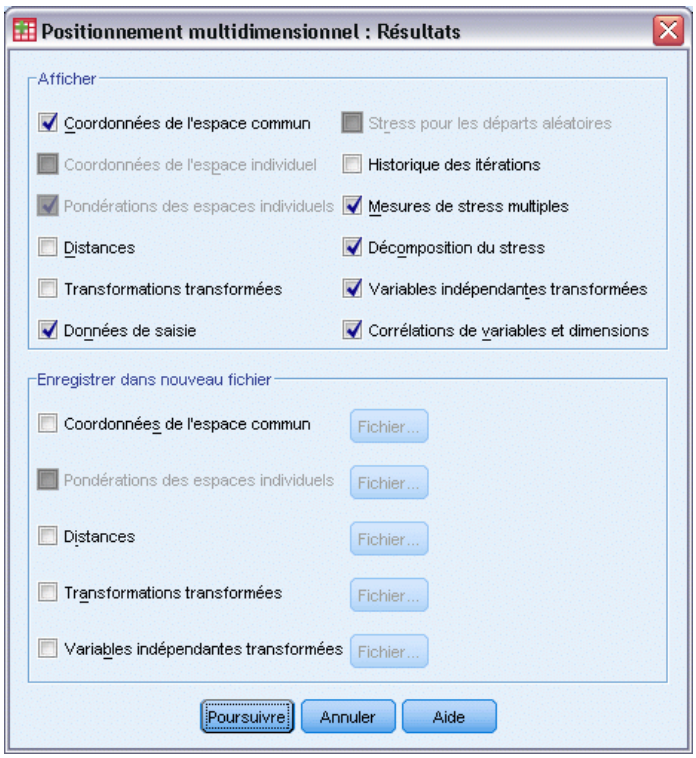
Figure 7-11
Boîte de dialogue Diagrammes, version 2



Positionnement multidimensionnel : Résultat

La boîte de dialogue Résultat vous permet de contrôler les résultats affichés et d'en enregistrer certains pour séparer des fichiers.

Figure 7-12
Résultat



Afficher : Sélectionnez l'un des items suivants à afficher :

- **Coordonnées de l'espace commun :** Affiche les coordonnées de l'espace commun.
- **Coordonnées de l'espace individuel :** Les coordonnées de l'espace individuel sont affichées uniquement si le modèle n'est pas le modèle d'identité.
- **Pondérations des espaces individuels :** Affiche les pondérations des espaces individuels, uniquement si l'un des modèles de différences individuelles est spécifié. En fonction du modèle, les pondérations des espaces sont décomposées en pondérations de rotation et de dimension, lesquelles sont également affichées.
- **Distances :** Affiche les distances entre les objets de la configuration.
- **Transformations transformées :** Affiche les proximités transformées entre les objets de la configuration.
- **Données de saisie :** Inclut les proximités originales, et si elles existent, les pondérations de données, la configuration initiale et les coordonnées fixées des variables indépendantes.
- **Stress pour les départs aléatoires :** Affiche le générateur de nombre aléatoire et la valeur du stress brut standardisé de chaque départ aléatoire.
- **Historique des itérations :** Affiche l'historique des itérations de l'algorithme principal.

- **Mesures de stress multiples :** Affiche les différentes valeurs de stress. Le tableau contient des valeurs pour le stress brut standardisé, le stress-I, le stress-II, le stress-S, la dispersion représentée (DAF) et enfin le coefficient de congruence de Tucker.
- **Décomposition du stress :** Affiche la décomposition par objet et par source du stress brut final standardisé, notamment la moyenne par objet et par source.
- **Variables indépendantes transformées :** Si une restriction de combinaisons linéaires a été sélectionnée, les variables indépendantes transformées et les pondérations de régression correspondantes sont affichées.
- **Corrélations de variables et dimensions :** Si une restriction de combinaisons linéaires a été sélectionnée, les corrélations entre les variables indépendantes et les dimensions de l'espace commun sont affichées.

Enregistrer dans nouveau fichier : Vous pouvez enregistrer les coordonnées de l'espace commun, les pondérations des espaces individuels, les distances, les proximités transformées et les variables indépendantes transformées pour séparer les fichiers de données IBM® SPSS® Statistics.

Fonctionnalités supplémentaires de la commande PROXSCAL

Vous pouvez personnaliser l'analyse de votre positionnement multidimensionnel de proximité en collant vos sélections dans une fenêtre de syntaxe et en modifiant la syntaxe de commande `PROXSCAL`. Le langage de syntaxe de commande vous permet aussi de :

- Spécifier des listes de variables distinctes pour les diagrammes de transformation et résiduels (avec la sous-commande `PLOT`).
- Spécifier des listes de sources distinctes pour les diagrammes de pondération des espaces individuels, de transformation et résiduels (avec la sous-commande `PLOT`).
- Spécifier un sous-groupe des diagrammes de transformation de variables indépendantes à afficher (avec la sous-commande `PLOT`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
8

Dépliage multidimensionnel (PREFSCAL)

La procédure de dépliage multidimensionnel tente de trouver une échelle quantitative commune vous permettant d'examiner les relations entre deux ensembles d'objets de manière visuelle.

Exemples : Vous demandez à 21 personnes de classer 15 aliments constituant un petit-déjeuner selon leurs préférences, de 1 à 15. Le dépliage multidimensionnel vous permet de déterminer que la logique discriminatoire des individus suit deux schémas primaires : entre les pains mous et les pains durs et entre les aliments gras et allégés.

Autre exemple : vous demandez à un groupe de conducteurs de noter 26 modèles de voitures sur 10 critères selon une échelle de 6 points, allant de 1=« pas vrai du tout » à 6=« tout à fait vrai ». En effectuant la moyenne des résultats de l'ensemble des individus, on constate une certaine similarité des valeurs. Le dépliage multidimensionnel vous permet cependant de distinguer des regroupements de modèles similaires et les critères avec lesquels ils sont le plus souvent associés.

Diagrammes et statistiques : La procédure de dépliage multidimensionnel permet de produire un historique des itérations, les mesures de stress, la décomposition du stress, les coordonnées de l'espace commun, les distances des objets dans la configuration finale, les pondérations des espaces individuels, les espaces individuels, les proximités transformées, les diagrammes de stress, les diagrammes de dispersion des espaces communs, les diagrammes de dispersion de pondération des espaces individuels, les diagrammes de dispersion des espaces individuels, les diagrammes de transformation et les diagrammes résiduels de Shepard.

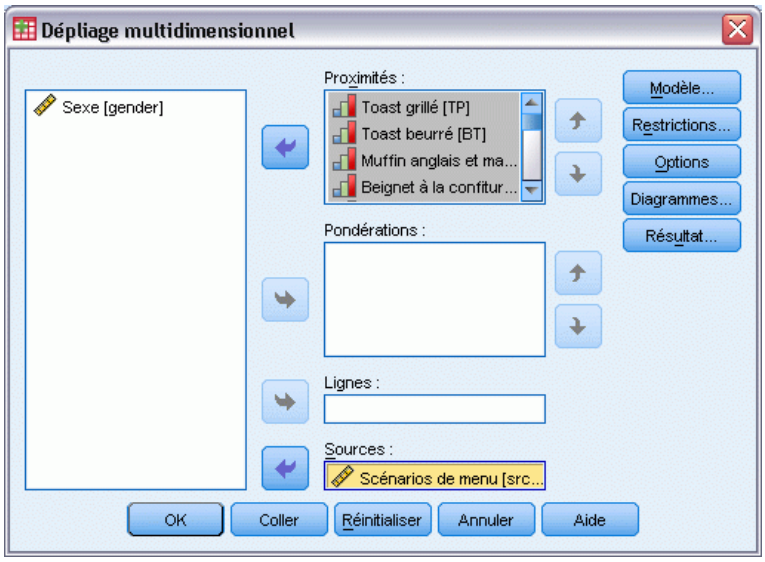
Données : Les données sont fournies sous forme de matrices de proximité rectangulaires. Chaque colonne est considérée comme un objet de colonne distinct. Chaque ligne d'une matrice de proximité est considérée comme un objet de ligne distinct. Lorsqu'il existe plusieurs sources de proximités, les matrices sont empilées.

Hypothèses : Deux variables au moins doivent être spécifiées. Le nombre de dimensions de la solution ne doit pas dépasser le nombre d'objets moins un. Si vous indiquez une source seulement, tous les modèles équivalent au modèle d'identité, puis l'analyse sélectionne par défaut le modèle d'identité.

Obtenir un dépliage multidimensionnel

- A partir des menus, sélectionnez :
  Analyse > Echelle > Dépliage multidimensionnel (PREFSCAL)...

Figure 8-1
Boîte de dialogue principale Dépliage multidimensionnel



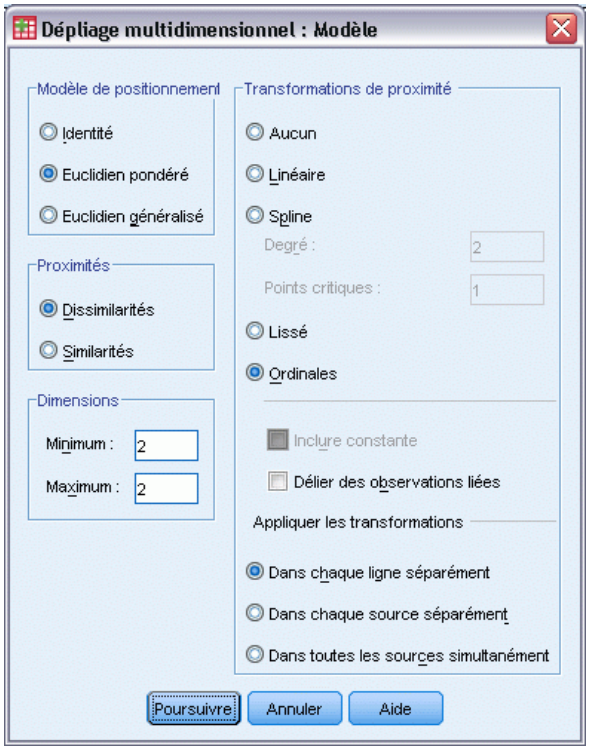
- Sélectionnez deux variables ou plus identifiant les colonnes dans la matrice de proximité rectangulaire. Chaque variable représente un objet de colonne distinct.
- Sélectionnez éventuellement un nombre de variables de pondération égal au nombre de variables d'objets de colonnes. L'ordre des variables de pondération doit être le même que celui des objets de colonnes qu'elles pondèrent.
- Eventuellement, choisissez une variable de ligne. Les valeurs (ou étiquettes de valeur) de cette variable sont utilisées pour étiqueter les objets de lignes du résultat.
- S'il existe plusieurs sources, sélectionnez éventuellement une variable de sources. Le nombre d'observations dans chaque fichier de données doit être égal au nombre d'objets de lignes multiplié par le nombre de sources.

  De plus, vous pouvez définir un modèle pour un dépliage multidimensionnel, placer les restrictions dans l'espace commun, définir les critères de convergence, spécifier la configuration initiale à utiliser et enfin, choisir des diagrammes et des résultats.

Définir un modèle de dépliage multidimensionnel

La boîte de dialogue Modèle vous permet de spécifier un modèle de positionnement, un nombre minimum et maximum de dimensions, la structure de la matrice de proximité, la transformation à utiliser sur les proximités, et de déterminer si les proximités sont transformées avec condition sur la ligne, avec condition sur la source ou sans condition sur la source.

Figure 8-2
Boîte de dialogue Modèle



Modèle de positionnement. Choisissez parmi les options suivantes :

- **Identité :** Toutes les sources ont la même configuration.
- **Euclidien pondéré :** Ce modèle est un modèle des différences individuelles. Chaque source comporte un espace individuel dans lequel chaque dimension de l'espace commun est pondérée de façon différentielle.
- **Euclidien généralisé :** Ce modèle est un modèle des différences individuelles. Chaque source comporte un espace individuel qui est égal à une rotation de l'espace commun, suivie d'une pondération différenciée des dimensions.

Proximités : Spécifiez si votre matrice de proximité contient des mesures de similarité ou de dissimilarité.

Dimensions : Par défaut, une solution est calculée dans deux dimensions (minimum =2, maximum =2). Vous pouvez choisir un entier minimum et maximum depuis 1 jusqu'au nombre d'objets moins 1, tant que le minimum reste inférieur ou égal au maximum. La procédure calcule une solution des dimensions maximales, puis réduit le nombre de dimensions en matière d'étapes, jusqu'à ce que la plus petite soit atteinte.

Transformations de proximité : Choisissez parmi les options suivantes :

- **Aucune.** Les proximités ne sont pas transformées. Vous pouvez éventuellement sélectionner Inclure une constante afin de décaler les proximités d'une constante définie.

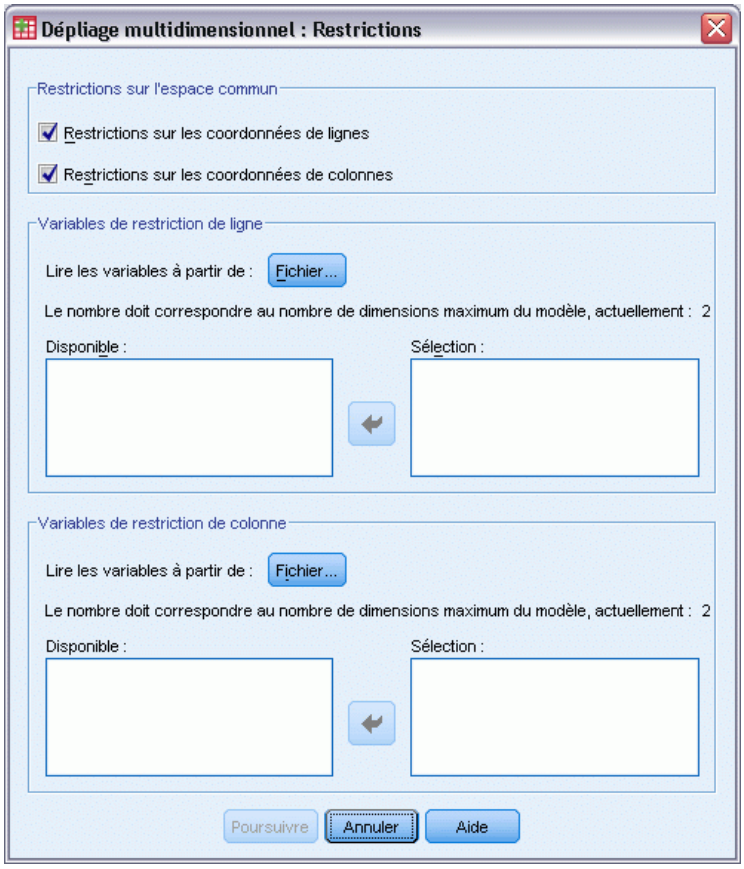
- **Linéaire :** Les proximités transformées sont proportionnelles aux proximités d'origine : la fonction de transformation estime une pente et la constante est définie sur 0. C'est ce qu'on appelle une transformation de ratio. Vous pouvez éventuellement sélectionner Inclure une constante afin de décaler les proximités d'une constante définie. Ce procédé est également appelé transformation d'intervalle.
- **Spline.** Les proximités transformées représentent une transformation polynomiale non décroissante lissée des proximités originales. Vous spécifiez le degré de la fonction polynomiale ainsi que le nombre de points critiques. Vous pouvez éventuellement sélectionner Inclure une constante afin de décaler les proximités d'une constante définie.
- **Lissé.** Les proximités transformées présentent le même ordre que les proximités d'origine, y compris la restriction qui prend en compte les différences entre les valeurs successives. Il en résulte une transformation « ordinale lissée ». Vous pouvez spécifier si les proximités liées doivent être gardées liées ou autorisées à ne plus l'être.
- **Ordinal :** Les proximités transformées ont le même ordre que les originales. Vous pouvez spécifier si les proximités liées doivent être gardées liées ou autorisées à ne plus l'être.

Appliquer les transformations : Spécifiez si les proximités sont comparées l'une à l'autre dans chaque ligne ou dans chaque source ou si les comparaisons sont sans condition sur la ligne ou sur la source, c'est à dire si les transformations sont effectuées par ligne, par source ou sur toutes les proximités en une fois.

Restrictions du dépliage multidimensionnel

La boîte de dialogue Restrictions vous permet de placer les restrictions sur l'espace commun.

Figure 8-3
Boîte de dialogue Restrictions



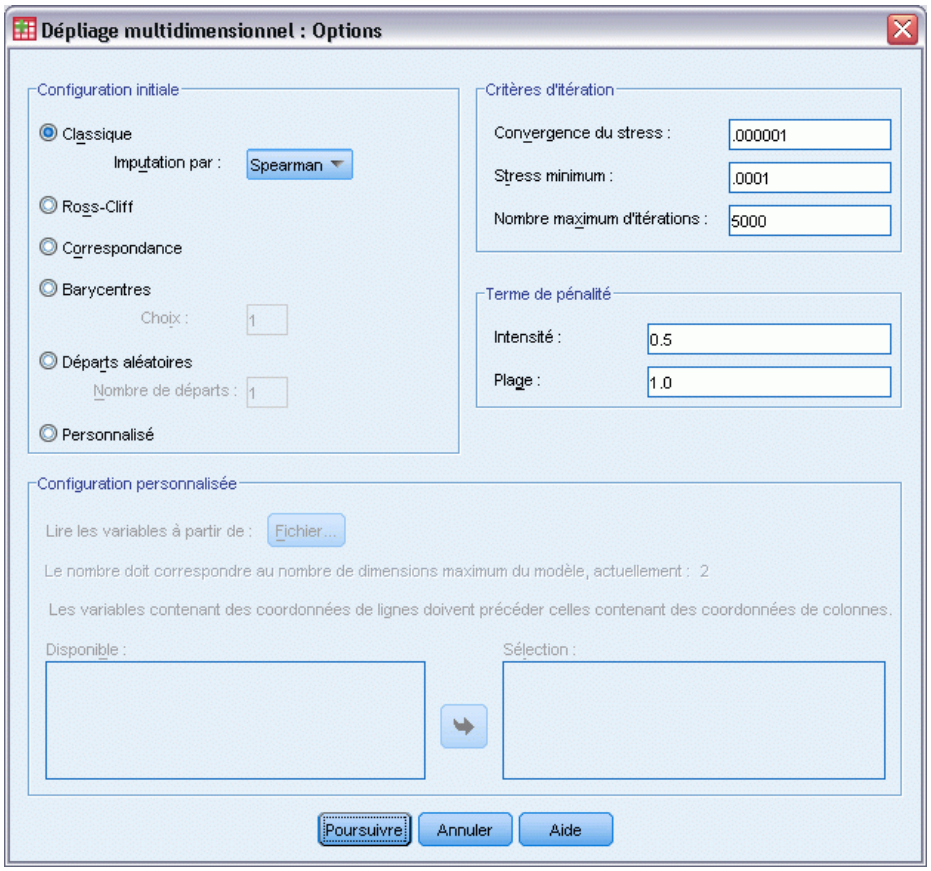
Restrictions sur l'espace commun : Vous pouvez choisir de fixer les coordonnées des objets de lignes et/ou de colonnes dans l'espace commun.

Variables de restriction des lignes/colonnes. Sélectionnez le fichier contenant les restrictions et sélectionnez les variables définissant les restrictions de l'espace commun. La première variable sélectionnée contient les coordonnées des objets sur la première dimension ; la seconde correspond aux coordonnées des objets sur la deuxième dimension, et ainsi de suite. Une valeur manquante indique qu'une coordonnée sur une dimension est libre. Le nombre de variables sélectionnées doit être égal au nombre maximum de dimensions requis. Le nombre d'observations dans chaque variable doit être égal au nombre d'objets.

Options de dépliage multidimensionnel

La boîte de dialogue Options vous permet de sélectionner le style de configuration initiale, de spécifier les critères d'itération et de convergence et de configurer le terme de pénalité pour le stress.

Figure 8-4
Options



Configuration initiale : Choisissez l'une des options suivantes :

- **Classique.** La matrice de proximité rectangulaire est utilisée pour compléter les valeurs intrablocs (valeurs entre les lignes et entre les colonnes) de la matrice MDS symétrique complète. Une fois la matrice complète formée, une solution de positionnement classique est utilisée pour la configuration initiale. Les valeurs intrablocs peuvent être calculées à l'aide de l'inégalité triangulaire ou des distances de Spearman.
- **Ross-Cliff.** Le départ de Ross-Cliff utilise les résultats de la décomposition d'une valeur singulière sur une matrice de proximité à double centre carrée comme valeurs initiales pour les objets de lignes et de colonnes.
- **Correspondance.** Le départ par correspondance utilise les résultats d'une analyse de correspondance sur les données inversées (similitudes au lieu des différences) avec une normalisation symétrique des écarts des lignes et des colonnes.
- **Barycentres.** La procédure démarre avec le positionnement des objets de lignes dans la configuration à l'aide de la décomposition de la valeur propre. Les objets de colonnes sont ensuite positionnés dans le barycentre des choix spécifiés. Pour le nombre de choix, spécifiez un entier positif entre 1 et le nombre de variables de proximité.

- **Départs aléatoires multiples :** Les solutions sont calculées pour plusieurs configurations initiales sélectionnées de manière aléatoire et celle présentant la mesure de stress pénalisée la plus basse représente la meilleure.
- **Personnalisé :** Vous sélectionnez les variables contenant les coordonnées de votre configuration initiale. Le nombre de variables sélectionnées doit être égal au nombre de dimensions spécifié, la première variable correspondant aux coordonnées sur la dimension 1, la seconde correspondant aux coordonnées sur la dimension 2, etc. Le nombre d'observations dans chaque variable doit être égal au nombre combiné d'objets de lignes et de colonnes. Les coordonnées des lignes et des colonnes doivent être empilées, avec les coordonnées des colonnes à la suite des coordonnées des lignes.

Critères d'itération : Spécifiez les valeurs des critères d'itération.

- **Convergence du stress :** L'algorithme interrompt son itération lorsque la différence relative des valeurs des mesures de stress pénalisé consécutives est inférieure au nombre spécifié ici, lequel ne peut pas être négatif.
- **Stress minimum :** L'algorithme s'arrête lorsque la mesure de stress pénalisé est inférieure au nombre spécifié ici, qui ne peut pas être négatif.
- **Nombre maximum d'itérations :** L'algorithme exécute le nombre d'itérations spécifié ici, à moins que l'un des critères ci-dessus ne soit déjà satisfait.

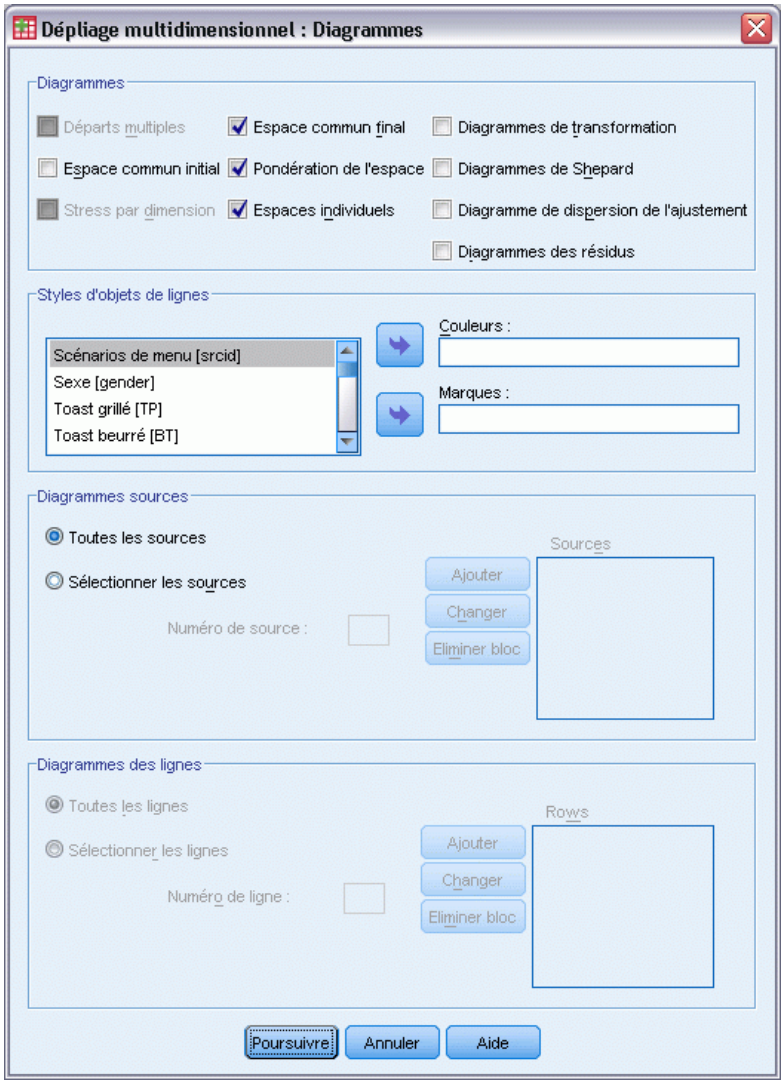
Terme de pénalité. L'algorithme tente de minimiser la mesure de stress pénalisé, qui est une mesure de la qualité d'ajustement égale au produit du Stress-I de Kruskal par un terme de pénalité basé sur le coefficient de variation des proximités transformées. Ces contrôles vous permettent de définir l'intensité et l'intervalle du terme du pénalité.

- **Intensité.** Plus la valeur du paramètre d'intensité est petite, plus la pénalité est intense. Spécifiez une valeur comprise entre 0,0 et 1,0.
- **Intervalle :** Ce paramètres définit le moment auquel la pénalité devient active. Si vous le définissez sur 0,0, la pénalité est inactive. L'augmentation de la valeur entraîne la recherche par l'algorithme d'une solution présentant une plus grande variation parmis les proximités transformées. Spécifiez une valeur non négative.

Diagrammes de dépliage multidimensionnel

La boîte de dialogue Diagrammes vous permet de spécifier quels diagrammes doivent être produits.

Figure 8-5
Boîte de dialogue Diagrammes



Diagrammes : Les diagrammes suivants sont disponibles :

- **Départs multiples.** Affiche un histogramme empilé de la mesure de stress pénalisé, affichant à la fois le stress et la pénalité.
- **Espace commun initial.** Affiche une matrice de diagramme de dispersion des coordonnées de l'espace commun initial.
- **Stress par dimension.** Produit un diagramme en lignes de la mesure du stress pénalisé en fonction des dimensions. Ce diagramme est uniquement généré si le nombre maximum de dimensions est supérieur au nombre minimum.
- **Espace commun final.** Une matrice de diagramme de dispersion des coordonnées de l'espace commun est affiché.

- **Pondération de l'espace.** Un diagramme de dispersion est produit à partir des pondérations des espaces individuels. Cela est uniquement possible si l'un des modèles de différences individuels est spécifié dans la boîte de dialogue Modèle. Pour le modèle Euclidien pondéré, les pondérations de toutes les sources sont affichées dans un diagramme avec une dimension sur chaque axe. Pour le modèle Euclidien généralisé, un diagramme est produit par dimension, indiquant à la fois la rotation et sa pondération pour chaque source.
- **Espaces individuels :** Une matrice de diagramme de dispersion des coordonnées de l'espace individuel de chaque source est affichée. Cela est uniquement possible si l'un des modèles de différences individuels est spécifié dans la boîte de dialogue Modèle.
- **Diagrammes de transformation :** Un diagramme de dispersion est généré à partir des proximités d'origine par opposition aux proximités transformées. Selon l'application des transformations, une couleur distincte est assignée à chaque ligne ou source. Une transformation inconditionnelle génère une seule couleur.
- **Diagrammes de Shepard.** Les proximités d'origine en fonction des proximités transformées et des distances. Les distances sont représentées par des points et les proximités transformées par une ligne. Selon l'application des transformations, une ligne distincte est générée pour chaque ligne ou source. Une transformation inconditionnelle génère une seule ligne.
- **Diagramme de dispersion de l'ajustement.** Un diagramme de dispersion des proximités transformées en fonction des distances est affiché. Une couleur distincte est assignée à chaque source lorsque plusieurs sources sont spécifiées.
- **Diagrammes des résidus.** Un diagramme de dispersion des proximités transformées en fonction des résidus (proximités transformées moins les distances) est affiché. Une couleur distincte est assignée à chaque source lorsque plusieurs sources sont spécifiées.

Styles d'objets de lignes. Les styles vous apportent un contrôle supplémentaire pour l'affichage des objets de lignes dans les diagrammes. Les valeurs des variables de couleurs facultatives sont utilisées pour passer en revue toutes les couleurs. Les valeurs des variables de marques facultatives sont utilisées pour passer en revue toutes les marques possibles.

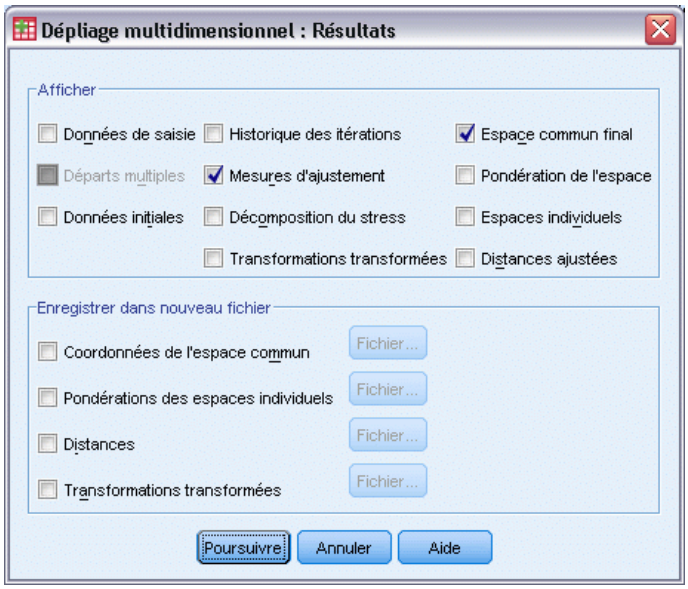
Diagrammes sources. Pour les espaces individuels, diagrammes de dispersion de l'ajustement et Diagrammes des résidus, ainsi que pour les diagrammes de transformations— et diagrammes de Shepard—, si les transformations sont appliquées par la source, vous pouvez spécifier les sources pour lesquelles les diagrammes doivent être générés. Les numéros de source entrés doivent être des valeurs de la variable de sources spécifiée dans la boîte de dialogue principale et être classés de 1 jusqu'au nombre de sources.

Diagrammes des lignes. Si des transformations sont appliquées par lignes, vous pouvez spécifier la ligne pour laquelle les diagrammes doivent être générés pour les Diagrammes de transformations et les Diagrammes de Shepard. Les numéros de lignes doivent être compris entre 1 et le nombre de lignes.

Résultat du dépliage multidimensionnel

La boîte de dialogue Résultat vous permet de contrôler les résultats affichés et d'en enregistrer certains pour séparer des fichiers.

Figure 8-6
Résultat



Afficher : Sélectionnez l'un des éléments suivants à afficher :

- **Données de saisie :** Inclut les proximités d'origine et, si elles existent, les pondérations de données, la configuration initiale et les coordonnées fixées.
- **Départs multiples.** Affiche le générateur de nombre aléatoire et la valeur du stress pénalisé de chaque départ aléatoire.
- **Données initiales.** Affiche les coordonnées de l'espace commun initial.
- **Historique des itérations :** Affiche l'historique des itérations de l'algorithme principal.
- **Mesures d'ajustement.**Affiche différentes mesures. Le tableau contient plusieurs mesures de qualité de l'ajustement, de défaut de l'ajustement, de corrélation, de variation et de non-dégénérescence.
- **Décomposition du stress :** Affiche la mesure du stress pénalisé de la décomposition d'un objet, d'une ligne ou d'une source, y compris les moyennes et les écarts-types de la ligne, de la colonne ou de la source.
- **Transformations transformées :** Affiche les proximités transformées.
- **Espace commun final.** Affiche les coordonnées de l'espace commun.
- **Pondération de l'espace.** Affiche les pondérations de l'espace individuel. Cette option est uniquement disponible lorsque l'un des modèles de différences individuelles est spécifié. En fonction du modèle, les pondérations des espaces sont décomposées en pondérations de rotation et de dimension, lesquelles sont également affichées.
- **Espaces individuels :** Les coordonnées de l'espace individuel sont affichées. Cette option est uniquement disponible lorsque l'un des modèles de différences individuelles est spécifié.
- **Distances ajustées.** Affiche les distances entre les objets de la configuration.

Enregistrer dans nouveau fichier : Vous pouvez enregistrer les coordonnées de l'espace commun, les pondérations des espaces individuels, les distances et les proximités transformées dans des fichiers de données IBM® SPSS® Statistics distincts.

Fonctionnalités supplémentaires de la commande PREFSCAL

Vous pouvez personnaliser l'analyse des proximités du dépliage multidimensionnel en collant vos sélections dans une fenêtre de syntaxe et en modifiant la syntaxe de commande `PROXSCAL` résultante. Le langage de syntaxe de commande vous permet aussi de :

- Spécifiez plusieurs listes sources pour les Espaces individuels, les Diagrammes de dispersion de l'ajustement et les Diagrammes des résidus, ainsi que pour les Diagrammes de transformations et les Diagrammes de Shepard dans le cas de transformations conditionnelles d'une matrice, lorsque plusieurs sources sont disponibles (avec la sous-commande `PLOT`).
- Spécifiez plusieurs listes de lignes pour les Diagrammes de transformations et les Diagrammes de Shepard dans le cas de transformations conditionnelles par lignes (avec la sous-commande `PLOT`).
- Spécifiez un nombre de colonnes au lieu d'une variable ID de colonne (avec la commande `INPUT`).
- Spécifiez un nombre de sources au lieu d'une variable ID de source (avec la commande `INPUT`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Partie II: Exemples

Chapitre
9

Régression nominale

L'objectif de la régression nominale avec codage optimal est de décrire la relation entre une variable de réponse et un groupe de variables prédites. La quantification de cette relation permet de prévoir les valeurs de la réponse pour n'importe quelle combinaison de variables prédites.

Dans ce chapitre, deux exemples illustrent les analyses impliquées dans la régression avec codage optimal. Le premier exemple utilise un ensemble de données réduit pour illustrer les concepts de base. Le second exemple utilise un ensemble plus vaste de variables et d'observations dans une application pratique.

Exemple : Données relatives à la shampouineuse

Dans un exemple courant (Green et Wind, 1973), une société intéressée par la commercialisation d'une nouvelle shampouineuse souhaite examiner l'influence de cinq critères sur la préférence du consommateur : la conception du conditionnement, la marque, le prix, une étiquette *Economique* et une garantie Satisfait ou remboursé. Il existe trois niveaux de critère pour la conception du conditionnement, suivant l'emplacement de l'applicateur, trois marques (*K2R*, *Glory* et *Bissell*), trois niveaux de prix et deux niveaux (non ou oui) pour chacun des deux derniers critères. Le tableau suivant indique les variables utilisées dans l'étude sur la shampouineuse, ainsi que les valeurs et étiquettes correspondantes.

Table 9-1
Variables explicatives de l'étude sur la shampouineuse

Nom de variable	l'étiquette Variable	Etiquette de valeur
conditionnement	Conception du conditionnement	A*, B*, C*
marque	Nom de la marque	K2R, Glory, Bissell
prix	Prix	$1.19, $1.39, $1.59
étiquette	Etiquette *Economique*	Non, oui
argent	Garantie Satisfait ou remboursé	Non, oui

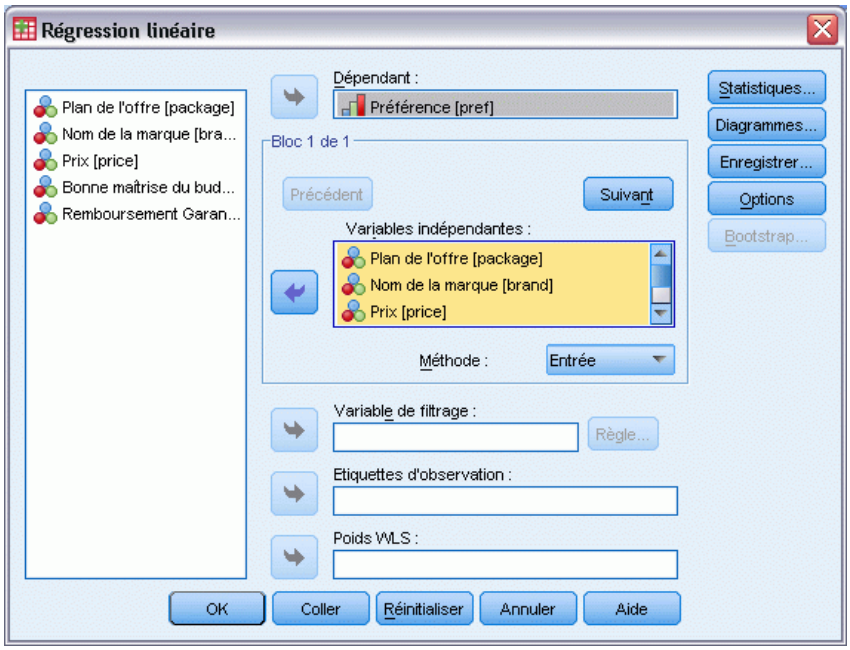
Dix consommateurs classent 22 profils définis par ces critères. La variable *Préférence* indique le classement des rangs moyens de chaque profil. Un rang faible correspond à une préférence élevée. Cette variable reflète une mesure globale de préférence pour chaque profil. A l'aide de la régression nominale, vous allez examiner le rapport entre la préférence et les cinq critères. Cet ensemble de données est disponible dans *carpet.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*.

Analyse de régression linéaire standard

- Pour générer le résultat d'une régression linéaire standard, dans les menus, choisissez : Analyse > Régression > Linéaire

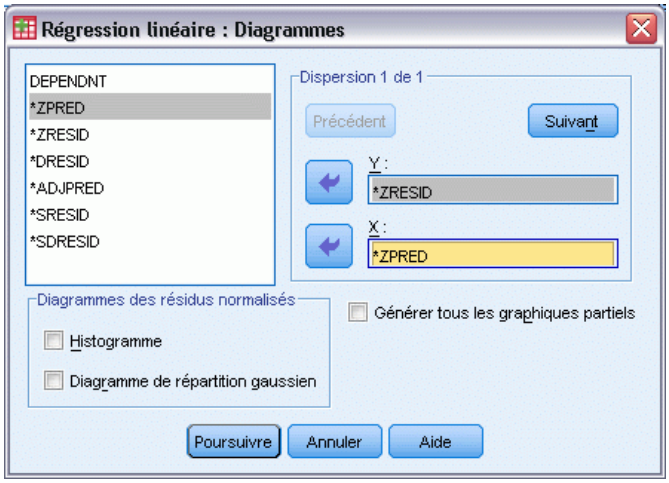
  Remarque : Cette fonction nécessite l'option Statistiques de base.

Figure 9-1
Boîte de dialogue Régression linéaire



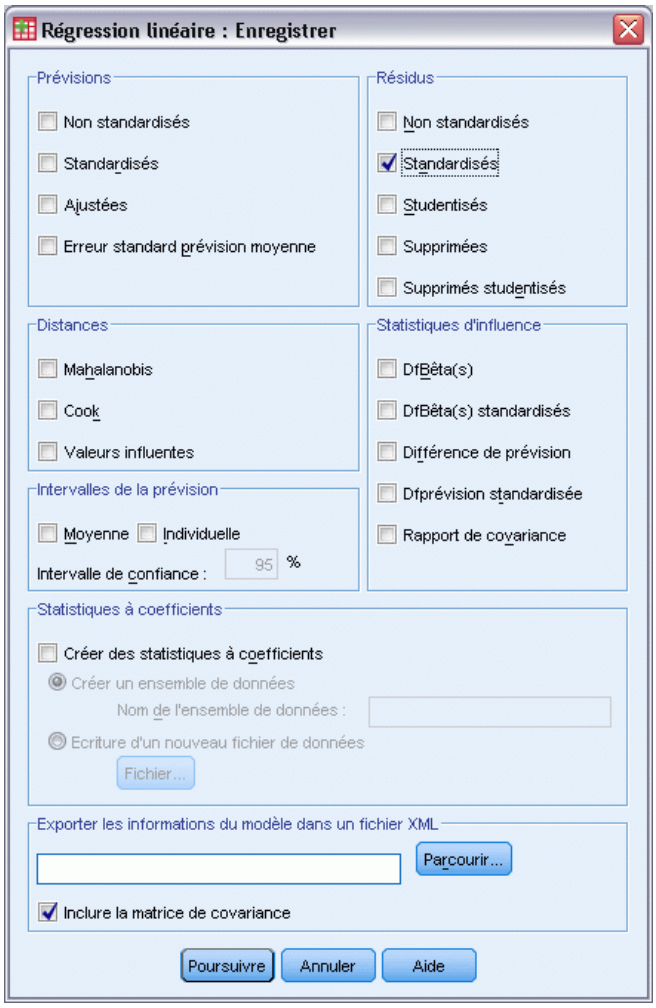
- Sélectionnez l'option *Préférence* comme variable dépendante.
- Sélectionnez comme variables indépendantes les options allant de *Conception du conditionnement* à *Garantie satisfait ou remboursé*.
- Cliquez sur Diagrammes.

Figure 9-2
Boîte de dialogue Diagrammes



- Sélectionnez l'option **ZRESID* comme variable de l'axe *y*.
- Sélectionnez l'option **ZPRED* comme variable de l'axe *x*.
- Cliquez sur Poursuivre.
- Cliquez sur le bouton Enregistrer dans la boîte de dialogue Régression linéaire.

Figure 9-3
Boîte de dialogue Enregistrer



- Sélectionnez l'option Standardisés dans le groupe Résidus.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Régression linéaire.

Récapitulatif des modèles

Figure 9-4
Récapitulatif du modèle de la régression linéaire standard

Modele	R	R deux	R deux ajusté	Erreur std de l'estim.
1	.841[a]	.707	.615	3.99810

a. Hauteur de base d'inversion - Quantification Garantie satisfait ou remboursé Prix

L'approche standard de la description des relations dans ce cas de figure est la régression linéaire. La mesure la plus courante de l'ajustement d'un modèle de régression aux données est R^2. Cette statistique indique la quantité de variance, dans la réponse, explicable par la combinaison pondérée des variables prédites. Plus la mesure R^2 tend vers 1, meilleur est l'ajustement du modèle. La régression de la variable *Préférence* sur les cinq variables prédites aboutit à une mesure R^2 de 0,707, ce qui indique qu'environ 71 % de la variance dans les rangs de préférence sont explicables par les variables prédites dans la régression linéaire.

Coefficients

Le tableau répertorie les coefficients standardisés. Le signe du coefficient indique si la réponse prévue augmente ou diminue lorsque la variable prédite augmente, toutes les autres variables prédites étant constantes. Dans le cas des données qualitatives, le codage des modalités détermine la signification de l'augmentation d'une variable prédite. Par exemple, une augmentation de la variable *Garantie satisfait ou remboursé*, *Conception du conditionnement* ou *Etiquette Economique* provoque une diminution du rang de préférence prévue. La variable *Garantie satisfait ou remboursé* a le code 1 pour *aucune garantie Satisfait ou remboursé* et 2 pour la *garantie Satisfait ou remboursé*. Une augmentation de la variable *Garantie satisfait ou remboursé* correspond à l'ajout d'une garantie Satisfait ou remboursé. Par conséquent, l'ajout d'une garantie Satisfait ou remboursé réduit le rang de préférence prévue, ce qui correspond à une augmentation de la préférence prévue.

Figure 9-5
Coefficients de régression

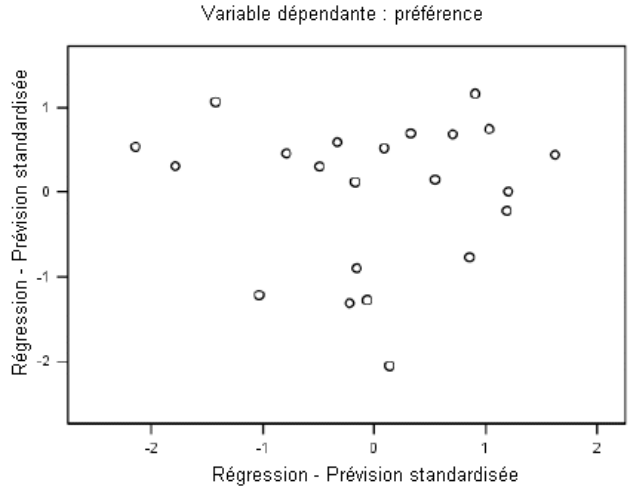
Modele	Coefficients non standardisés		Coefficients standardisés		
	B	Std. Error	Beta	t	Sig.
(Constant)	22.529	5.177		4.352	.000
Conception du conditionnement	-4.159	1.036	-.560	-4.015	.001
Marque	.429	1.054	.056	.407	.689
Prix	2.703	1.009	.366	2.681	.016
Etiquette Economique	-4.314	1.780	-.330	-2.423	.028
Garantie satisfait ou remboursé	-2.779	1.921	-.197	-1.447	.167

La valeur du coefficient reflète la quantité de modifications survenues dans le rang de préférence prévue. A partir de coefficients standardisés, les interprétations sont basées sur les écarts-types des variables. Chaque coefficient indique le nombre d'écarts-types que la réponse prévue remplace par un écart-type de 1 dans une variable prédite, toutes les autres variables prédites demeurant constantes. Par exemple, une modification d'écart-type de 1 dans la variable *Nom de marque* provoque une augmentation d'écart-type de 0,056 dans la préférence prévue. L'écart-type de la

variable *Préférence* étant 6,44, la variable *Préférence* augmente de $0,056 \times 6,44 = 0,361$. Les modifications de la variable *Conception du conditionnement* provoquent les changements les plus importants dans la préférence prévue.

Diagrammes de dispersion des résidus

Figure 9-6
Résidus et prévisions

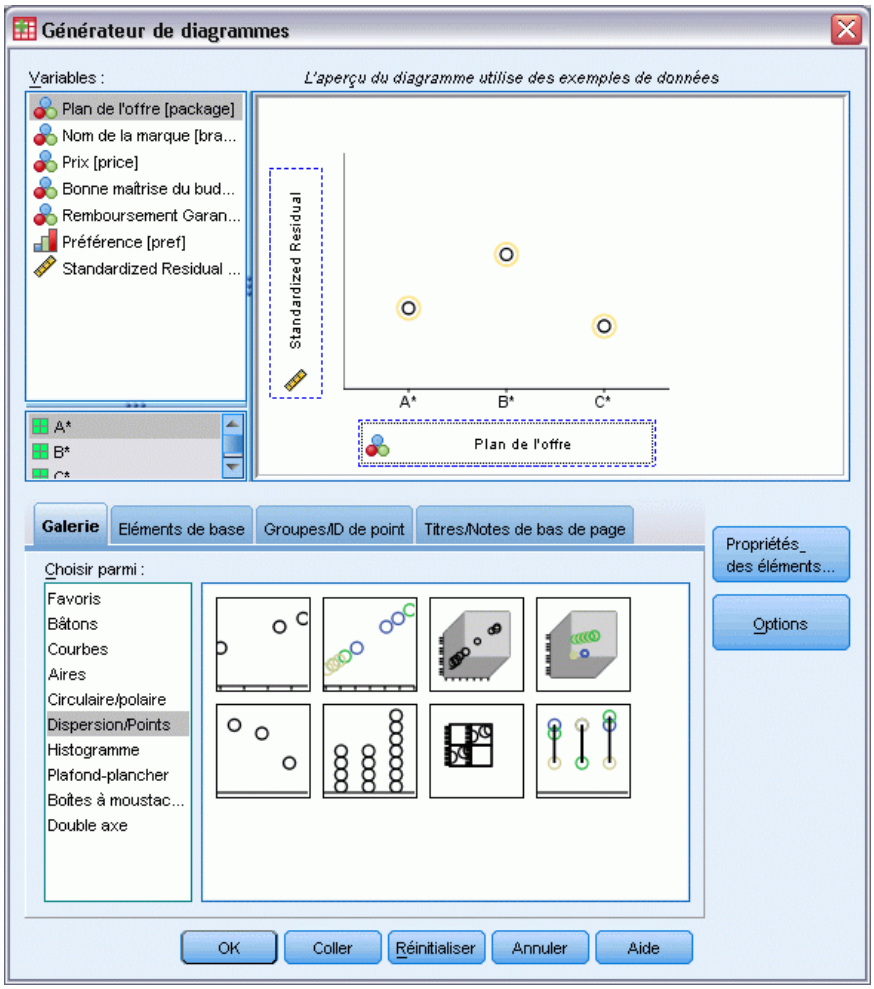


Les résidus standardisés sont représentés par rapport aux prévisions standardisées. Aucun motif ne doit être présent si le modèle s'ajuste correctement. Ici, vous pouvez constater une forme en U dans laquelle les prévisions standardisées basses et élevées possèdent des résidus positifs. Les prévisions standardisées proches de 0 tendent à détenir des résidus négatifs.

- Pour générer un diagramme de dispersion des résidus à partir de la variable prédite *Conception du conditionnement*, choisissez les options suivantes dans les menus :

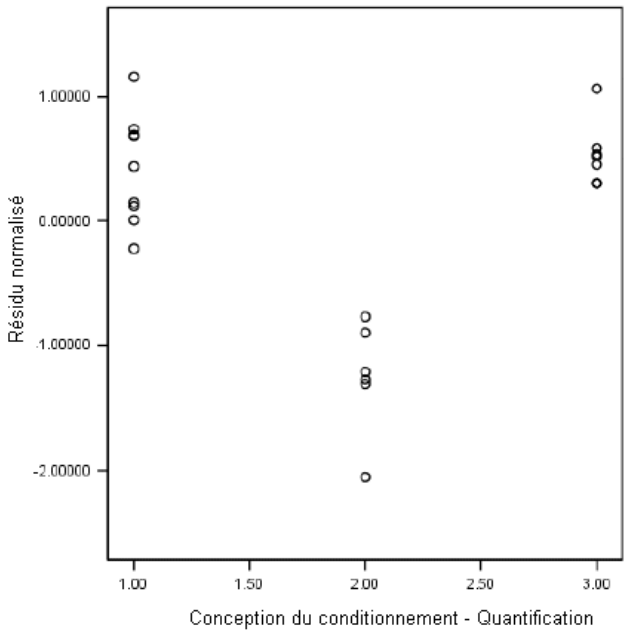
  Graphes > Générateur de diagrammes...

Figure 9-7
Générateur de diagrammes



- Sélectionnez la galerie Dispersion/Points, puis choisissez Dispersion simple.
- Sélectionnez l'option *Résidus standardisés* comme variable de l'axe *y* et l'option *Conception du conditionnement* comme variable de l'axe *x*.
- Cliquez sur OK.

Figure 9-8
Résidus et conception du conditionnement



La forme en U est davantage prononcée dans le diagramme des résidus standardisés établi par rapport au conditionnement. Chaque résidu de la conception B* est négatif, tandis que tous les résidus, à l'exception d'un seul, sont positifs pour les deux autres conceptions. Etant donné que le modèle de régression linéaire ajuste un paramètre par variable, la relation ne peut pas être capturée par l'approche standard.

Analyse de régression nominale

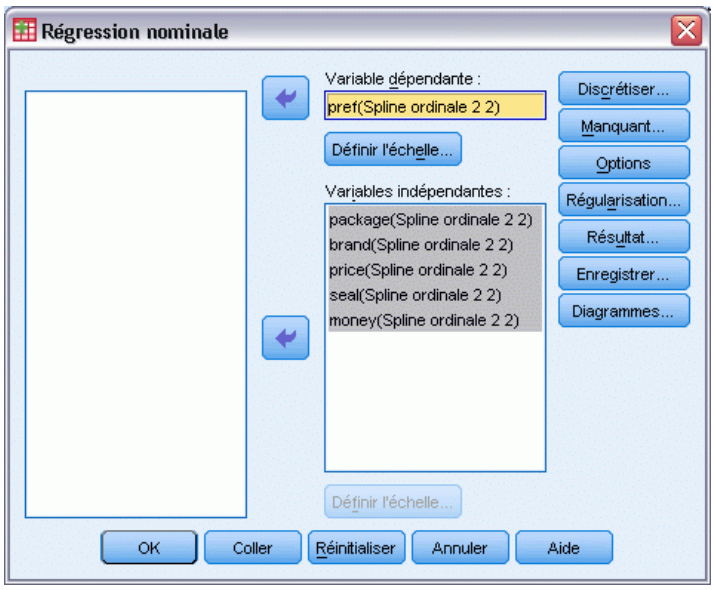
La nature qualitative des variables et la relation non linéaire entre les variables *Préférence* et *Conception du conditionnement* laissent supposer que la régression sur des quantifications optimales peut s'avérer meilleure que la régression standard. La forme en U des diagrammes résiduels indique la nécessité de recourir à un traitement nominal de la variable *Conception du conditionnement*. Toutes les autres variables prédites seront traitées au niveau du codage numérique.

La variable de réponse garantit une considération particulière. Vous souhaitez prévoir les valeurs de la variable *Préférence*. Par conséquent, il est souhaitable de récupérer autant de propriétés que possible de ses modalités. L'utilisation d'un niveau de codage ordinal ou nominal ignore les différences entre les modalités de réponse. Toutefois, la transformation linéaire des modalités de réponse préserve les différences de modalité. Par conséquent, le codage numérique de la réponse est généralement privilégié et sera employé ici.

Exécution de l'analyse

- Pour exécuter une analyse de régression nominale, choisissez les options suivantes dans les menus : Analyse > Régression > Codage optimal (CATREG)…

Figure 9-9
Boîte de dialogue Régression nominale



- Sélectionnez l'option *Préférence* comme variable dépendante.
- Sélectionnez comme variables indépendantes les options allant de *Conception du conditionnement* à *Garantie satisfait ou remboursé*.
- Sélectionnez l'option *Préférence*, puis cliquez sur Définir l'échelle.

Figure 9-10
Boîte de dialogue Définir l'échelle

- Sélectionnez l'option Numérique comme niveau de codage optimal.
- Cliquez sur Poursuivre.

- Sélectionnez l'option *Conception du conditionnement*, puis cliquez sur Définir l'échelle dans la boîte de dialogue Régression nominale.

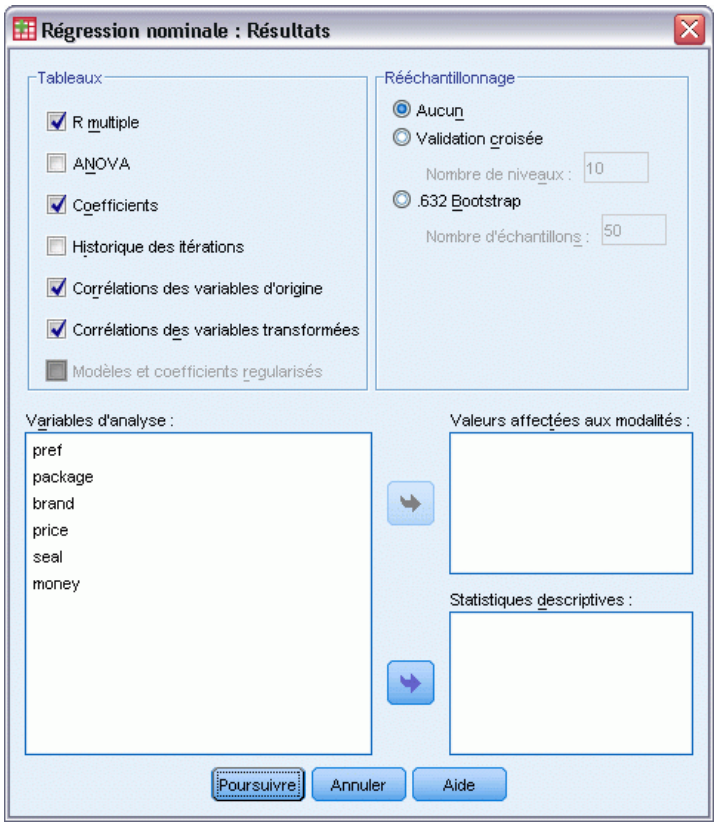
Figure 9-11
Boîte de dialogue Définir l'échelle

- Sélectionnez l'option Nominal comme niveau de codage optimal.
- Cliquez sur Poursuivre.
- Sélectionnez les options allant de *Nom de marque* à *Garantie satisfait ou remboursé*, puis cliquez sur Définir l'échelle dans la boîte de dialogue Régression nominale.

Figure 9-12
Boîte de dialogue Définir l'échelle

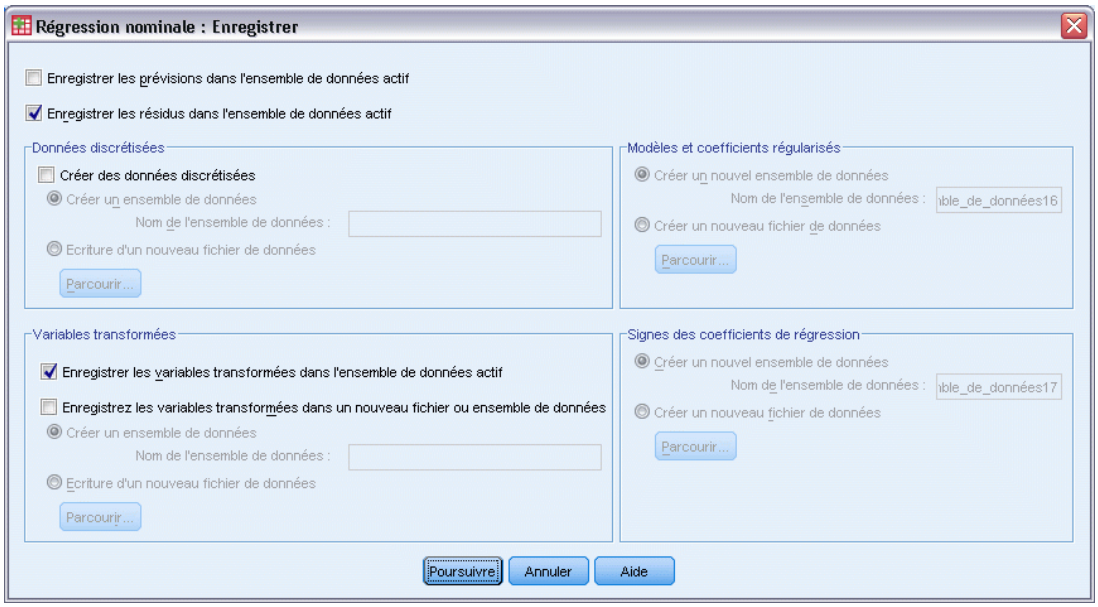
- Sélectionnez l'option Numérique comme niveau de codage optimal.
- Cliquez sur Poursuivre.
- Cliquez sur Résultat dans la boîte de dialogue Régression nominale.

Figure 9-13
Résultat



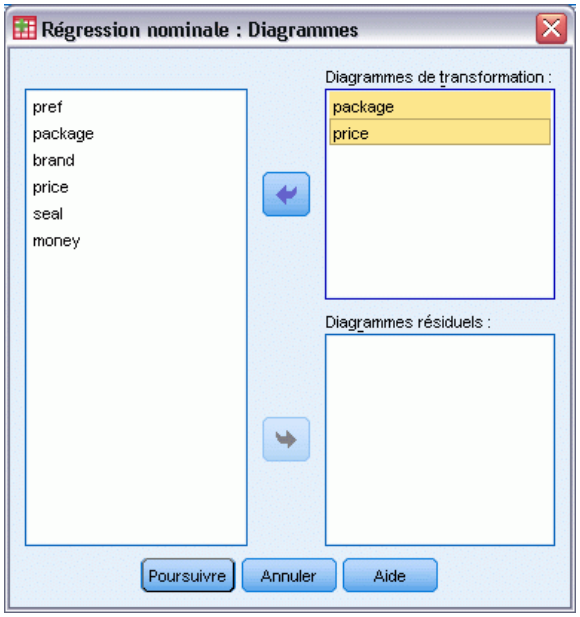
- Sélectionnez les options Corrélations des variables d'origine et Corrélations des variables transformées.
- Désélectionnez l'option ANOVA.
- Cliquez sur Poursuivre.
- Cliquez sur le bouton Enregistrer dans la boîte de dialogue Régression nominale.

Figure 9-14
Boîte de dialogue Enregistrer



- Sélectionnez Enregistrer les résidus dans l'ensemble de données actif.
- Sélectionnez Enregistrer les variables transformées dans l'ensemble de données actif dans le groupe Variables transformées.
- Cliquez sur Poursuivre.
- Cliquez sur Diagrammes dans la boîte de dialogue Régression nominale.

Figure 9-15
Boîte de dialogue Diagrammes



- Appliquez la création de diagrammes de transformation au conditionnement (*conditionnement*) et au prix (*prix*).
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Régression nominale.

Intercorrélations

Les intercorrélations existant entre les variables prédites permettent d'identifier la multicolinéarité dans la régression. Les variables en étroite corrélation aboutissent à des estimations de régression instables. Toutefois, en raison de leur corrélation élevée, l'omission de l'une d'elles dans le modèle n'affecte que très légèrement la prévision. Dans la réponse, la variance explicable par la variable omise est expliquée par la variable corrélée restante. Toutefois, les corrélations simples sont sensibles aux valeurs éloignées et, en outre, ne peuvent pas identifier la multicolinéarité en raison d'une corrélation élevée entre une variable prédite et une combinaison d'autres variables prédites.

Figure 9-16
Corrélations des variables prédites initiales

	Conception du conditionnement	Marque	Prix	Etiquette Economique	Garantie satisfait ou remboursé
Conception du condt.	1.000	-.189	-.126	.081	.066
Marque	-.189	1.000	.065	-.042	-.034
Prix	-.126	.065	1.000	.000	.000
Etiquette Economique	.081	-.042	.000	1.000	-.039
Garantie satisfait ou rem	.066	-.034	.000	-.039	1.000
Dimension	1	2	3	4	5
Valeur	1.291	1.038	.980	.905	.785

Figure 9-17
Corrélations des variables prédites transformées

	Conception du conditionnement	Marque	Prix	Etiquette Economique	Garantie satisfait ou remboursé
Conception du condt.	1.000	-.156	-.089	.032	.102
Marque	-.156	1.000	.065	-.042	-.034
Prix	-.089	.065	1.000	.000	.000
Etiquette Economique	.032	-.042	.000	1.000	-.039
Garantie satisfait ou rem.	.102	-.034	.000	-.039	1.000
Dimension	1	2	3	4	5
Valeur	1.248	1.043	.983	.905	.821

Les intercorrélations des variables prédites pour les variables prédites non transformées et transformées sont affichées. Toutes les valeurs sont proches de 0, ce qui indique que la multicolinéarité entre les différentes variables n'est pas préoccupante.

Les seules corrélations qui changent concernent la variable *Conception du conditionnement*. Etant donné que toutes les autres variables prédites sont traitées de manière numérique, les différences entre les modalités et leur ordre sont conservées pour ces variables. Par conséquent, les corrélations ne peuvent pas changer.

Qualité de l'ajustement et coefficients

La procédure de régression nominale génère une mesure R^2 de 0,948, indiquant que près de 95 % de la variance dans les rangs de préférence transformée sont explicables par la régression sur les variables prédites transformées de manière optimale. La transformation des variables prédites améliore l'ajustement par rapport à l'approche standard.

Figure 9-18
Récapitulatif du modèle de la régression nominale

Multiple	R deux	R deux ajusté
.974	.948	.927

Variable dépendante : préférence
Variables indépendantes - Conception du conditionnement Marque Prix Etiquette Economique

Le tableau suivant répertorie les coefficients de régression standardisés. Etant donné que la régression nominale standardise les variables, seuls les coefficients standardisés sont indiqués. Ces valeurs sont divisées par les erreurs standard correspondantes, aboutissant à un test *F* pour chaque variable. Toutefois, le test de chaque variable dépend des autres variables iprédites présentes dans le modèle. En d'autres termes, le test détermine si l'omission d'une variable prédite dans le modèle, alors que toutes les autres y sont présentes, détériore sensiblement les capacités prévisionnelles de celui-ci. Ces valeurs ne doivent pas être utilisées pour omettre simultanément

plusieurs variables dans un modèle ultérieur. En outre, l'utilisation de moindres carrés alternés optimise les quantifications, ce qui implique que ces tests doivent être interprétés avec prudence.

Figure 9-19
Coefficients standardisés des variables prédites transformées

	Coefficients standardisés				
	Bêta	Erreur standard	ddl	F	Sig.
package	-.748	.060	2	155.289	.000
brand	.045	.060	1	.578	.459
price	.371	.059	1	39.312	.000
seal	-.350	.059	1	35.299	.000
money	-.159	.059	1	7.175	.017

Variable dépendante : pref

Le coefficient le plus élevé concerne la variable *Conception du conditionnement*. Une augmentation d'écart-type de 1 dans la variable *Conception du conditionnement* provoque une diminution d'écart-type de 0,748 dans le rang de préférence prévue. Toutefois, la variable *Conception du conditionnement* étant traitée de manière nominale, il n'est pas nécessaire qu'une augmentation des quantifications corresponde à une augmentation des codes de modalité initiaux.

Les coefficients standardisés sont souvent interprétés comme révélateurs de l'importance de chaque variable prédite. Toutefois, les coefficients de régression ne peuvent pas décrire entièrement l'impact d'une variable prédite ou les relations entre les variables prédites. Vous devez recourir à d'autres statistiques, conjointement aux coefficients standardisés, pour explorer complètement les effets des variables prédites.

Corrélations et importance

Le seul examen des coefficients de régression est insuffisant pour interpréter les contributions des variables prédites à la régression. En outre, les corrélations, les corrélations partielles et les mesures doivent être examinées. Le tableau suivant illustre les mesures de corrélation pour chaque variable.

La corrélation simple est la corrélation existant entre la variable prédite et la réponse transformées. Pour ces données, la corrélation la plus élevée concerne la variable *Conception du conditionnement*. Toutefois, si vous pouvez expliquer une partie de la variation dans la variable prédite ou dans la réponse, vous obtenez une meilleure représentation de la qualité de la variable prédite.

Figure 9-20
Corrélations simples, mesures et corrélations partielles (variables transformées)

	Corrélations			Importance	Tolérance	
	Ordre zéro	Partielle	Partie		Après transformation	Avant transformation
package	.816	.955	.733	.644	.959	.942
brand	.206	.193	.045	.010	.971	.961
price	.440	.851	.369	.172	.989	.982
seal	-.370	-.838	-.349	.137	.996	.991
money	-.223	-.569	-.158	.037	.987	.993

Variable dépendante : pref

D'autres variables du modèle peuvent fausser l'effet d'une variable prédite donnée lors de la prévision de la réponse. Le coefficient de corrélation partielle supprime les effets linéaires des autres variables prédites de la réponse et de la variable prédite. Cette mesure équivaut à la corrélation entre les résidus issus de la régression de la variable prédite sur les autres variables prédites et ceux issus de la régression de la réponse sur les autres variables prédites. La corrélation partielle carrée correspond à la proportion de la variance expliquée par rapport à la variance résiduelle de la réponse, après suppression des effets des autres variables. Par exemple, la corrélation partielle de la variable *Conception du conditionnement* est égale à –0,955. Une fois les effets des autres variables supprimés, la variable *Conception du conditionnement* explique 91 % ($(–0,955)^2 = 0,91$) de la variation des rangs de préférence. Les variables *Prix* et *Etiquette Economique* expliquent également une large partie de la variance si les effets des autres variables sont supprimés.

Au lieu de supprimer les effets de variables de la réponse et d'une variable prédite, vous pouvez vous contenter de les supprimer de la variable prédite. La corrélation entre la réponse et les résidus issus de la régression d'une variable prédite sur les autres variables prédites est la mesure. L'élévation au carré de cette valeur donne une mesure de la proportion de variance expliquée par rapport à la variance totale de la réponse. Si vous supprimez les effets des variables *Nom de marque*, *Etiquette Economique*, *Garantie satisfait ou remboursé* et *Prix* de la variable *Conception du conditionnement*, la partie restante de cette dernière explique 54 % ($(–0,733)^2 = 0,54$) de la variation des rangs de préférence.

Importance

Outre les coefficients de régression et les corrélations, la mesure d'importance relative de Pratt(Pratt, 1987) facilite l'interprétation des contributions des variables prédites à la régression. Des importances élevées par rapport aux autres importances correspondent à des variables prédites cruciales pour la régression. De même, la présence de variables suppressives est signalée par une importance faible dans le cas d'une variable dont la taille du coefficient est similaire à celle du coefficient des variables prédites importantes.

Par opposition aux coefficients de régression, cette mesure définit l'importance des variables prédites de manière additive, c'est-à-dire que l'importance d'un groupe de variables prédites est la somme de l'importance de chacune de ces variables. La mesure de Pratt équivaut au produit du coefficient de régression et de la corrélation simple d'une variable prédite. Ces produits s'ajoutent à R^2, ils sont donc divisés par R^2, ce qui génère une somme égale à 1. Le groupe de variables prédites *Conception du conditionnement* et *Nom de la marque*, par exemple, ont une importance de 0,654. L'importance la plus élevée correspond à la variable *Conception du conditionnement*, les variables *Conception du conditionnement*, *Prix* et *Etiquette Economique* représentant 95 % de l'importance de cette combinaison de variables prédites.

Multicolinéarité

Les corrélations élevées entre variables prédites réduisent sensiblement la stabilité d'un modèle de régression. Les variables prédites corrélées aboutissent à des estimations de paramètre instables. La tolérance reflète le degré de linéarité de la relation entre les variables indépendantes. Cette mesure constitue la proportion de la variance d'une variable qui n'est pas expliquée par d'autres variables indépendantes de l'équation. Si les autres variables prédites peuvent expliquer une large partie de la variance d'une variable prédite, celle-ci n'est pas requise dans le modèle. Une

valeur de tolérance proche de 1 indique que la variable ne peut pas être prévue très correctement à partir des autres variables prédites. En revanche, une variable à très faible tolérance apporte peu d'informations à un modèle et peut entraîner des problèmes de calcul. En outre, des valeurs négatives élevées de la mesure d'importance de Pratt indiquent une multicolinéarité.

Toutes les mesures de tolérance sont très élevées. Aucune des variables prédites n'est prévue très correctement par les autres variables prédites et il n'y a pas de multicolinéarité.

Diagrammes de transformation

La représentation des valeurs de modalité initiales par rapport aux quantifications correspondantes peut mettre en évidence des tendances qu'une liste de quantifications ne laisse pas forcément transparaître. Ces types de diagramme sont communément appelés Diagrammes de transformation. Vous devez prêter une attention particulière aux modalités qui reçoivent des quantifications similaires. Ces modalités affectent la réponse prévue de la même manière. Toutefois, le type de transformation détermine l'aspect de base du diagramme.

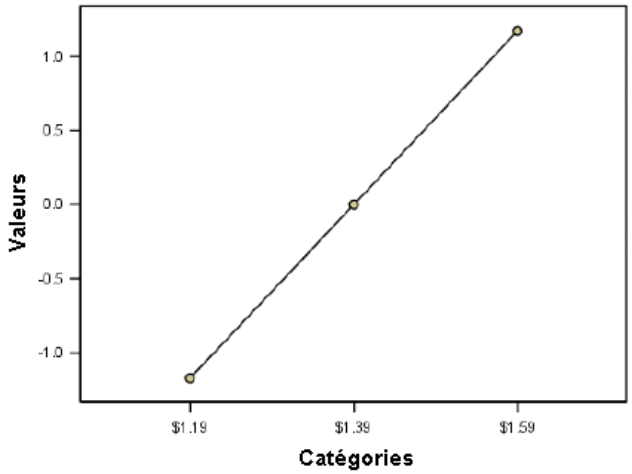
Les variables traitées en tant que données numériques aboutissent à une relation linéaire entre les quantifications et les modalités initiales, qui apparaît sous la forme d'une ligne droite dans le diagramme de transformation. L'ordre des modalités initiales et leurs différences sont conservés dans les quantifications.

L'ordre des quantifications des variables traitées en tant que données ordinales correspond à l'ordre des modalités initiales. Toutefois, les différences entre les modalités ne sont pas conservées. Par conséquent, le diagramme de transformation est non décroissant, mais n'est pas nécessairement une ligne droite. Si des modalités consécutives correspondent à des quantifications similaires, la distinction entre elles peut s'avérer superflue et les modalités peuvent être combinées. Ces modalités se traduisent par un palier dans le diagramme de transformation. Toutefois, ce motif peut également résulter de l'application d'une structure ordinale à une variable à traiter comme donnée nominale. Si un traitement nominal ultérieur de la variable met en évidence le même motif, la combinaison des modalités est garantie. En outre, si les quantifications d'une variable traitée en tant que donnée ordinale s'alignent sur une ligne droite, une transformation numérique peut s'avérer plus appropriée.

Dans le cas des variables traitées en tant que données nominales, l'ordre des modalités le long de l'axe horizontal correspond à l'ordre des codes utilisés pour les représenter. Les interprétations de l'ordre des modalités ou de la distance les séparant sont sans fondement. Le diagramme peut prendre toute forme non linéaire ou linéaire. En présence d'une tendance ascendante, un traitement ordinal doit être tenté. Si le diagramme de transformation nominale montre une tendance linéaire, une transformation numérique peut s'avérer plus appropriée.

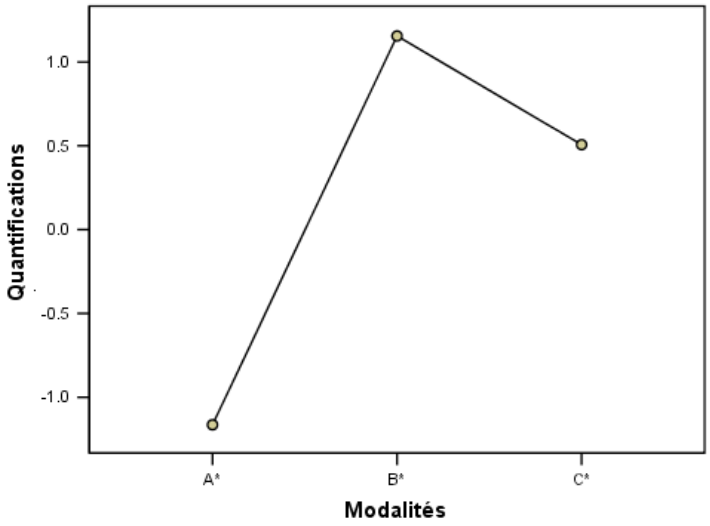
Le schéma ci-après illustre le diagramme de transformation de la variable *Prix*, qui a été traitée en tant que donnée numérique. L'ordre des modalités le long de la ligne droite correspond à l'ordre des modalités d'origine. En outre, la différence entre les quantifications de *1,19 $* et *1,39 $* (–1,173 et 0) est la même que celle entre les quantifications de *1,39 $* et *1,59 $* (0 et 1,173). Le fait que les modalités 1 et 3 soient à égale distance de la modalité 2 est conservé dans les quantifications.

Figure 9-21
Diagramme de transformation de prix (numérique)



La transformation nominale de la variable *Conception du conditionnement* génère le diagramme de transformation ci-après. Notez la forme non linéaire distincte sous laquelle la deuxième modalité détient la quantification la plus élevée. En matière de régression, la deuxième modalité diminue le rang de préférence prévue, tandis que les première et troisième modalités ont l'effet inverse.

Figure 9-22
Diagramme de transformation de conception du conditionnement (nominal)

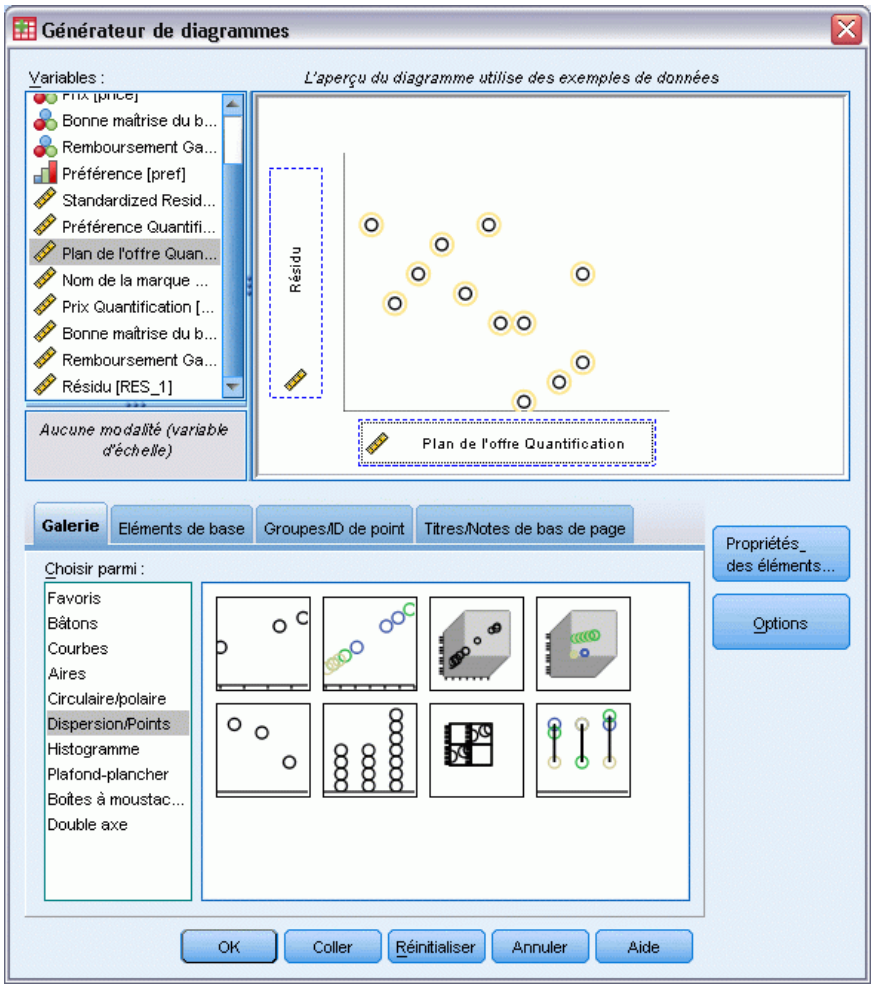


Analyse des résidus

A l'aide des données transformées et des résidus que vous avez enregistrés dans l'ensemble de données actif, vous pouvez créer un diagramme de dispersion des prévisions à partir des valeurs transformées de la variable *Conception du conditionnement*.

Pour obtenir ce type de diagramme de dispersion, rappelez le Générateur de diagrammes, puis cliquez sur le bouton Réinitialiser pour effacer vos sélections précédentes et restaurer les options par défaut.

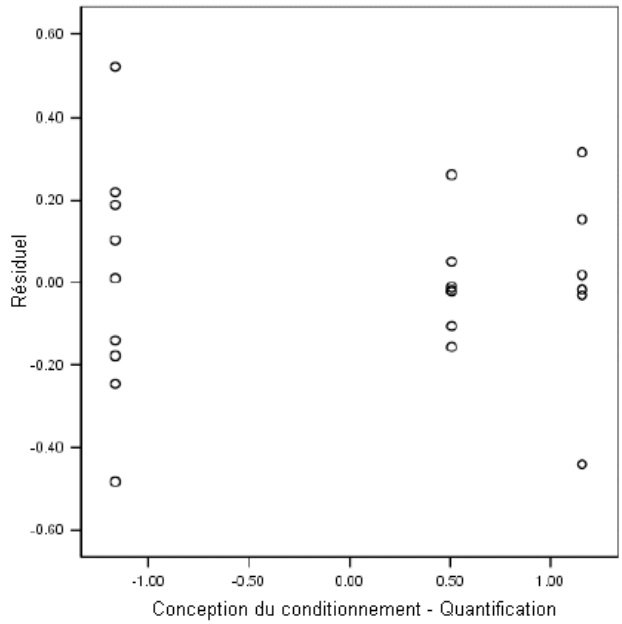
Figure 9-23
Générateur de diagrammes



- Sélectionnez la galerie Dispersion/Points, puis choisissez Dispersion simple.
- Sélectionnez l'option *Résidus* comme variable de l'axe *y*.
- Sélectionnez l'option *Conception du conditionnement - Quantification* comme variable de l'axe *x*.
- Cliquez sur OK.

Le diagramme de dispersion représente les résidus standardisés par rapport aux quantifications optimales de la variable *Conception du conditionnement*. Tous les résidus figurent dans deux écarts-types de valeur 0. Une dispersion aléatoire de points remplace la forme en U présente dans le diagramme de dispersion issu de la régression linéaire standard. La quantification optimale des modalités améliore les capacités prévisionnelles.

Figure 9-24
Résidus de la régression nominale



Exemple : Données d'ozone

Dans cet exemple, vous allez utiliser un plus grand ensemble de données pour illustrer la sélection et les effets des transformations de codage optimal. Les données comprennent 330 observations sur six variables météorologiques précédemment analysées par Breiman et Friedman(Breiman et Friedman, 1985), ainsi que par Hastie et Tibshirani(Hastie et Tibshirani, 1990), entre autres. Le tableau ci-après décrit les variables initiales. Votre régression nominale essaie de prévoir la concentration d'ozone à partir des autres variables. Les chercheurs précédents ont décelé parmi ces variables des non-linéarités qui pénalisent les approches standard de la régression.

Table 9-2
Variables initiales

Variable	Description
ozon	niveau quotidien d'ozone ; classé dans l'une des 38 modalités
h base inv	hauteur de base d'inversion
gr press	gradient de pression (mm Hg)
vis	visibilité (miles)
temp	température (degrés F)
jour année	jour de l'année

Cet ensemble de données est disponible dans le fichier *ozone.sav*.Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*.

Discrétisation des variables

Si une variable détient une quantité excessive de modalités rendant difficile leur interprétation, vous devez modifier les modalités à l'aide de la boîte de dialogue Discrétisation de manière à réduire l'intervalle des modalités et obtenir ainsi une quantité plus facile à gérer.

La variable *Jour de l'année* possède la valeur minimale 3 et la valeur maximale 365. Le recours à cette variable dans une régression nominale correspond à l'utilisation d'une variable avec 365 modalités. De même, l'intervalle de valeurs de la variable *Visibilité (miles)* est compris entre 0 et 350. Pour simplifier l'interprétation des analyses, discrétisez ces variables en intervalles égaux de longueur 10.

L'intervalle de valeurs de la variable *Hauteur de base d'inversion* est compris entre 111 et 5 000. Une variable dotée d'autant de modalités aboutit à des relations très complexes. Toutefois, la discrétisation de cette variable en intervalles égaux de longueur 100 génère approximativement 50 modalités. L'utilisation d'une variable de 50 modalités plutôt que d'une variable de 5 000 modalités simplifie sensiblement les interprétations.

L'intervalle de valeurs de la variable *Gradient de pression (mm Hg)* est compris entre –69 et 107. La procédure retire de l'analyse toutes les modalités codées avec des nombres négatifs, mais la discrétisation de cette variable en intervalles égaux de longueur 10 génère approximativement 19 modalités.

L'intervalle de valeurs de la variable *Température (degrés F)* est compris entre 25 et 93 sur l'échelle Fahrenheit. Pour analyser les données comme si elles figuraient sur l'échelle Celsius, discrétisez cette variable en intervalles égaux de longueur 1,8.

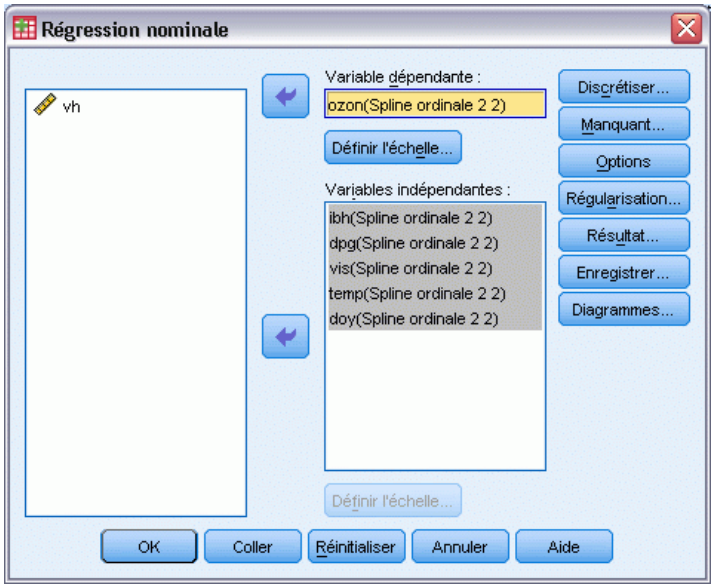
Une discrétisation différente des variables peut être souhaitable. Les choix que vous faites ici sont totalement subjectifs. Pour obtenir moins de modalités, choisissez des intervalles plus grands. Par exemple, la variable *Jour de l'année* aurait pu être divisée en mois de l'année ou en saisons.

Sélection du type de transformation

Différents niveaux d'analyse sont disponibles pour chaque variable. Toutefois, l'objectif étant la prévision de la réponse, vous devez coder celle-ci "en l'état" en utilisant le niveau de codage numérique optimal. Par conséquent, l'ordre des modalités et leurs différences seront conservés dans la variable transformée.

- Pour exécuter une analyse de régression nominale, choisissez les options suivantes dans les menus : Analyse > Régression > Codage optimal (CATREG)…

Figure 9-25
Boîte de dialogue Régression nominale



- Sélectionnez l'option *Niveau quotidien d'ozone* comme variable dépendante.
- Sélectionnez les options allant de *Hauteur de base d'inversion* à *Jour de l'année* comme variables indépendantes.
- Sélectionnez l'option *Niveau quotidien d'ozone*, puis cliquez sur Définir l'échelle.

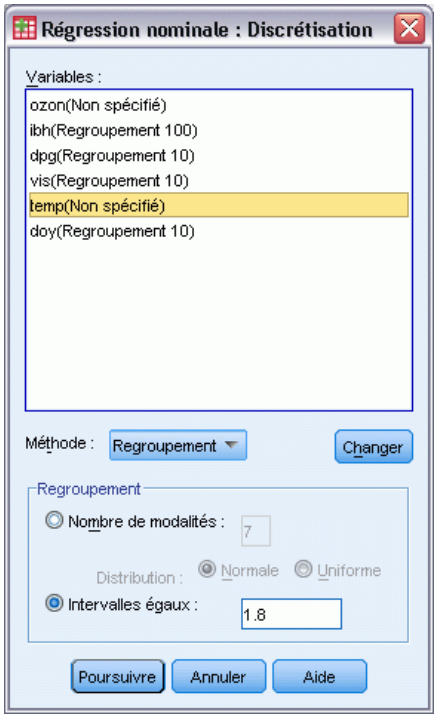
Figure 9-26
Boîte de dialogue Définir l'échelle

- Sélectionnez l'option Numérique comme niveau de codage optimal.
- Cliquez sur Poursuivre.
- Sélectionnez les options allant de *Hauteur de base d'inversion* à *Jour de l'année*, puis cliquez sur Définir l'échelle dans la boîte de dialogue Régression nominale.

Figure 9-27
Boîte de dialogue Définir l'échelle

- Sélectionnez l'option Nominal comme niveau de codage optimal.
- Cliquez sur Poursuivre.
- Cliquez sur Discrétiser dans la boîte de dialogue Régression nominale.

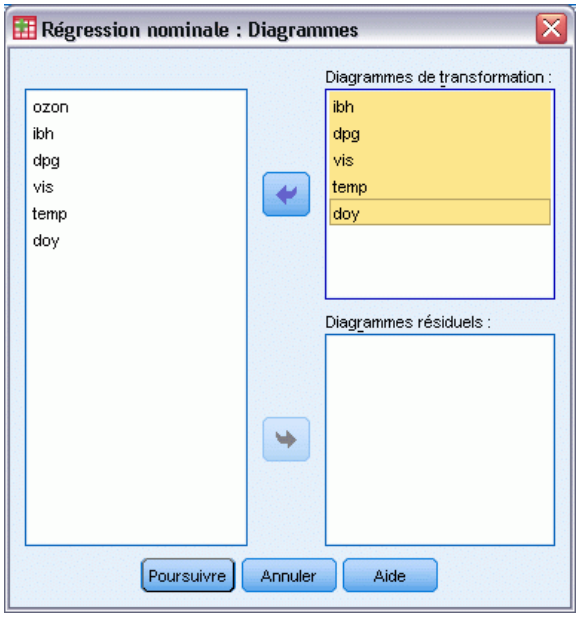
Figure 9-28
Discrétisation



- Sélectionnez l'option *h base inv*.
- Sélectionnez l'option Intervalles égaux, puis tapez 100 comme longueur de l'intervalle.
- Cliquez sur Changer.

- Sélectionnez les options *gr press*, *vis* et *jour année*.
- Tapez 10 comme longueur de l'intervalle.
- Cliquez sur Changer.
- Sélectionnez l'option *temp*.
- Tapez 1.8 comme longueur de l'intervalle.
- Cliquez sur Changer.
- Cliquez sur Poursuivre.
- Cliquez sur Diagrammes dans la boîte de dialogue Régression nominale.

Figure 9-29
Boîte de dialogue Diagrammes



- Sélectionnez les diagrammes de transformation pour la variable *Hauteur de base d'inversion* dans la variable *Jour de l'année*.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Régression nominale.

Figure 9-30
Récapitulatif du modèle

	R-deux multiples	R-deux	R-deux ajusté	Erreur de prévision apparente
Donnéess normalisées	.938	.880	.785	.120

Variable dépendante : Daily ozone level
Variables indépendantes : Inversion base height Pressure gradient (mm Hg) Visibility (miles) Temperature (degrees F) Day of the year

Le traitement de toutes les variables prédites en tant que données nominales génère une mesure R^2 égale à 0,880. Cette quantité élevée de variance représentée n'est pas surprenante dans la mesure où le traitement nominal n'impose aucune restriction sur les quantifications. Toutefois, l'interprétation des résultats peut s'avérer assez difficile.

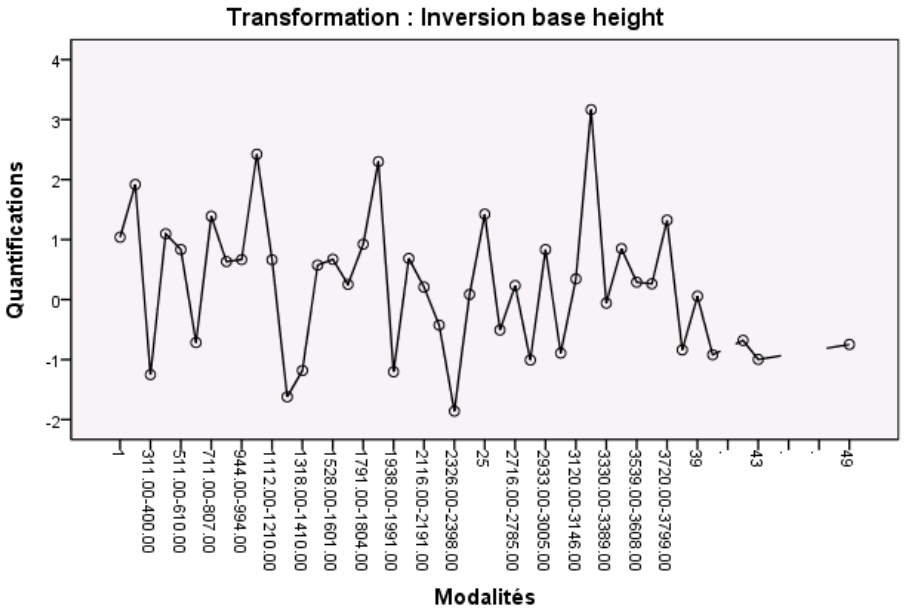
Figure 9-31
Coefficients de régression (toutes les variables prédites traitées en tant que données nominales)

	Coefficients standardisés				
	Bêta	Bootstrap (1000) Estimation de l'erreur standard	ddl	D	Sig.
Inversion base height	.297	.052	42	33.101	.000
Pressure gradient (mm Hg)	.326	.056	16	33.522	.000
Visibility (miles)	.229	.050	17	20.972	.000
Temperature (degrees F)	.577	.088	35	42.714	.000
Day of the year	.420	.072	36	34.425	.000

Variable dépendante : Daily ozone level

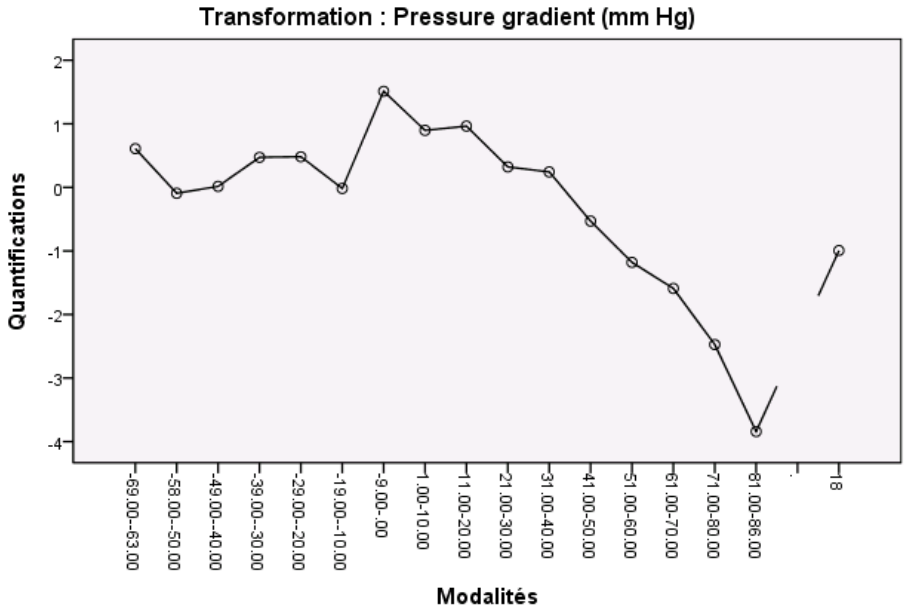
Ce tableau indique les coefficients de régression standardisés des variables prédites. Une erreur fréquente lors de l'interprétation de ces valeurs consiste à se concentrer sur les coefficients au détriment des quantifications. Vous ne pouvez pas simplement affirmer qu'une valeur positive de *Hauteur de base d'inversion*, implique que lorsque la variable prédite augmente, l'*Ozone* prévu augmente. Toutes les interprétations doivent être en rapport avec les variables transformées. Par conséquent, lorsque les quantifications de *Hauteur de base d'inversion* augmentent, l'*Ozone* prévu augmente. Pour examiner les effets des variables initiales, vous devez définir les relations entre les modalités et les quantifications.

Figure 9-32
Diagramme de transformation de la variable Hauteur de base d'inversion (nominal)



Le diagramme de transformation de la variable *Hauteur de base d'inversion* ne montre aucun motif apparent. Comme l'atteste la nature irrégulière du diagramme, le passage des modalités inférieures aux modalités supérieures génère des fluctuations des quantifications dans les deux sens. Par conséquent, la description des effets de cette variable requiert une analyse des différentes modalités. Le fait d'imposer des restrictions ordinales ou linéaires aux quantifications de cette variable peut sensiblement réduire l'ajustement.

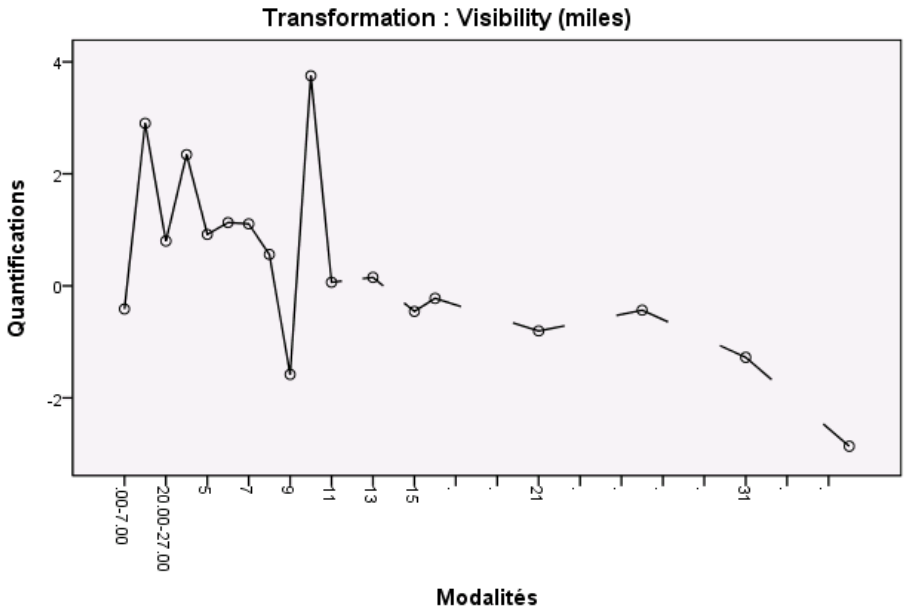
Figure 9-33
Diagramme de transformation de la variable Gradient de pression (nominal)



Ce schéma illustre le diagramme de transformation de la variable *Gradient de pression*. Les modalités discrétisées initiales (*1* à *6*) reçoivent des quantifications réduites, si bien qu'elles contribuent de façon minimale à la réponse prévue. Les trois modalités suivantes reçoivent des valeurs positives un peu plus élevées, générant une augmentation modérée de l'ozone prévu.

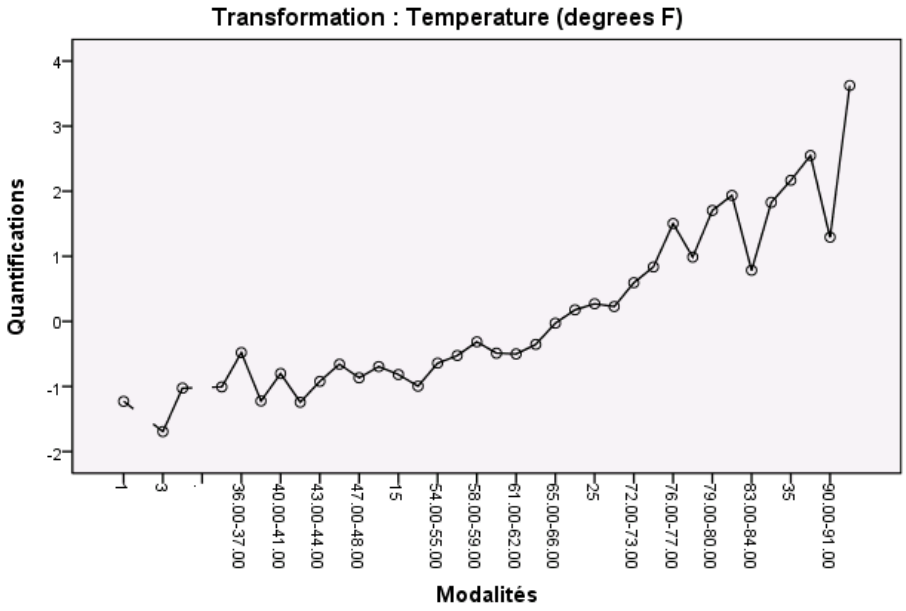
Les quantifications diminuent jusqu'à la modalité *16*, où la variable *Gradient de pression* produit son effet de diminution le plus important sur l'ozone prévu. Bien que la courbe remonte après cette modalité, l'utilisation d'un niveau de codage ordinal pour la variable *Gradient de pression* risque de ne pas beaucoup réduire l'ajustement, tout en simplifiant les interprétations des effets. Toutefois, la mesure d'importance 0,04 et le coefficient de régression de la variable *Gradient de pression* indiquent que cette variable n'est pas très utile dans la régression.

Figure 9-34
Diagramme de transformation de la variable Visibilité (nominal)



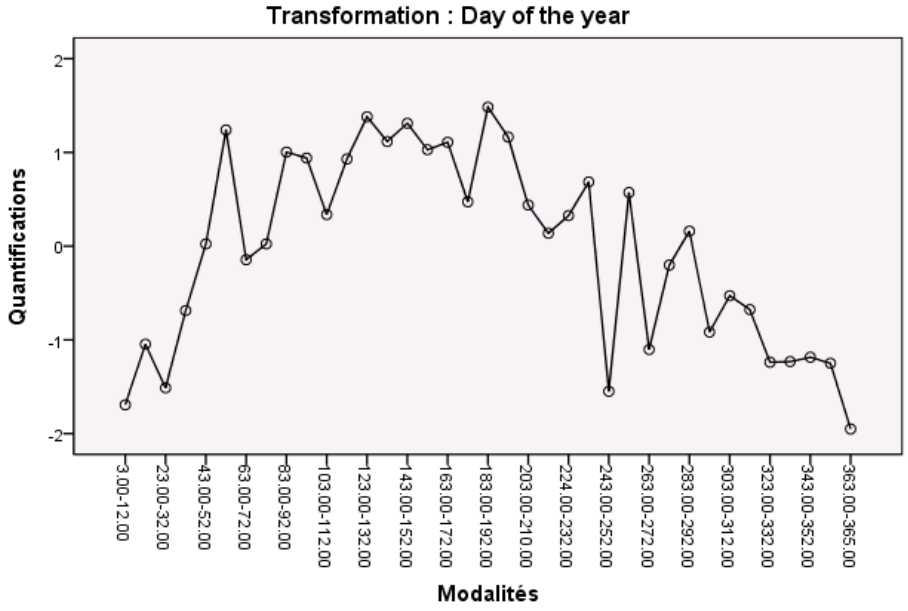
Le diagramme de transformation de la variable *Visibilité*, à l'instar de celui de la variable *Hauteur de base d'inversion*, ne montre aucun motif apparent. Le fait d'imposer des restrictions ordinales ou linéaires aux quantifications de cette variable peut sensiblement réduire l'ajustement.

Figure 9-35
Diagramme de transformation de la variable Température (nominal)



Le diagramme de transformation de la variable *Température* montre un autre motif. A mesure que les modalités augmentent, les quantifications tendent à s'accroître. Par conséquent, à mesure que la variable *Température* augmente, l'ozone prévu tend à s'accroître. Ce motif suggère le codage de la variable *Température* au niveau ordinal.

Figure 9-36
Diagramme de transformation de la variable Jour de l'année (nominal)



Ce schéma illustre le diagramme de transformation de la variable *Jour de l'année*. Les quantifications tendent à augmenter jusqu'au centre du graphique, point à partir duquel elles tendent à diminuer, générant une forme en U inversé. D'après le signe du coefficient de régression de la variable *Jour de l'année*, les modalités initiales reçoivent des quantifications ayant un effet réducteur sur l'ozone prévu. Pour les modalités intermédiaires, l'effet des quantifications sur l'ozone prédit augmente, atteignant son maximum autour du centre du graphique.

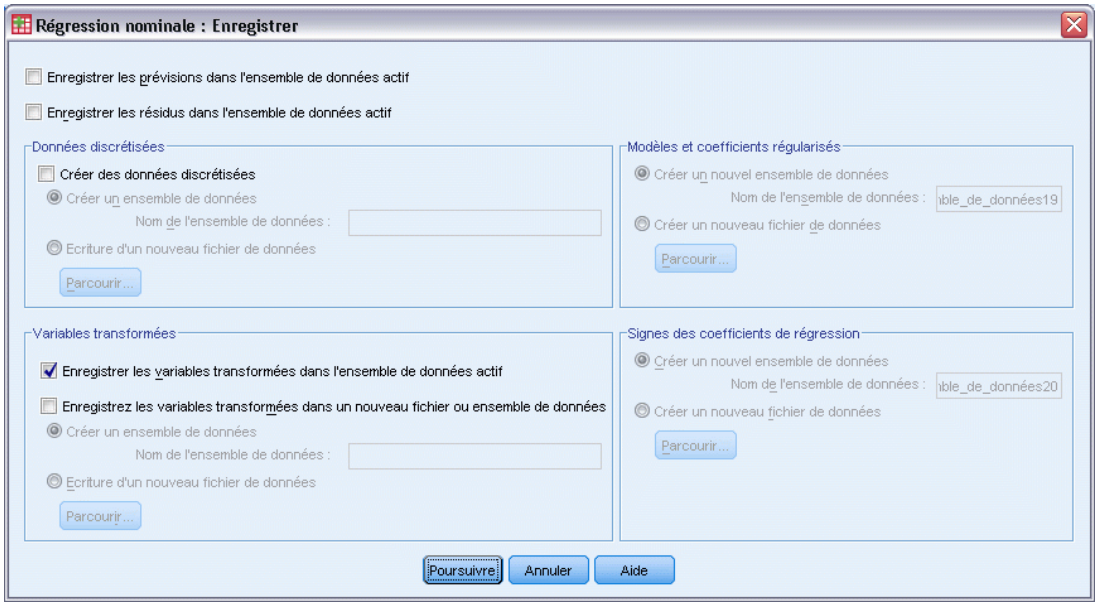
Au-delà de ce point, les quantifications tendent à diminuer l'ozone prévu. Bien que la courbe soit assez irrégulière, la forme générale reste identifiable. Par conséquent, les diagrammes de transformation suggèrent le codage de la variable *Température* au niveau ordinal avec conservation du codage nominal pour toutes les autres variables prédites.

Pour calculer de nouveau la régression en codant la variable *Température* au niveau ordinal, rappelez la boîte de dialogue Régression nominale.

Figure 9-37
Boîte de dialogue Définir l'échelle

- Sélectionnez l'option *Température*, puis cliquez sur Définir l'échelle.
- Sélectionnez l'option Ordinal comme niveau de codage optimal.
- Cliquez sur Poursuivre.
- Cliquez sur le bouton Enregistrer dans la boîte de dialogue Régression nominale.

Figure 9-38
Boîte de dialogue Enregistrer



- Sélectionnez Enregistrer les variables transformées dans l'ensemble de données actif dans le groupe Variables transformées.
- Cliquez sur Poursuivre.

▶ Cliquez sur OK dans la boîte de dialogue Régression nominale.

Figure 9-39
Récapitulatif du modèle de la régression, avec traitement de la variable Température en tant que donnée ordinale

Récapitulatif des modèles

	R-deux multiples	R-deux	R-deux ajusté	Erreur de prévision apparente
Donnéess normalisées	.934	.872	.787	.128

Variable dépendante : Daily ozone level
Variables indépendantes : Inversion base height Pressure gradient (mm Hg) Visibility (miles) Temperature (degrees F) Day of the year

Ce modèle génère une mesure R^2 égale à 0.872, si bien que la variance représentée diminue de façon négligeable lorsque les quantifications de la variable *Température* sont limitées à être ordonnées.

Figure 9-40
Coefficients de régression avec traitement de la variable Température en tant que donnée ordinale

Coefficients

	Coefficients standardisés				
	Bêta	Bootstrap (1000) Estimation de l'erreur standard	ddl	D	Sig.
Inversion base height	.298	.042	42	50.969	.000
Pressure gradient (mm Hg)	.301	.047	16	40.283	.000
Visibility (miles)	.224	.043	17	27.716	.000
Temperature (degrees F)	.609	.084	21	52.317	.000
Day of the year	.373	.051	36	53.134	.000

Variable dépendante : Daily ozone level

Ce tableau répertorie les coefficients du modèle dans lequel la variable *Température* est soumise à un codage ordinal. La comparaison des coefficients à ceux du modèle dans lequel la variable *Température* est soumise à un codage nominal ne laisse pas apparaître de différences significatives.

Figure 9-41
Corrélations, importance et tolérance

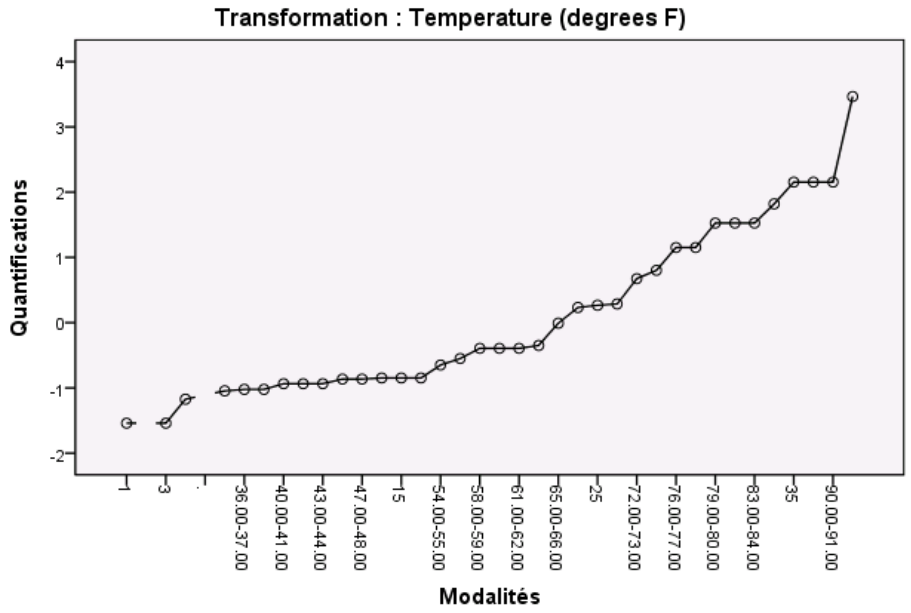
Corrélations et tolérance

	Corrélations			Importance	Tolérance	
	Ordre zéro	Partielle	Partie		Après transformation	Avant transformation
Inversion base height	.438	.627	.288	.150	.930	.596
Pressure gradient (mm Hg)	.128	.606	.272	.044	.815	.858
Visibility (miles)	.365	.518	.216	.094	.933	.752
Temperature (degrees F)	.804	.843	.559	.562	.842	.580
Day of the year	.352	.677	.329	.151	.777	.802

Variable dépendante : Daily ozone level

En outre, les mesures d'importance suggèrent que la variable *Température* reste beaucoup plus importante pour la régression que les autres variables. Toutefois, en raison du niveau de codage ordinal de la variable *Température* et du coefficient de régression positif, vous pouvez désormais affirmer que l'ozone prévu augmente à mesure que la variable *Température* s'accroît.

Figure 9-42
Diagramme de transformation de la variable Température (ordinal)



Le diagramme de transformation illustre la restriction ordinale appliquée aux quantifications de la variable *Température*. La courbe irrégulière issue de la transformation nominale est remplacée ici par une courbe ascendante douce. En outre, l'absence de longs paliers indique que la fusion des modalités n'est pas nécessaire.

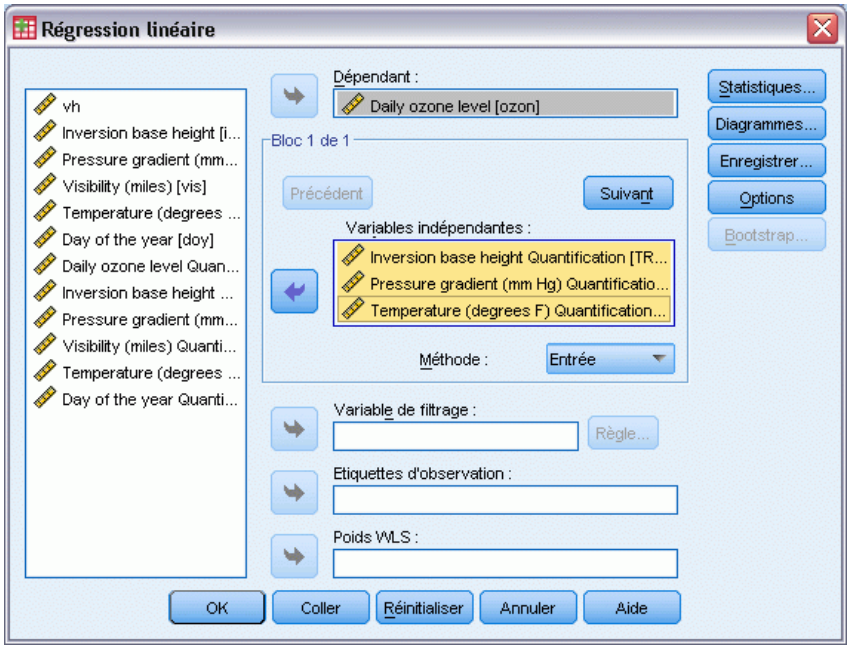
Optimisation des quantifications

Les variables transformées issues d'une régression nominale peuvent être utilisées dans une régression linéaire standard et aboutir à des résultats identiques. Toutefois, les quantifications ne sont optimales que pour le modèle qui les a générées. L'utilisation d'un sous-ensemble de variables prédites dans une régression linéaire ne correspond pas à une régression avec codage optimal sur le même sous-ensemble.

Par exemple, la régression nominale que vous avez calculée présente une mesure R^2 égale à 0,875. Vous avez enregistré les variables transformées. Par conséquent, pour ajuster une régression linéaire uniquement à l'aide des options *Température*, *Gradient de pression* et *Hauteur de base d'inversion* comme variables prédites, dans les menus, choisissez :

Analyse > Régression > Linéaire

Figure 9-43
Boîte de dialogue Régression linéaire



- Sélectionnez l'option *Niveau quotidien d'ozone - Quantification* comme variable dépendante.
- Sélectionnez les options *Hauteur de base d'inversion - Quantification*, *Gradient de pression (mm Hg) - Quantification* et *Température (degrés F) – Quantification* comme variables prédites.
- Cliquez sur OK.

Figure 9-44
Récapitulatif du modèle de régression avec un sous-ensemble de variables prédites codées de façon optimale

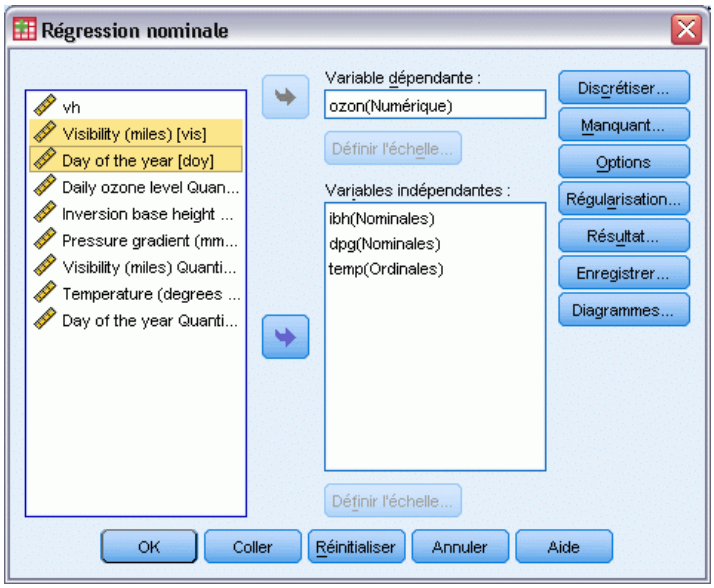
Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.856[a]	.732	.729	4.16711

a. Valeurs prédites : (constantes), Inversion base height Quantification, Pressure gradient (mm Hg) Quantification, Temperature (degrees F) Quantification

Grâce à l'utilisation des quantifications pour la réponse, les variables *Température*, *Gradient de pression* et *Hauteur de base d'inversion* dans une régression linéaire standard génèrent un ajustement égal à 0.732. Pour comparer ce dernier à l'ajustement d'une régression nominale en utilisant uniquement ces trois variables prédites, rappelez la boîte de dialogue Régression nominale.

Figure 9-45
Boîte de dialogue Régression nominale



- Désélectionnez les options *Visibilité (miles)* et *Jour de l'année* comme variables prédites.
- Cliquez sur OK.

Figure 9-46
Récapitulatif du modèle de régression nominale sur trois variables prédites

	R-deux multiples	R-deux	R-deux ajusté	Erreur de prévision apparente
Donnéess normalisées	.892	.796	.735	.204

Variable dépendante : Daily ozone level
Variables indépendantes : Inversion base height Pressure gradient (mm Hg) Temperature (degrees F)

L'analyse de régression nominale possède un ajustement égal à 0.796, meilleur que l'ajustement de 0.732. Cela démontre la propriété des codages selon laquelle les quantifications obtenues dans la régression initiale ne sont optimales que lorsque les cinq variables sont incluses dans le modèle.

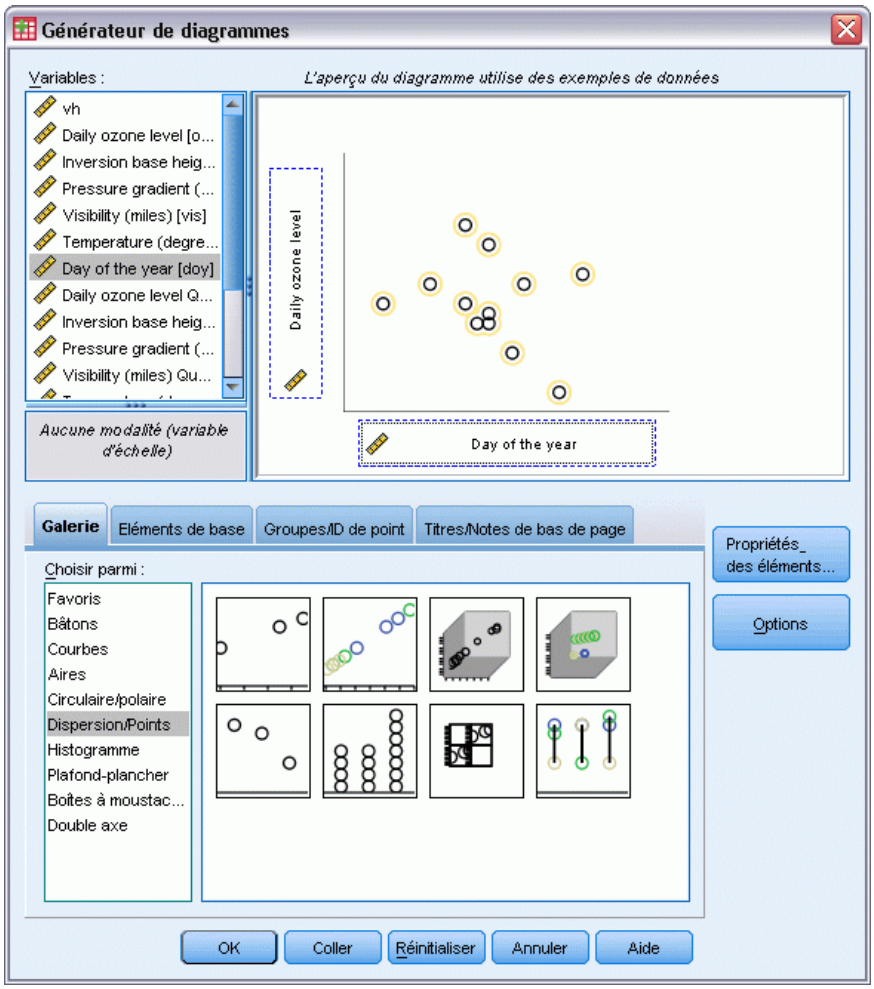
Effets des transformations

La transformation d'une série de variables rend linéaire, pour celles-ci, une relation non linéaire entre la réponse initiale et le groupe initial de variables prédites. Toutefois, en présence de plusieurs variables prédites, les relations par paire sont confondues par les autres variables du modèle.

Pour focaliser votre analyse sur la relation entre les variables *Niveau quotidien d'ozone* et *Jour de l'année*, commencez par observer un diagramme de dispersion. A partir des menus, sélectionnez :

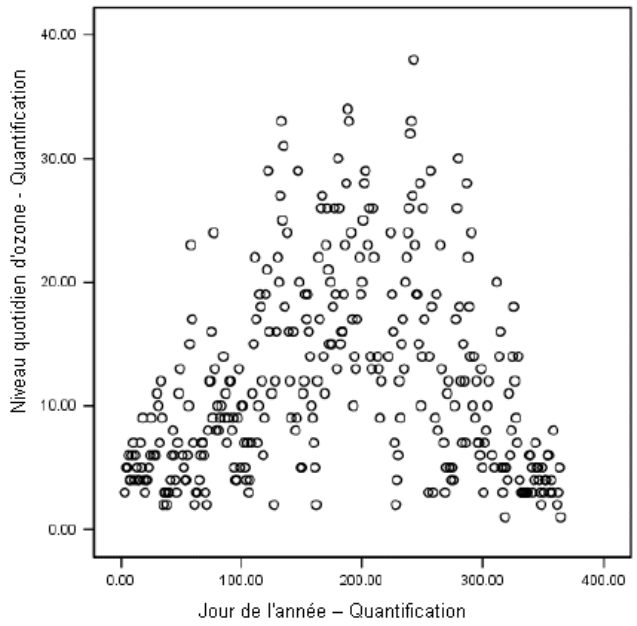
Graphes > Générateur de diagrammes...

Figure 9-47
Boîte de dialogue Générateur de diagrammes



- Sélectionnez la galerie Dispersion/Points, puis choisissez Dispersion simple.
- Sélectionnez l'option *Niveau quotidien d'ozone* comme variable de l'axe *y* et l'option *Jour de l'année* comme variable de l'axe *x*.
- Cliquez sur OK.

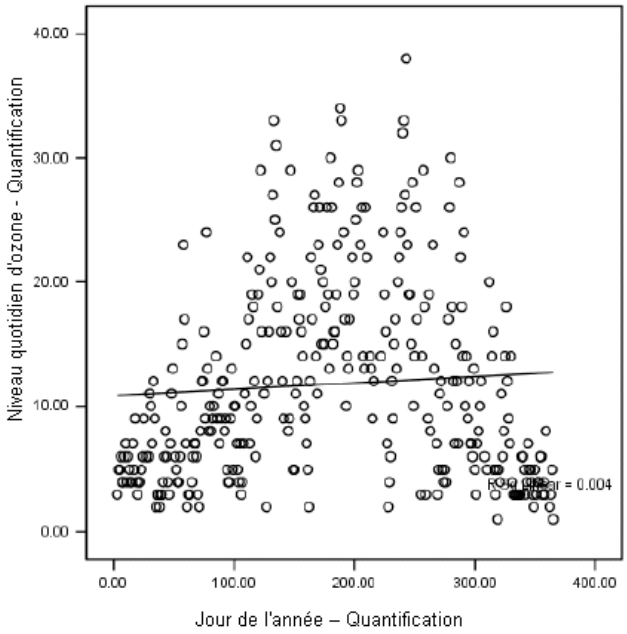
Figure 9-48
Diagramme de dispersion du niveau d'ozone quotidien par jour de l'année



Ce schéma illustre la relation entre les variables *Niveau quotidien d'ozone* et *Jour de l'année*. A mesure que la variable *Jour de l'année* augmente jusqu'à environ 200, la variable *Niveau quotidien d'ozone* s'accroît. Toutefois, pour les valeurs de la variable *Jour de l'année* supérieures à 200, la variable *Niveau quotidien d'ozone* diminue. Ce motif en U inversé suggère une relation quadratique entre les deux variables. Une régression linéaire ne peut pas capturer cette relation.

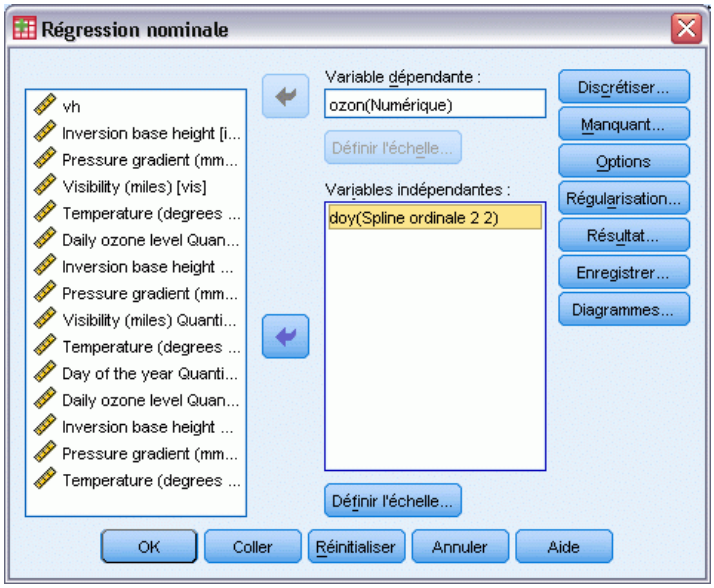
- Pour qu'une courbe optimisée relie les points du diagramme de dispersion, activez le graphique en double-cliquant dessus.
- Sélectionnez un point dans l'éditeur de diagrammes.
- Cliquez sur l'outil Ajouter une courbe d'ajustement au total, puis fermez Chart Editor.

Figure 9-49
Diagramme de dispersion contenant la courbe d'ajustement la plus appropriée



Une régression linéaire de la variable *Niveau quotidien d'ozone* sur la variable *Jour de l'année* génère une mesure R^2 égale à 0,004. Cet ajustement suggère que la variable *Jour de l'année* ne possède aucune valeur prévisionnelle pour la variable *Niveau quotidien d'ozone*. Cela n'est pas surprenant, au vu du motif du schéma. Toutefois, vous pouvez recourir au codage optimal pour linéariser la relation quadratique et utiliser la variable *Jour de l'année* transformée pour prévoir la réponse.

Figure 9-50
Boîte de dialogue Régression nominale



Pour obtenir une régression nominale de la variable *Niveau quotidien d'ozone* sur la variable *Jour de l'année*, rappelez la boîte de dialogue Régression nominale.

- Désélectionnez les options allant de *Hauteur de base d'inversion* à *Température (degrés F)* comme variables indépendantes.
- Sélectionnez l'option *Jour de l'année* comme variable indépendante.
- Cliquez sur Définir l'échelle.

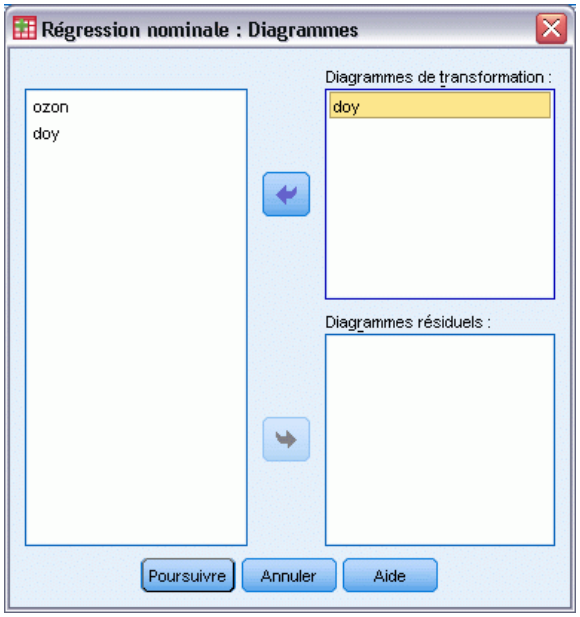
Figure 9-51
Boîte de dialogue Définir l'échelle

- Sélectionnez l'option Nominal comme niveau de codage optimal.
- Cliquez sur Poursuivre.
- Cliquez sur Discrétiser dans la boîte de dialogue Régression nominale.

Figure 9-52
Discrétisation

- Sélectionnez l'option *jour année*.
- Sélectionnez l'option Intervalles égaux.
- Tapez 10 comme longueur de l'intervalle.
- Cliquez sur Changer.
- Cliquez sur Poursuivre.
- Cliquez sur Diagrammes dans la boîte de dialogue Régression nominale.

Figure 9-53
Boîte de dialogue Diagrammes



- Sélectionnez l'option *jour année* pour les diagrammes de transformation.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Régression nominale.

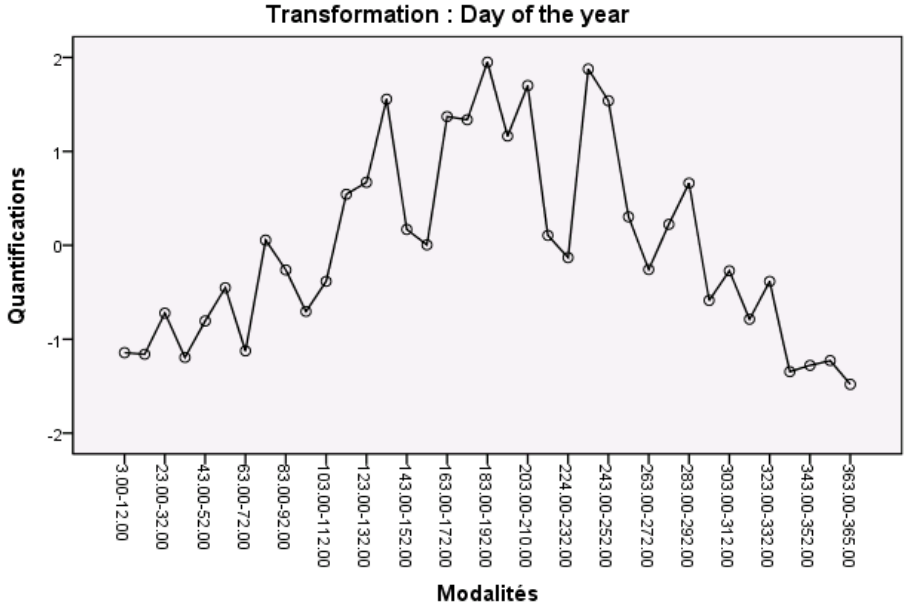
Figure 9-54
Récapitulatif du modèle de régression nominale de la variable Niveau quotidien d'ozone sur la variable Jour de l'année

	R-deux multiples	R-deux	R-deux ajusté	Erreur de prévision apparente
Donnéess normalisées	.741	.549	.494	.451

Variable dépendante : Daily ozone level
Variable indépendante\ : Day of the year

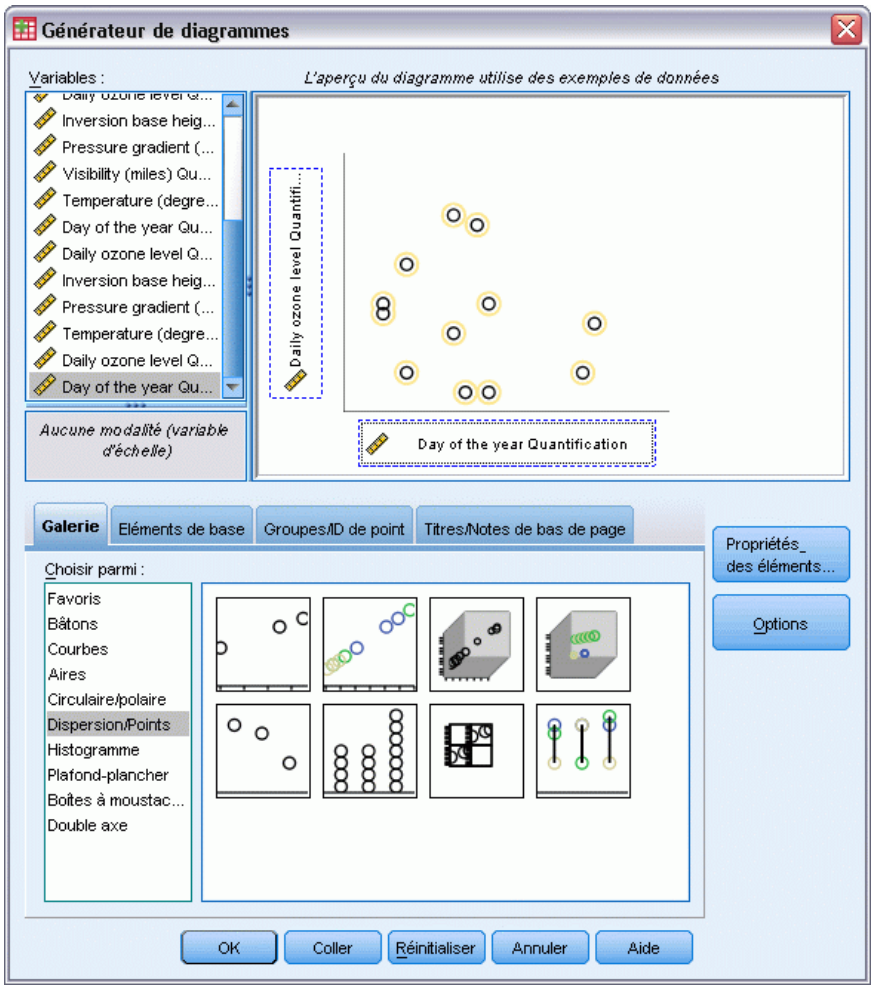
La régression avec codage optimal traite la variable *Niveau quotidien d'ozone* en tant que donnée numérique et la variable *Jour de l'année* en tant que donnée nominale. Cette opération génère une mesure R^2 égale à 0,549. Bien que seulement 55 % de la variation de la variable *Niveau quotidien d'ozone* soient représentés par la régression nominale, cela constitue une amélioration significative par rapport à la régression initiale. La transformation de la variable *Jour de l'année* permet de prévoir la variable *Niveau quotidien d'ozone*.

Figure 9-55
Diagramme de transformation de la variable Jour de l'année (nominal)



Ce schéma affiche le diagramme de transformation de la variable *Jour de l'année*. Les deux extrêmes de la variable *Jour de l'année* reçoivent des quantifications négatives, tandis que les valeurs centrales possèdent des quantifications positives. Une fois cette transformation appliquée, les valeurs basse et haute de la variable *Jour de l'année* ont des effets similaires sur la variable *Niveau quotidien d'ozone* prévue.

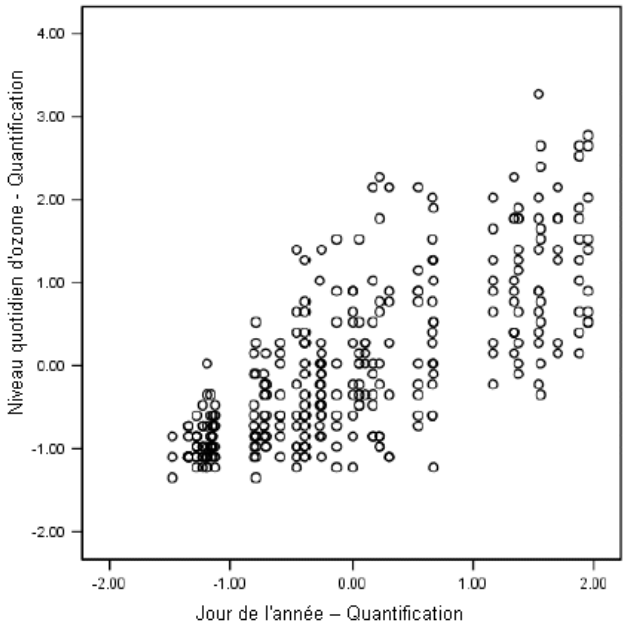
Figure 9-56
Générateur de diagrammes



Pour obtenir un diagramme de dispersion des variables transformées, rappelez le Générateur de diagrammes, puis cliquez sur le bouton Réinitialiser afin d'effacer vos sélections antérieures.

- Sélectionnez la galerie Dispersion/Points, puis choisissez Dispersion simple.
- Sélectionnez l'option *Niveau quotidien d'ozone - Quantification [TRA1_3]* comme variable de l'axe *y* et l'option *Jour de l'année – Quantification [TRA2_3]* comme variable de l'axe *x*.
- Cliquez sur OK.

Figure 9-57
Diagramme de dispersion des variables transformées



Ce schéma décrit la relation entre les variables transformées. Une tendance à l'augmentation remplace la forme en U inversée. La ligne de régression possède une pente positive, ce qui indique que le *Niveau d'ozone quotidien* prévu augmente à mesure que la variable *Jour de l'année* transformée s'accroît. L'utilisation du codage optimal linéarise la relation et autorise des interprétations qui seraient passées inaperçues.

Lectures recommandées

Pour plus d'informations sur la régression nominale, reportez-vous aux documents suivants :

Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics*, 18, .

Hastie, T., R. Tibshirani, et A. Buja. 1994. Flexible discriminant analysis. *Journal of the American Statistical Association*, 89, .

Hayashi, C. 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statitical Mathematics*, 2, .

Kruskal, J. B. 1965. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society Series B*, 27, .

Meulman, J. J. 2003. Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, 4, .

Ramsay, J. O. 1989. Monotone regression splines in action. *Statistical Science*, 4, .

Van der Kooij, A. J., et J. J. Meulman. 1997. MURALS: Multiple regression and optimal scaling using alternating least squares. Dans : *Softstat '97,* F. Faulbaum, et W. Bandilla, éds. Stuttgart: Gustav Fisher.

Winsberg, S., et J. O. Ramsay. 1980. Monotonic transformations to additivity using splines. *Biometrika*, 67, .

Winsberg, S., et J. O. Ramsay. 1983. Monotone spline transformations for dimension reduction. *Psychometrika*, 48, .

Young, F. W., J. De Leeuw, et Y. Takane. 1976. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .

Chapitre
10

Analyse en composantes principales qualitatives

L'analyse en composantes principales qualitatives peut être considérée comme une méthode de réduction des dimensions. Un ensemble de variables est analysé de manière à mettre en évidence les principales dimensions de variation. L'ensemble de données initial peut ensuite être remplacé par un nouvel ensemble plus petit avec une perte d'informations minimale. La méthode met en évidence les relations entre les variables, entre les observations et entre les variables et les observations.

Le critère utilisé par l'analyse en composantes principales qualitatives pour la quantification des données observées est le suivant : les coordonnées principales (scores des composantes) doivent avoir des corrélations élevées avec chacune des variables quantifiées. Une solution est appropriée dans la mesure où ce critère est respecté.

Deux exemples d'analyse en composantes principales qualitatives seront présentés. Le premier emploie un ensemble de données plutôt réduit permettant d'illustrer les concepts de base et les interprétations associées à la procédure. Le second exemple examine une application pratique.

Exemple : Examen des relations entre systèmes sociaux

Cet exemple étudie l'adaptation d'un tableau de Guttman (Guttman, 1968) par Bell (Bell, 1961). Les données sont également présentées par Lingoes (Lingoes, 1968).

Bell a présenté un tableau pour illustrer les groupes sociaux possibles. Guttman a utilisé une partie de ce tableau, dans lequel cinq variables décrivant des éléments tels que l'interaction sociale, le sentiment d'appartenance à un groupe, la proximité physique des membres et la formalité de la relation, ont été croisées avec sept groupes sociaux théoriques, dont les foules (par exemple, le public d'un match de football), l'audience (par exemple, au cinéma ou dans une salle de classe), le public (par exemple, les journaux ou la télévision), les bandes (proche d'une foule, mais qui serait caractérisée par une interaction beaucoup plus intense), les groupes primaires (intimes), les groupes secondaires (volontaires) et la communauté moderne (groupement lâche issu d'une forte proximité physique et d'un besoin de services spécialisés).

Le tableau suivant indique les variables de l'ensemble de données résultant de la classification en sept groupes sociaux utilisés dans les données Guttman-Bell, ainsi que les étiquettes de variable correspondantes et les étiquettes de valeur (modalités) associées aux niveaux de chaque variable. Cet ensemble de données est disponible dans le fichier *guttman.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*. Outre les variables à inclure dans le calcul de l'analyse des composantes principales qualitatives, vous pouvez sélectionner celles utilisées pour étiqueter les objets sur les diagrammes. Dans cet exemple, les cinq premières variables des données sont incluses dans l'analyse, tandis que la classe est exclusivement utilisée comme variable d'étiquetage. Lorsque vous spécifiez une analyse des composantes principales qualitatives, vous devez définir le niveau de codage optimal

de chaque variable d'analyse. Dans cet exemple, un niveau ordinal est spécifié pour toutes les variables d'analyse.

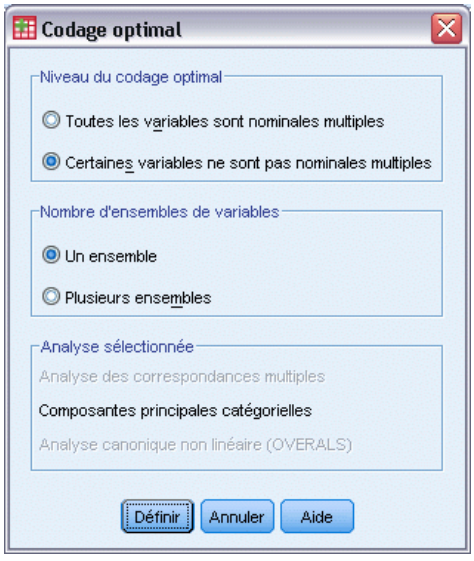
Table 10-1
Variables de l'ensemble de données Guttman-Bell

Nom de variable	l'étiquette Variable	Etiquette de valeur
intensité	Intensité de l'interaction	légère, faible, modérée, élevée
fréquence	Fréquence de l'interaction	Légère, non récurrente, rare, fréquente
appartenance	Sentiment d'appartenance	Aucun, léger, variable, élevé
proximité	Proximité physique	Distante, proche
formalité	Formalité de la relation	aucune relation, formelle, informelle
classe		Foules, audience, public, modèle d'objets, groupes primaires, groupes secondaires, communauté moderne

Exécution de l'analyse

- Pour générer un résultat de composants principaux qualitatifs pour cet ensemble de données, choisissez dans les menus :
  Analyse > Réduction des dimensions > Codage optimal

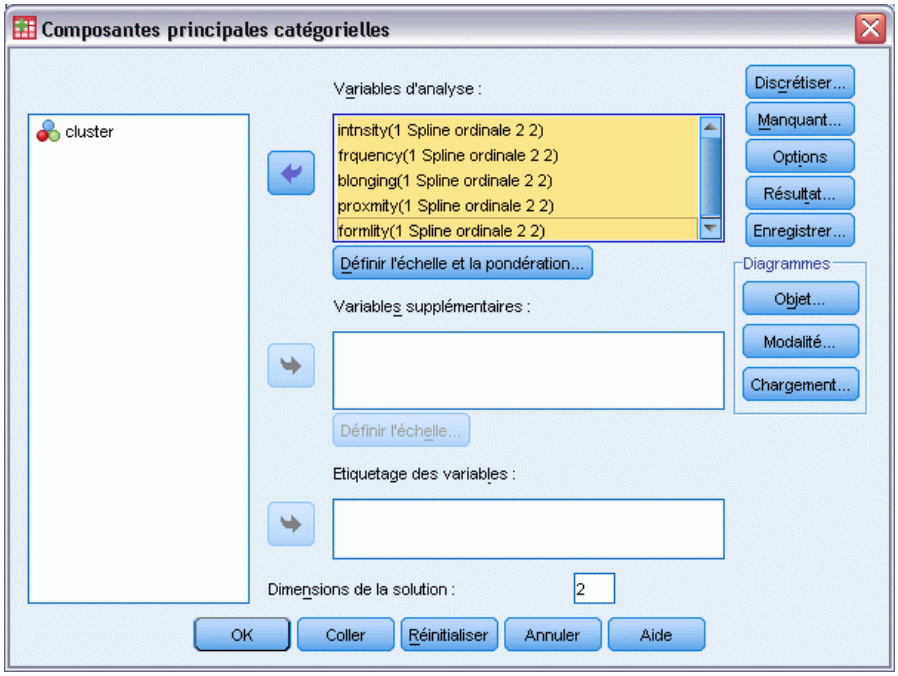
Figure 10-1
Boîte de dialogue Niveau du codage optimal



- Sélectionnez l'option Certaines variables non nominales multiples dans le groupe Niveau du codage optimal.

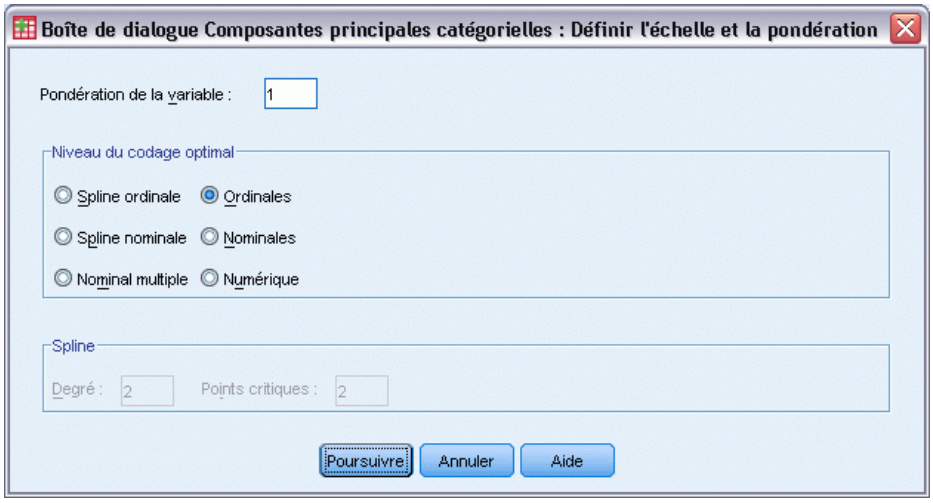
- Cliquez sur Définir.

Figure 10-2
Boîte de dialogue Composantes principales qualitatives



- Sélectionnez les options allant de *Intensité de l'interaction* à *Formalité de la relation* comme variables d'analyse.
- Cliquez sur Définir l'échelle et la pondération.

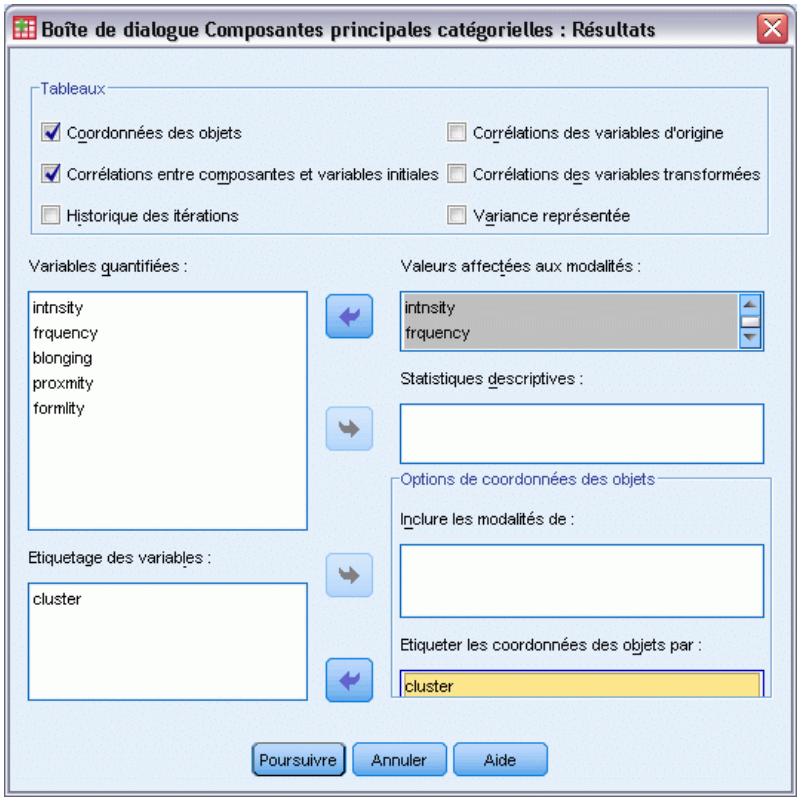
Figure 10-3
Définir l'échelle et la pondération



- Sélectionnez l'option Ordinal dans le groupe Niveau du codage optimal.
- Cliquez sur Poursuivre.

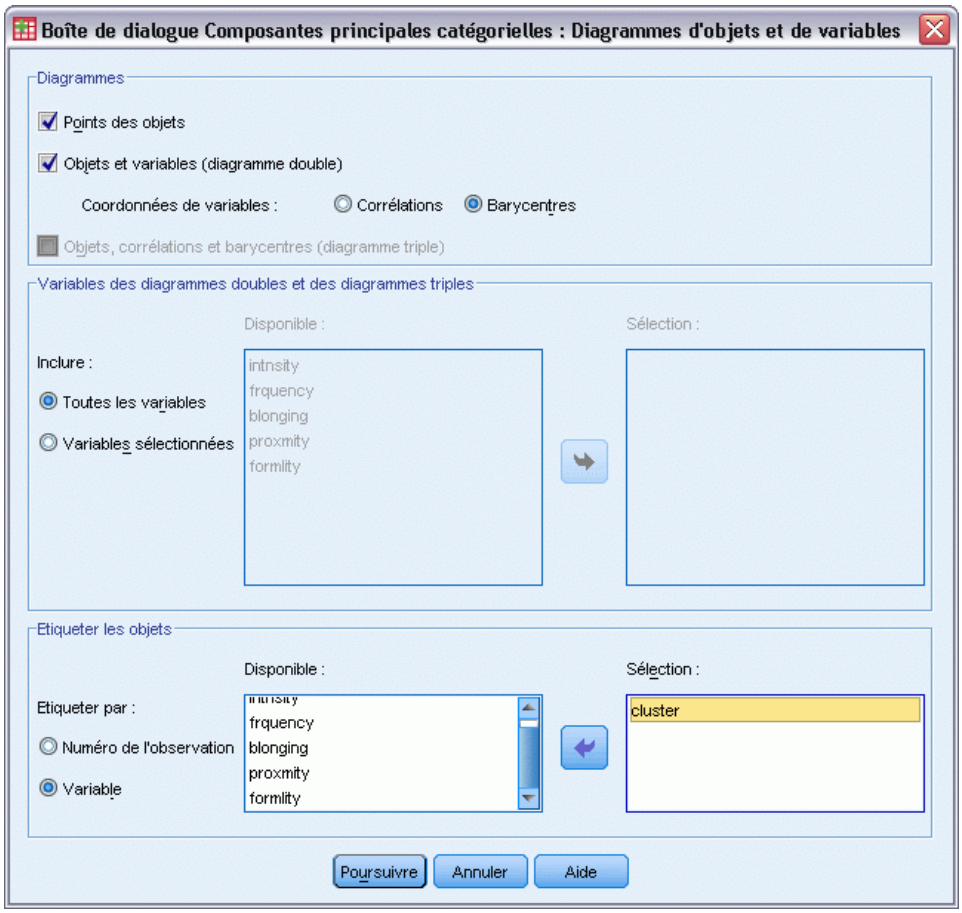
- Sélectionnez *grappe* comme variable d'étiquetage dans la boîte de dialogue Composantes principales qualitatives.
- Cliquez sur Résultat.

Figure 10-4
Résultat



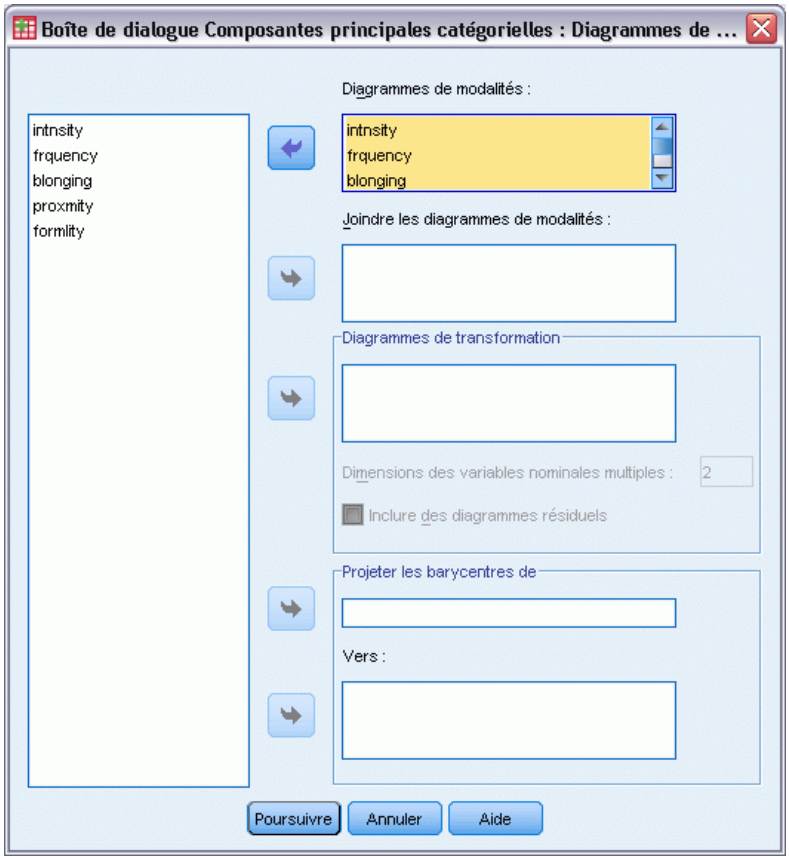
- Sélectionnez l'option Coordonnées principales et désélectionnez l'option Corrélations des variables transformées dans le groupe Tableaux.
- Appliquez la génération de quantifications de modalités aux options allant de *intnsité (Intensité de l'interaction)* à *formlité (Formalité de la relation)*.
- Choisissez d'étiqueter les coordonnées des objets par *grappe*.
- Cliquez sur Poursuivre.
- Cliquez sur Objet dans le groupe Diagrammes de la boîte de dialogue Composantes principales qualitatives.

Figure 10-5
Diagrammes d'objets et de variables



- Sélectionnez l'option Objets et variables (diagramme double) dans le groupe Diagrammes.
- Dans le groupe Objets d'étiquetage, choisissez l'option d'étiquetage des objets par Variable, puis sélectionnez l'option *Grappe* comme variable d'étiquetage des objets.
- Cliquez sur Poursuivre.
- Cliquez sur Modalité dans le groupe Diagrammes de la boîte de dialogue Composantes principales qualitatives.

Figure 10-6
Boîte de dialogue Diagrammes de modalités



- Appliquez l'opération Joindre les diagrammes de modalités aux options allant de *intnsité (Intensité de l'interaction)* à *formlité (Formalité de la relation)*.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Composantes principales qualitatives.

Nombre de dimensions.

Ces données montrent une partie du résultat initial de l'analyse en composantes principales nominales. Après l'historique des itérations de l'algorithme, le récapitulatif du modèle, y compris les valeurs propres de chaque dimension, apparaît. Ces valeurs propres sont équivalentes à celles de l'analyse en composantes principales classique. Elles permettent de mesurer la quantité de variance représentée par chaque dimension.

Figure 10-7
Historique des itérations

Nombre d'itérations	Variance expliquée		Perte		
	Total Variance Expliquée	Augmentation	Total Variance Non Expliquée	Coordonnées des centres de gravité	Restriction des centres de gravité aux coordonnées vectorielles
0[a]	4,515315	,000000	5,484685	4,075583	1,409101
31[b]	4,726009	,000008	5,273991	4,273795	1,000196

a. L'itération 0 présente les statistiques de la solution avec toutes les variables, à l'exclusion des variables présentant un niveau de codage optimal nominal multiple, considérées comme numériques.

b. Le processus d'itération s'est interrompu car la valeur test de la convergence a été atteinte.

Figure 10-8
Récapitulatif du modèle

Dimension	Alpha de Cronbach	Variance expliquée	
		Total (valeur propre)	Pourcentage de variance expliquée
1	,881	3,389	67,774
2	,315	1,337	26,746
Total	,986[a]	4,726	94,520

a. La valeur Alpha de Cronbach totale est basée sur la valeur propre totale.

Les valeurs propres permettent de déterminer le nombre de dimensions requises. Cet exemple utilise le nombre de dimensions par défaut (2). Ce nombre est-il correct ? En règle générale, lorsque toutes les variables sont nominales simples, ordinales ou numériques, la valeur propre d'une dimension doit être supérieure à 1. Dans la mesure où la solution bidimensionnelle représente 94,52 % de la variance, une troisième dimension n'apporterait probablement pas beaucoup plus d'informations.

Dans le cas des variables nominales multiples, il n'existe pas de principe de base simple permettant de déterminer le nombre de dimensions approprié. Si le nombre de variables est remplacé par le nombre total de modalités moins le nombre de variables, la règle ci-dessus demeure valable. Cependant, cette règle seule autoriserait probablement davantage de dimensions que le nombre requis. Lors du choix du nombre de dimensions, la conduite la plus utile consiste à définir un nombre suffisamment faible de manière à ce que des interprétations significatives soient possibles. En outre, le tableau récapitulatif du modèle indique l'alpha de Cronbach (mesure de fiabilité), qui est optimisé par la procédure.

Quantifications

Pour chaque variable, les quantifications, les coordonnées vectorielles et celles des centres de gravité de chaque dimension sont présentées. Les quantifications sont les valeurs affectées à chaque modalité. Les coordonnées des centres de gravité représentent la moyenne des coordonnées principales des objets d'une même modalité. Les coordonnées vectorielles sont les coordonnées des modalités qui figurent sur une ligne, afin de représenter la variable dans l'espace de l'objet. Ce dispositif est requis pour les variables dont le niveau de codage est ordinal ou numérique.

Figure 10-9
Quantifications de l'intensité de l'interaction

INTENSITE[a]

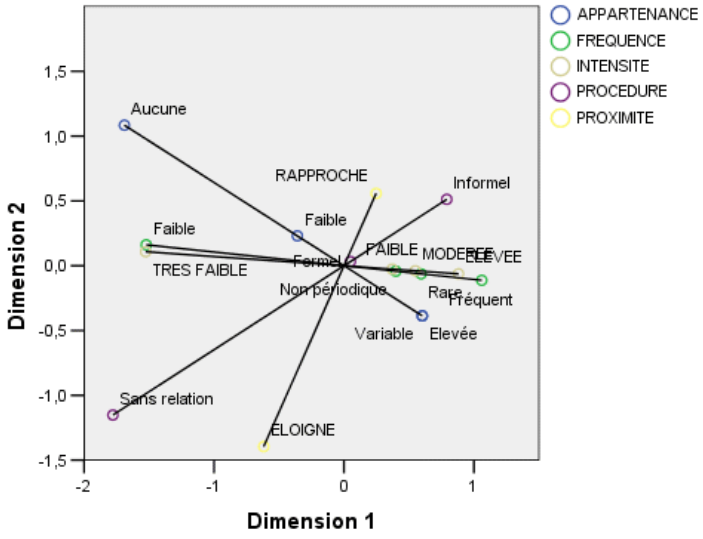
Modalité	Effectif	Quantification	Coordonnées des centres de gravité – Dimension 1	Coordonnées des centres de gravité – Dimension 2	Coordonnées vectorielles – Dimension 1	Coordonnées vectorielles – Dimension 2
TRES FAIBLE	2	-1,530	-1,496	,308	-1,510	,208
FAIBLE	2	,362	,392	,202	,358	-,049
MODEREE	1	,379	,188	-1,408	,374	-,051
ELEVEE	2	,978	1,010	,194	,965	-,133

Normalisation principale de la variable.

a. Niveau de codage optimal : Ordinal.

Les quantifications du diagramme joint des points de modalités indiquent que des modalités de certaines variables n'ont pas été aussi nettement séparées par l'analyse des composantes principales nominales que si l'opération avait eu recours à un niveau réellement ordinal. Les variables *Intensité de l'interaction* et *Fréquence d'interaction*, par exemple, présentent des quantifications égales ou pratiquement égales pour leurs deux modalités intermédiaires. Ce type de résultat peut amener à essayer d'autres analyses en composantes principales qualitatives, éventuellement en fusionnant certaines modalités ou en utilisant un autre niveau d'analyse, par exemple nominal (multiple).

Figure 10-10
Points de modalités du diagramme joint



Le diagramme joint des points de modalité ressemble au diagramme des contributions des facteurs, mais il indique également la position des extrema correspondant aux quantifications les plus faibles (par exemple, *Légère* pour *Intensité de l'interaction* et *aucun* pour *Sentiment d'appartenance*). Les deux variables mesurant l'interaction, *Intensité de l'interaction* et *Effectif*

d'interactions, sont très proches l'une de l'autre et représentent une grande partie de la variance de la dimension 1. La valeur *Formalité de la relation* se trouve également près de *Proximité physique*.

Les points de modalité permettent de discerner les relations plus clairement. Non seulement les variables *Intensité de l'interaction* et *Fréquence d'interaction* sont proches, mais les directions de leurs échelles sont similaires ; en d'autres termes, une intensité légère est proche d'une fréquence légère et une interaction fréquente est proche d'une intensité d'interaction élevée. Vous pouvez également constater que la forte proximité physique semble aller de pair avec un type informel de relation et que la distance physique est liée à l'absence de relation.

Coordonnées principales

En outre, vous pouvez demander une liste et un diagramme des coordonnées principales. Le diagramme des coordonnées principales peut être utile pour détecter des valeurs éloignées, repérer des groupes typiques d'objets ou mettre en évidence des modèles particuliers.

Le tableau des coordonnées principales répertorie les coordonnées principales étiquetées par groupe social pour les données Guttman-Bell. En examinant les valeurs des points des objets, vous pouvez identifier des objets spécifiques dans le diagramme.

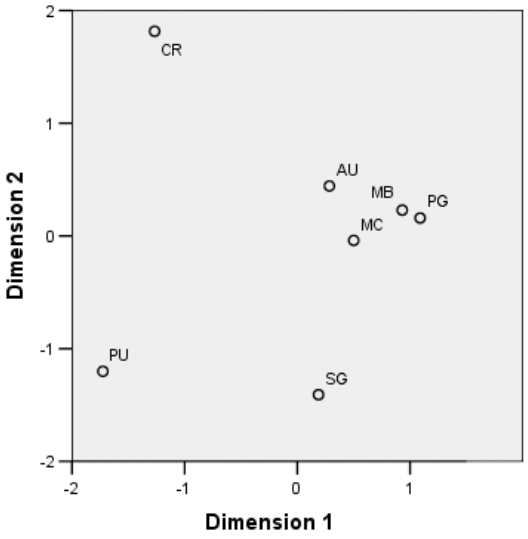
Figure 10-11
Coordonnées des objets

	Dimension	
CLASSE	1	2
CR	-1,266	1,816
AU	,284	,444
PU	-1,726	-1,201
MB	,931	,229
PG	1,089	,159
SG	,188	-1,408
MC	,500	-,039

Normalisation principale de la variable.

La première dimension sépare *FOULES* et *PUBLIC*, qui ont des scores négatifs relativement élevés, de *BANDES* et *GROUPES PRIMAIRES*, qui ont des scores positifs relativement élevés. La deuxième dimension possède trois groupes : *PUBLIC* et *GROUPES SECONDAIRES* avec des valeurs négatives élevées, *FOULES* avec des valeurs positives élevées, puis les autres groupes sociaux intermédiaires. L'inspection du diagramme des coordonnées principales met en évidence cette organisation.

Figure 10-12
Diagramme des coordonnées principales

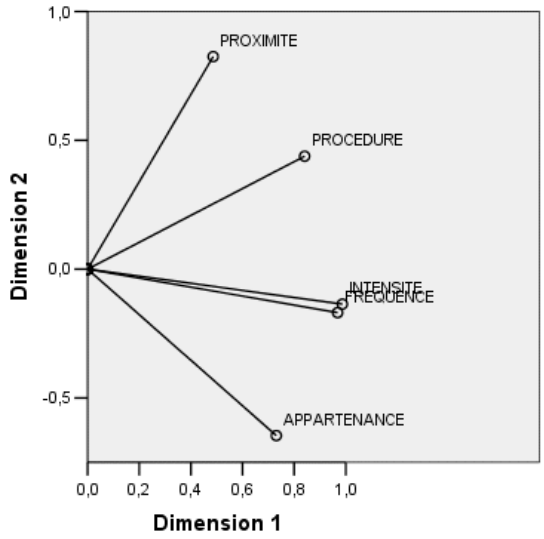


Dans le diagramme, *PUBLIC* et *GROUPES SECONDAIRES* apparaissent en bas, *FOULES* en haut et les autres groupes sociaux au milieu. L'examen des modèles parmi les différents objets dépend des informations supplémentaires disponibles pour les unités de l'analyse. Dans ce cas, vous connaissez la classification des objets. Dans d'autres cas, vous pouvez utiliser des variables supplémentaires pour étiqueter les objets. Vous pouvez également constater que l'analyse en composantes principales nominales ne sépare pas *BANDES* de *GROUPES PRIMAIRES*. Bien que la plupart des personnes ne considèrent généralement pas leurs familles comme des bandes, ces deux groupes obtiennent le même score sur quatre des cinq variables utilisées. Il va de soi que vous pouvez explorer les points faibles éventuels des variables et des modalités utilisées. Par exemple, une intensité d'interaction élevée et des relations informelles n'ont probablement pas la même signification pour ces deux groupes. Par ailleurs, vous pouvez envisager une solution impliquant davantage de dimensions.

Saturations

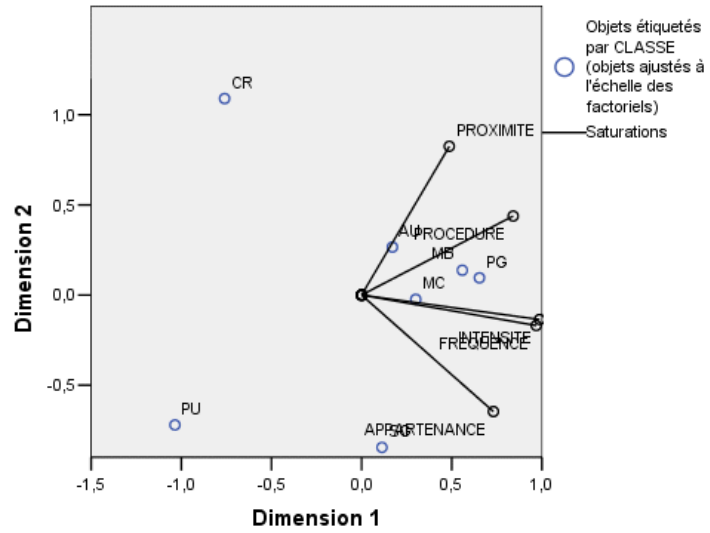
Ce schéma illustre le diagramme des corrélations entre composantes. Les vecteurs (lignes) sont relativement longs, ce qui est une nouvelle indication du fait que les deux premières dimensions représentent la majeure partie de la variance de toutes les variables quantifiées. Sur la première dimension, toutes les variables possèdent des corrélations entre composantes élevées (positives). La deuxième dimension est principalement corrélée avec les variables quantifiées *Sentiment d'appartenance* et *Proximité physique*, dans des sens opposés. Cela signifie que les objets ayant un score négatif élevé dans la dimension 2 auront un score élevé pour le sentiment d'appartenance et un score faible pour la proximité physique. Par conséquent, la deuxième dimension met en évidence un contraste entre ces deux variables tout en ayant peu de rapport avec les variables quantifiées *Intensité de l'interaction* et *Fréquence d'interaction*.

Figure 10-13
Corrélations entre composantes et variables initiales



Pour examiner la relation entre les objets et les variables, observez le diagramme double des objets et des saturations. Le vecteur d'une variable pointe en direction de la modalité la plus élevée de la variable. Par exemple, pour les variables *Proximité physique* et *Sentiment d'appartenance*, les modalités les plus élevées sont *forte* et *fort*, respectivement. Par conséquent, une forte proximité physique et l'absence de sentiment d'appartenance caractérisent les foules (*FOULES*), tandis qu'une proximité physique distante et un fort sentiment d'appartenance identifient les groupes secondaires (*GROUPES SECONDAIRES*).

Figure 10-14
Diagramme double



Dimensions supplémentaires

L'augmentation du nombre de dimensions accroît la quantité de variation prise en compte et peut mettre en évidence des différences masquées dans les solutions possédant un nombre réduit de dimensions. Comme indiqué précédemment, dans une solution bidimensionnelle, les groupes *BANDES* et *GROUPES PRIMAIRES* ne peuvent pas être séparés. Toutefois, vous pouvez augmenter le nombre de dimensions de manière à différencier les deux groupes.

Exécution de l'analyse

- ▶ Pour obtenir une solution tridimensionnelle, affichez de nouveau la boîte de dialogue Composantes principales nominales.
- ▶ Tapez 3 comme nombre de dimensions comprises dans la solution.
- ▶ Cliquez sur OK dans la boîte de dialogue Composantes principales qualitatives.

Récapitulatif des modèles

Figure 10-15
Récapitulatif du modèle

Dimension	Alpha de Cronbach	Variance expliquée	
		Total (valeur propre)	Pourcentage de variance expliquée
1	,885	3,424	68,480
2	-,232	,844	16,871
3	-,459	,732	14,649
Total	1,000[a]	5,000	99,999

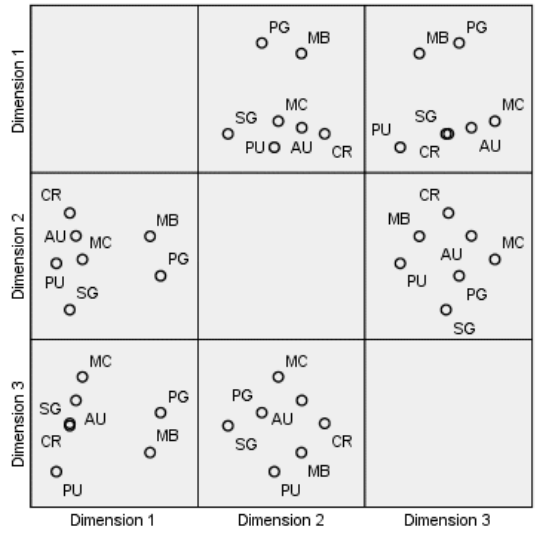
a. La valeur Alpha de Cronbach totale est basée sur la valeur propre totale.

Une solution tridimensionnelle possède les valeurs propres 3,424, 0,844 et 0,732, qui représentent la quasi-totalité de la variance.

Coordonnées principales

Les coordonnées principales de la solution tridimensionnelle sont représentées dans une matrice de diagramme de dispersion. Dans une matrice de diagramme de dispersion, chaque dimension est représentée par rapport à chacune des autres dimensions d'une série de diagrammes de dispersion bidimensionnelles. Les deux premières valeurs propres des trois dimensions diffèrent des valeurs propres de la solution bidimensionnelle ; en d'autres termes, les solutions ne sont pas emboîtées. Etant donné que les valeurs propres des dimensions 2 et 3 sont désormais inférieures à 1 (aboutissant à un alpha de Cronbach négatif), vous devez privilégier la solution bidimensionnelle. La solution tridimensionnelle est proposée à titre d'illustration.

Figure 10-16
Matrice de diagramme de dispersion des coordonnées principales tridimensionnelle



La ligne supérieure des diagrammes indique que la première dimension sépare les groupes *GROUPES PRIMAIRES* et *BANDES* des autres groupes. L'ordre des objets le long de l'axe vertical demeure inchangé d'un diagramme à l'autre dans la ligne supérieure ; chacun de ces diagrammes utilise la dimension 1 comme axe *y*.

La ligne intermédiaire des diagrammes permet d'interpréter la dimension 2. La deuxième dimension a légèrement évolué par rapport à la solution bidimensionnelle. Précédemment, la deuxième dimension possédait trois groupes distincts mais les objets sont désormais davantage répartis le long de l'axe.

La troisième dimension permet de séparer le groupe *BANDES* du groupe *GROUPES PRIMAIRES*, ce qui n'était pas le cas dans la solution bidimensionnelle.

Observez plus attentivement les diagrammes des dimensions 2 et 3 et ceux des dimensions 1 et 2. Dans le plan défini par les dimensions 2 et 3, les objets forment un rectangle approximatif ayant pour sommets *FOULES*, *COMMUNAUTE MODERNE*, *GROUPES SECONDAIRES* et *PUBLIC*. Dans ce plan, *BANDES* et *GROUPES PRIMAIRES* apparaissent comme des combinaisons convexes de *PUBLIC-FOULES* et de *GROUPES SECONDAIRES-COMMUNAUTE MODERNE*, respectivement. Toutefois, comme indiqué précédemment, ils sont séparés des autres groupes le long de la dimension 1. Le groupe *AUDIENCES* n'est pas séparé des autres groupes le long de la dimension 1 et apparaît sous la forme d'une combinaison des groupes *FOULES* et *COMMUNAUTE MODERNE*.

Saturations

Figure 10-17
Corrélations entre composantes tridimensionnelles

	Dimension		
	1	2	3
INTENSITE	,980	-,005	-,201
FREQUENCE	,521	-,643	,561
APPARTENANCE	,980	-,002	-,197
PROXIMITE	,519	,656	,549
PROCEDURE	,981	,004	-,193

Normalisation principale de la variable.

Le fait de savoir comment les objets sont séparés ne permet pas de connaître la correspondance entre variables et dimensions. Pour ce faire, vous devez recourir aux corrélations entre composantes. La première dimension correspond essentiellement aux groupes *Sentiment d'appartenance*, *Intensité de l'interaction* et *Formalité de la relation* ; la deuxième sépare les groupes *Fréquence d'interaction* et *Proximité physique*; la troisième dimension sépare ceux-ci des autres groupes.

Exemple : Symptomatologie des troubles du comportement alimentaire

Les troubles du comportement alimentaire sont des maux débilitants associés à un mauvais comportement alimentaire, à une grave déformation de l'image du corps et à une obsession du poids affectant simultanément l'esprit et le corps. Des millions de personnes, notamment les adolescents, sont affectées chaque année. Des traitements sont disponibles et la plupart d'entre eux sont efficaces si le trouble est identifié tôt.

Un médecin peut essayer de diagnostiquer un trouble du comportement alimentaire par le biais d'une évaluation psychologique et médicale. Toutefois, il peut s'avérer difficile de cataloguer un patient dans l'une des différentes classes de troubles du comportement alimentaire car il n'existe pas de symptomatologie standardisée du comportement anorexique/boulimique. Existe-t-il des symptômes qui permettent de classer facilement les patients dans l'un des quatre groupes ? Quels symptômes ont-ils en commun ?

Pour tenter de répondre à ces questions, des chercheurs (Van der Ham, Meulman, Van Strien, et Van Engeland, 1997) ont réalisé une étude sur 55 adolescents souffrant de troubles du comportement alimentaire connus (tableau ci-dessous).

Table 10-2
Diagnostics des patients

Diagnostic	**Nombre de patients**
Anorexie mentale	25
Anorexie avec boulimie mentale	9
Boulimie mentale après anorexie	14
Trouble atypique du comportement alimentaire	7
Total	55

Chaque patient a été observé quatre fois sur une période de quatre années, soit un total de 220 observations. A chaque observation, les patients ont été notés pour chacun des 16 symptômes présentés dans le tableau ci-après. En raison de l'absence de scores de symptôme pour le patient 71/visite 2, le patient 76/visite 2 et le patient 47/visite 3, le nombre d'observations valides est de 217. Les données sont disponibles dans *anorectic.sav*.Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*.

Table 10-3
Sous-échelles Morgan-Russell modifiées mesurant le bien-être

Nom de variable	l'étiquette Variable	Limite inférieure (score 1)	Limite supérieure (score 3 ou 4)
poids	Poids corporel	Hors de l'intervalle de poids normal	Normale
mens	Menstruation	Aménorrhée	Règles régulières
inappétance	Perte de l'appétit (inappétance)	Moins de 1 200 calories	Repas normaux/réguliers
frénésie alimentaire	Frénésie alimentaire	Plus d'une fois par semaine	Aucune frénésie alimentaire
vomissement	Vomissement	Plus d'une fois par semaine	Pas de vomissement
laxatifs	Laxatifs	Plus d'une fois par semaine	Pas de laxatifs
hyperactivité	Hyperactivité	Ne peut pas demeurer inactif	Pas d'hyperactivité
famille	Relations familiales	Mauvaises	Bonnes
éman	Emancipation par rapport à la famille	Forte dépendance	Suffisante
amis	Relations amicales	Pas de bons amis	Au moins deux bons amis
école	Antécédents scolaires/professionnels	A quitté l'école/le travail	Antécédents moyens à bons
atts	Attitude sexuelle	Inadéquate	Suffisante
comps	Comportement sexuel	Inadéquate	Apprécie les rapports sexuels
humeur	Etat mental (humeur)	Très déprimé	Normale
préo	Préoccupation nourriture et poids	Totale	Aucune préoccupation
corps	Perception du corps	Perturbée	Normale

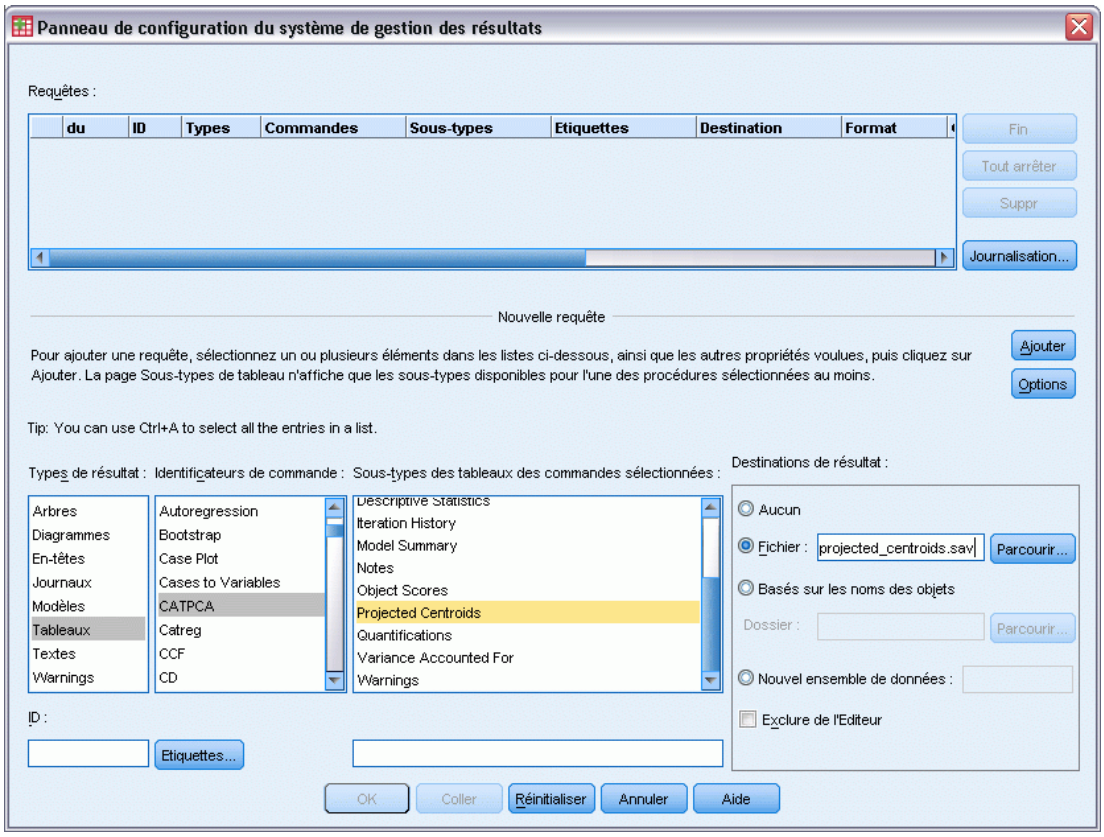
L'analyse en composantes principales est idéale pour cette situation, dans la mesure où la finalité de l'étude est d'établir les relations entre les symptômes et les différentes classes de troubles du comportement alimentaire. En outre, l'analyse en composantes principales qualitatives est susceptible d'être plus utile que l'analyse en composantes principales classique car les symptômes sont notés sur une échelle ordinale.

Exécution de l'analyse

Afin d'examiner correctement la structure de l'évolution de la maladie pour chaque diagnostic, vous pouvez faire en sorte que les résultats du tableau des centres de gravité projetés soient disponibles en tant que données pour les diagrammes de dispersion. Pour ce faire, utilisez le système de gestion des résultats (OMS).

▶ Pour démarrer une requête OMS, dans les menus, choisissez :
Utilitaires > Panneau de configuration du système de gestion des résultats...

Figure 10-18
Panneau de configuration du système de gestion des résultats



▶ Sélectionnez l'option Tableaux comme type de résultat.

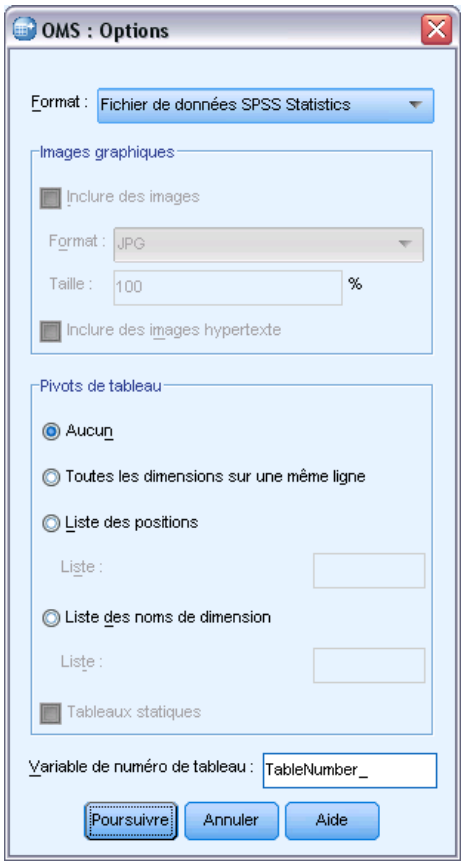
▶ Sélectionnez l'option CATPCA comme commande.

▶ Sélectionnez l'option Centres de gravité projetés comme type de tableau.

▶ Sélectionnez l'option Fichier dans le groupe Destinations de sortie, puis tapez projected_centroids.sav comme nom de fichier.

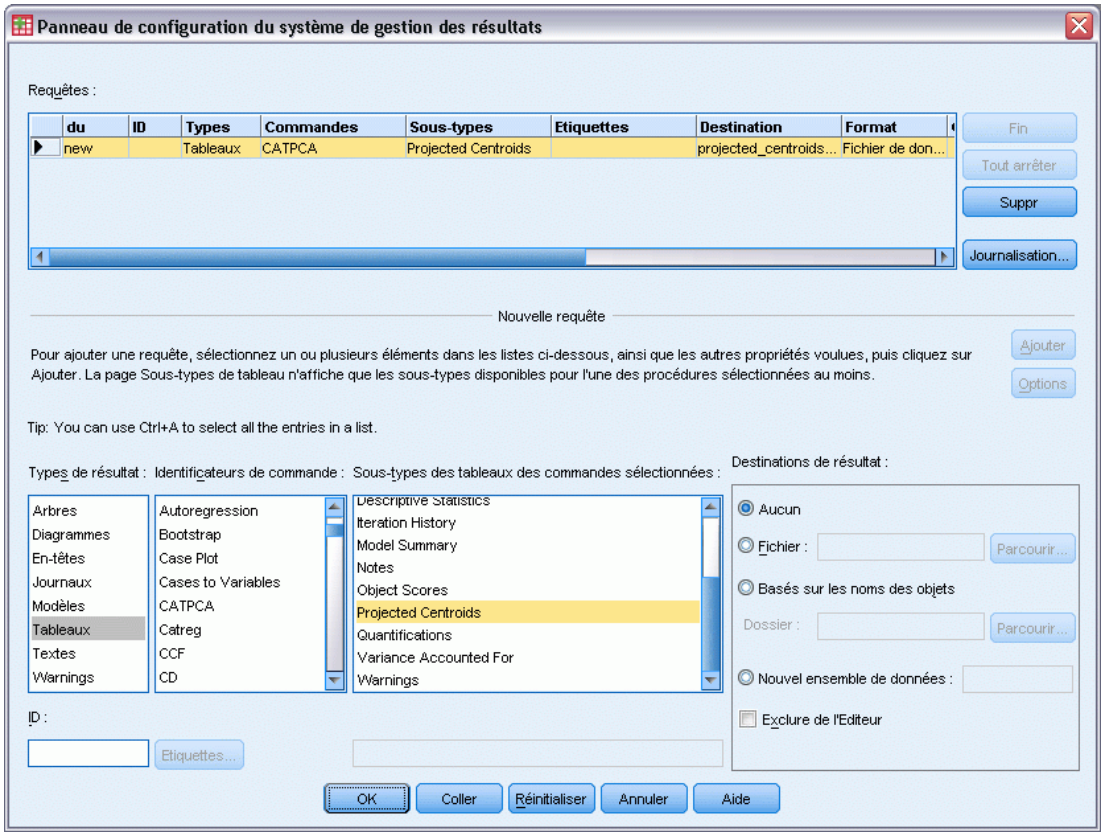
▶ Cliquez sur Options.

Figure 10-19
Boîte de dialogue Options



- Sélectionnez l'option Fichier de données IBM® SPSS® Statistics comme format de résultat.
- Tapez TableNumber_1 comme variable de numéro de tableau.
- Cliquez sur Poursuivre.

Figure 10-20
Panneau de configuration du système de gestion des résultats

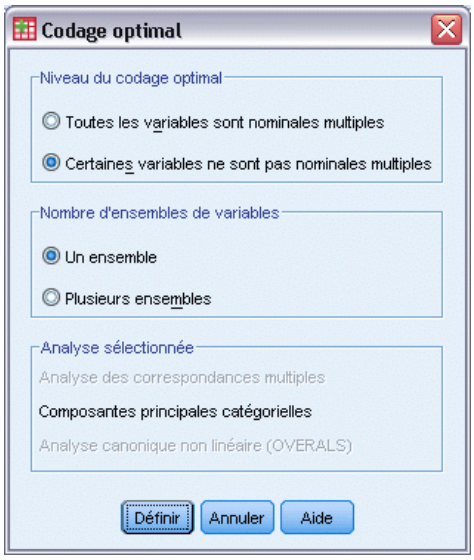


- Cliquez sur Ajouter.
- Cliquez sur OK, puis de nouveau sur OK pour confirmer la session OMS.

  Le système de gestion des résultats est désormais configuré pour écrire les résultats du tableau des centres de gravité projetés dans le fichier *projected_centroids.sav*.
- Pour générer un résultat de composants principaux qualitatifs pour cet ensemble de données, choisissez dans les menus :

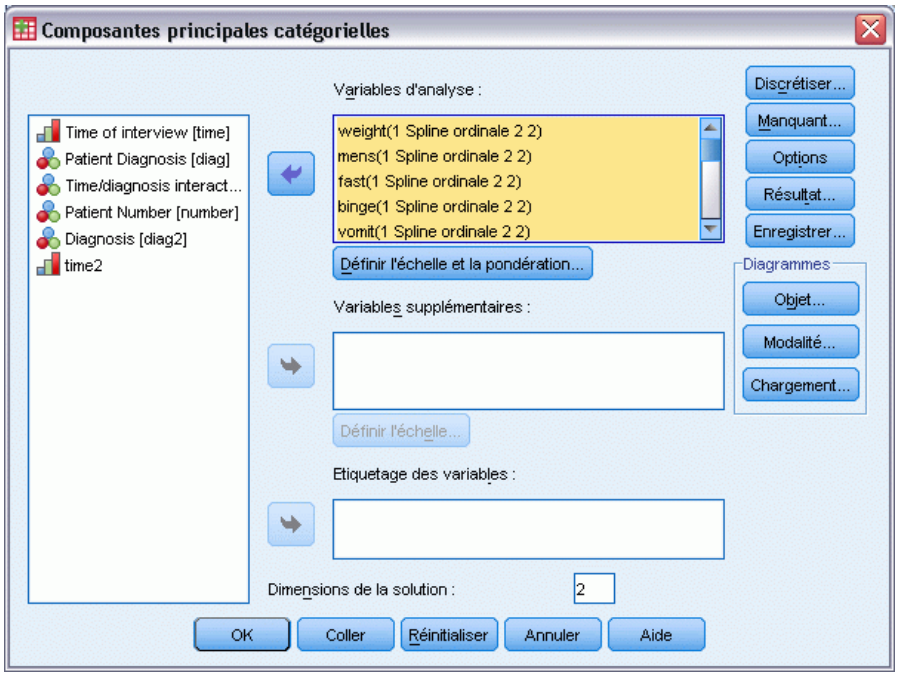
  Analyse > Réduction des dimensions > Codage optimal

Figure 10-21
Boîte de dialogue Niveau du codage optimal



- Sélectionnez l'option Certaines variables non nominales multiples dans le groupe Niveau du codage optimal.
- Cliquez sur Définir.

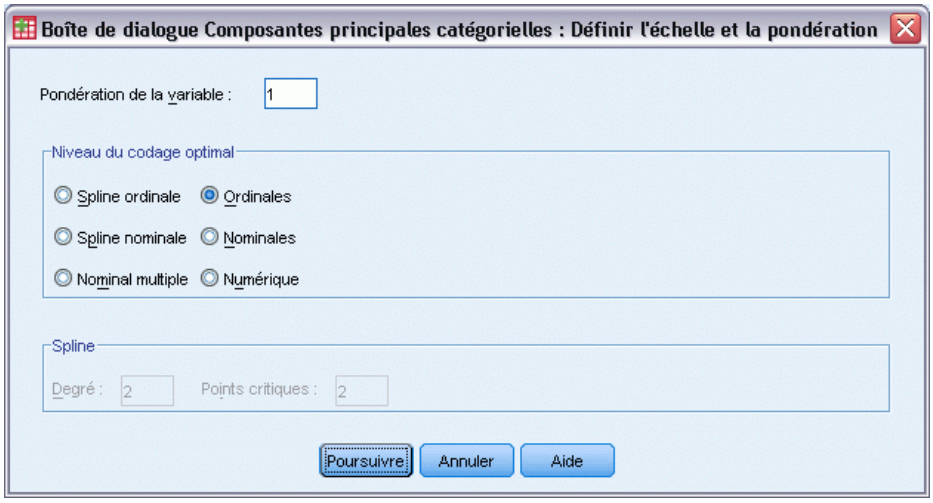
Figure 10-22
Boîte de dialogue Composantes principales qualitatives



- Sélectionnez les options allant de *Poids corporel* à *Perception du corps* comme variables d'analyse.

▶ Cliquez sur Définir l'échelle et la pondération.

Figure 10-23
Définir l'échelle et la pondération

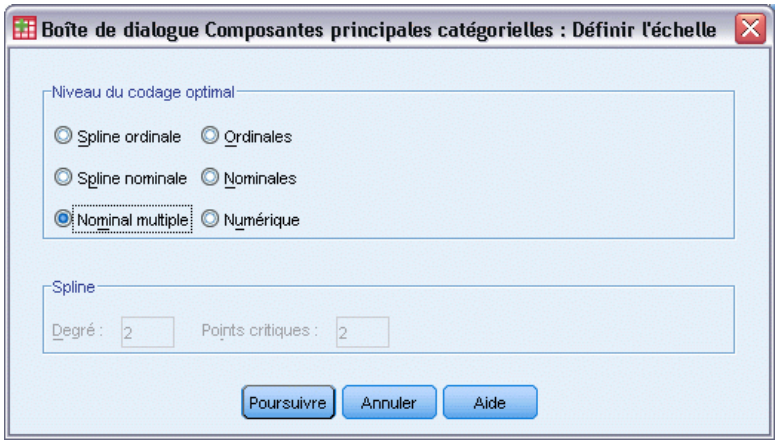


▶ Sélectionnez l'option Ordinal comme niveau de codage optimal.

▶ Cliquez sur Poursuivre.

▶ Sélectionnez l'option *Interaction moment/diagnostic* comme variable supplémentaire, puis cliquez sur Définir l'échelle dans la boîte de dialogue Composantes principales qualitatives.

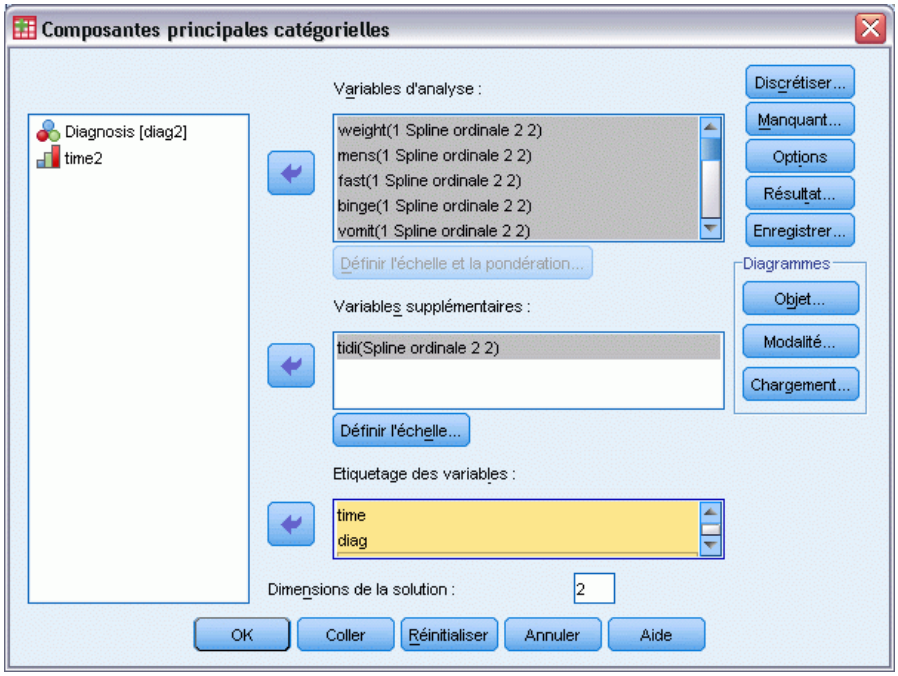
Figure 10-24
Boîte de dialogue Définir l'échelle



▶ Sélectionnez l'option Variables nominales multiples comme niveau de codage optimal.

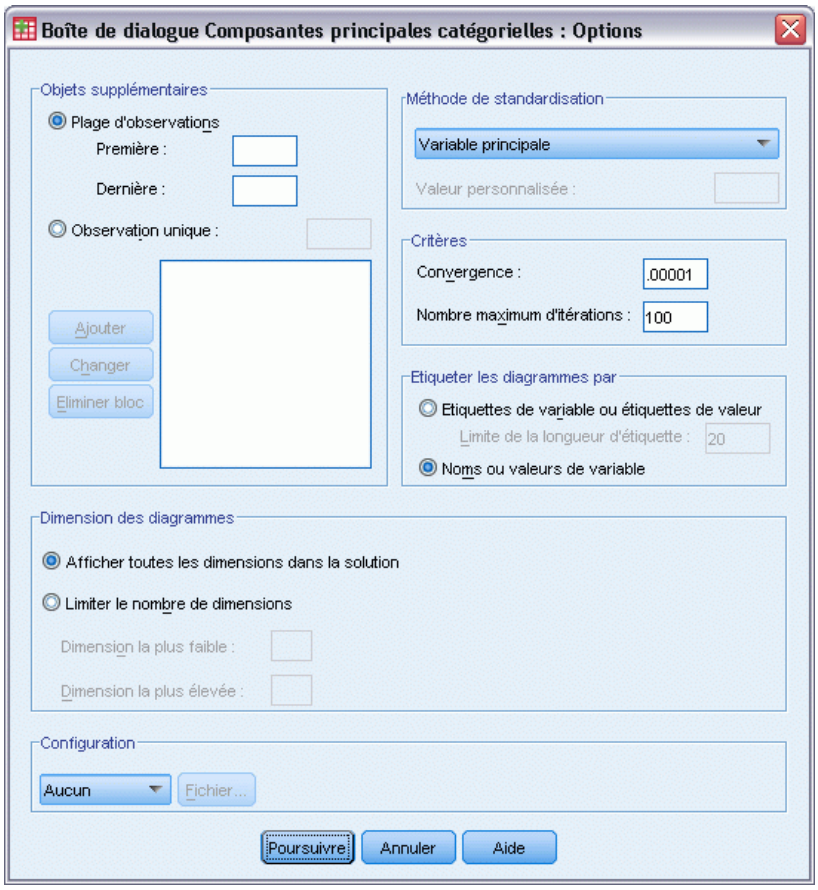
▶ Cliquez sur Poursuivre.

Figure 10-25
Boîte de dialogue Composantes principales qualitatives



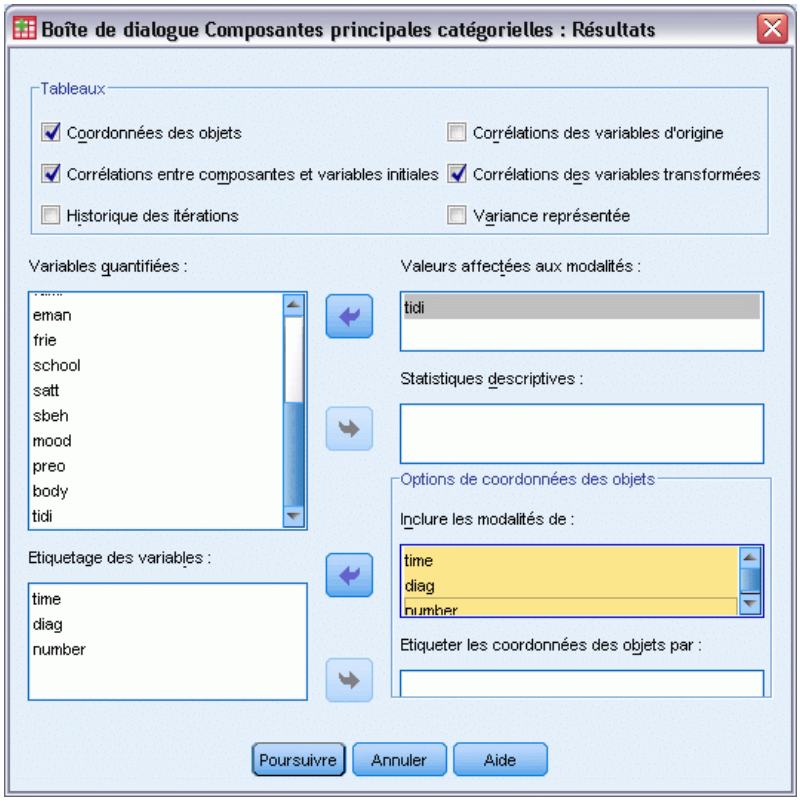
- Sélectionnez les options allant de *Moment de l'entrevue* à *Numéro de patient* comme variables d'étiquetage.
- Cliquez sur Options.

Figure 10-26
Boîte de dialogue Options



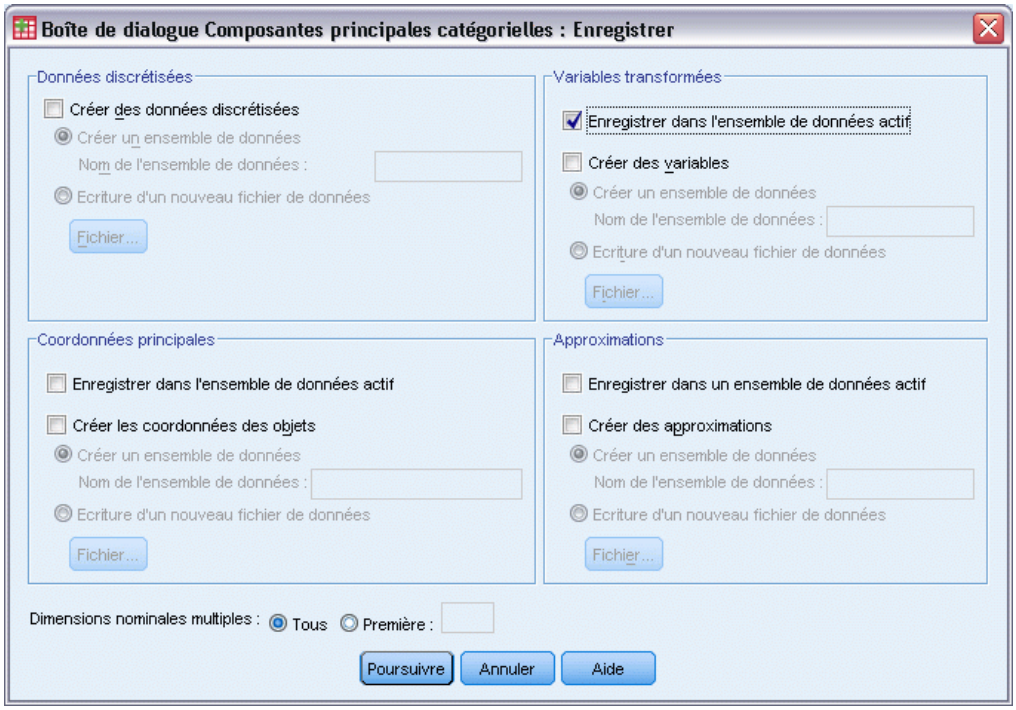
- Choisissez la méthode d'étiquetage des diagrammes Noms ou valeurs de variable.
- Cliquez sur Poursuivre.
- Cliquez sur Résultat dans la boîte de dialogue Composantes principales qualitatives.

Figure 10-27
Résultat



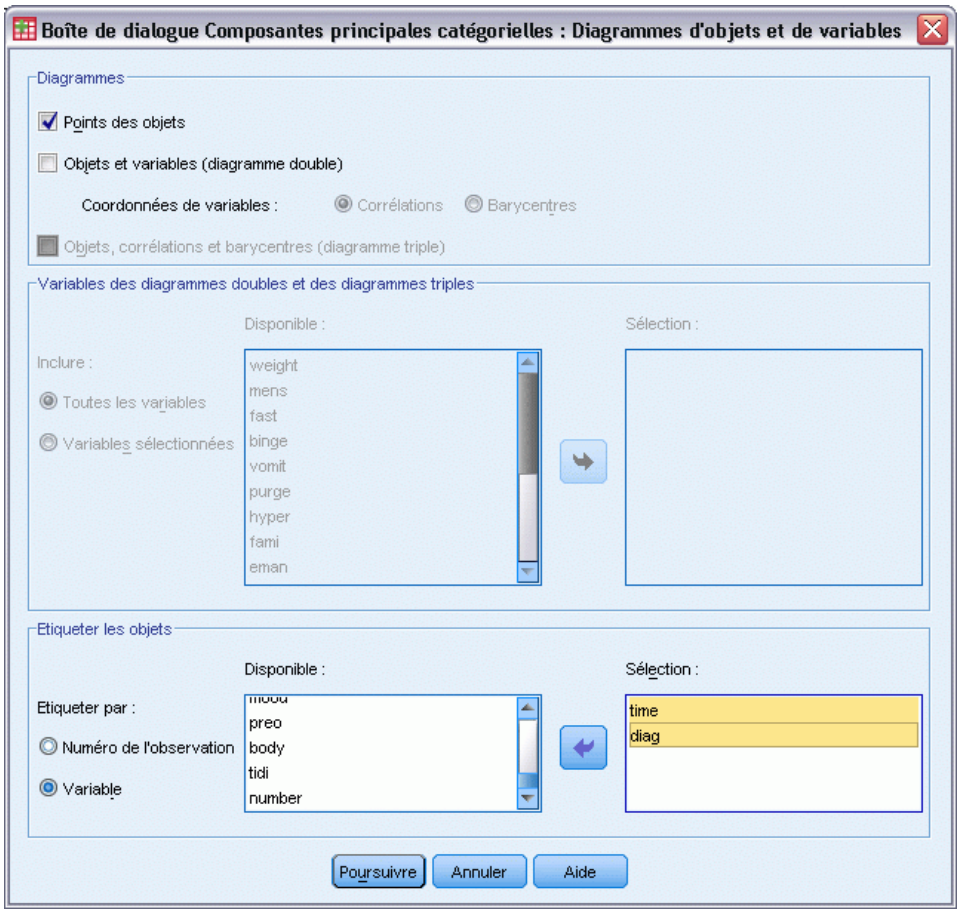
- Sélectionnez l'option Coordonnées principales dans le groupe Tableaux.
- Indiquez que vous souhaitez obtenir les valeurs affectées aux modalités pour la variable *moment/diagnostic*.
- Incluez les modalités *moment*, *diag* et *nombre*.
- Cliquez sur Poursuivre.
- Cliquez sur Enregistrer dans la boîte de dialogue Composantes principales qualitatives.

Figure 10-28
Boîte de dialogue Enregistrer



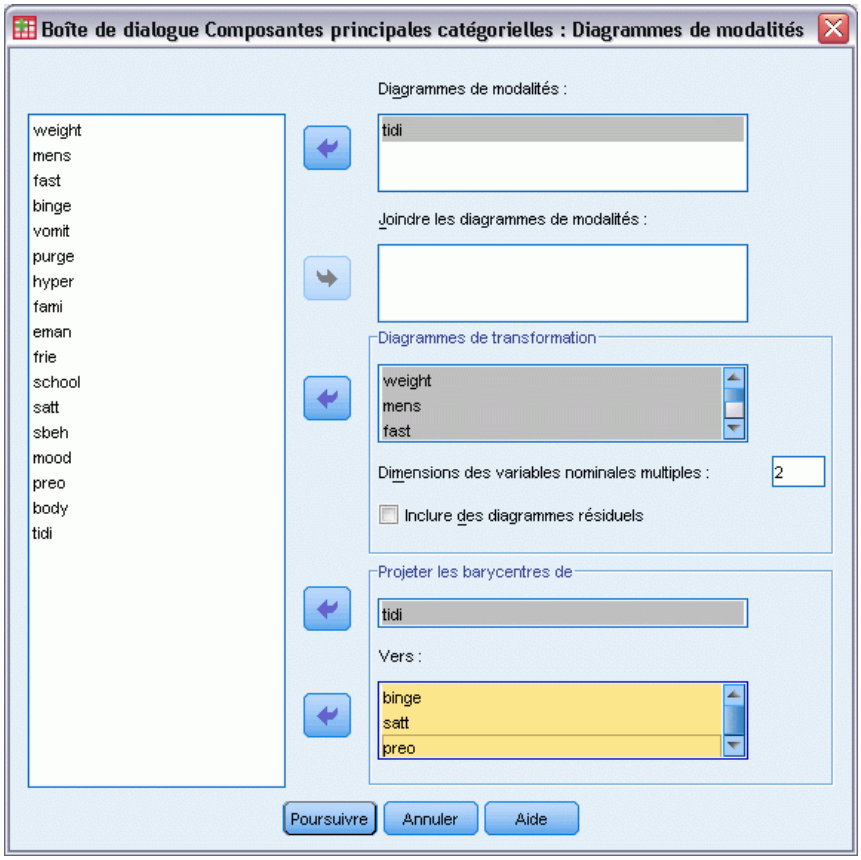
- Dans le groupe des variables transformées, sélectionnez Enregistrer dans l'ensemble de données actif.
- Cliquez sur Poursuivre.
- Cliquez sur Objet dans la boîte de dialogue Composantes principales qualitatives.

Figure 10-29
Diagrammes d'objets et de variables



- Choisissez l'option d'étiquetage des objets Variable.
- Sélectionnez les options *moment* et *diag* comme variables d'étiquetage des objets.
- Cliquez sur Poursuivre.
- Cliquez sur Modalité dans la boîte de dialogue Composantes principales qualitatives.

Figure 10-30
Boîte de dialogue Diagrammes de modalités

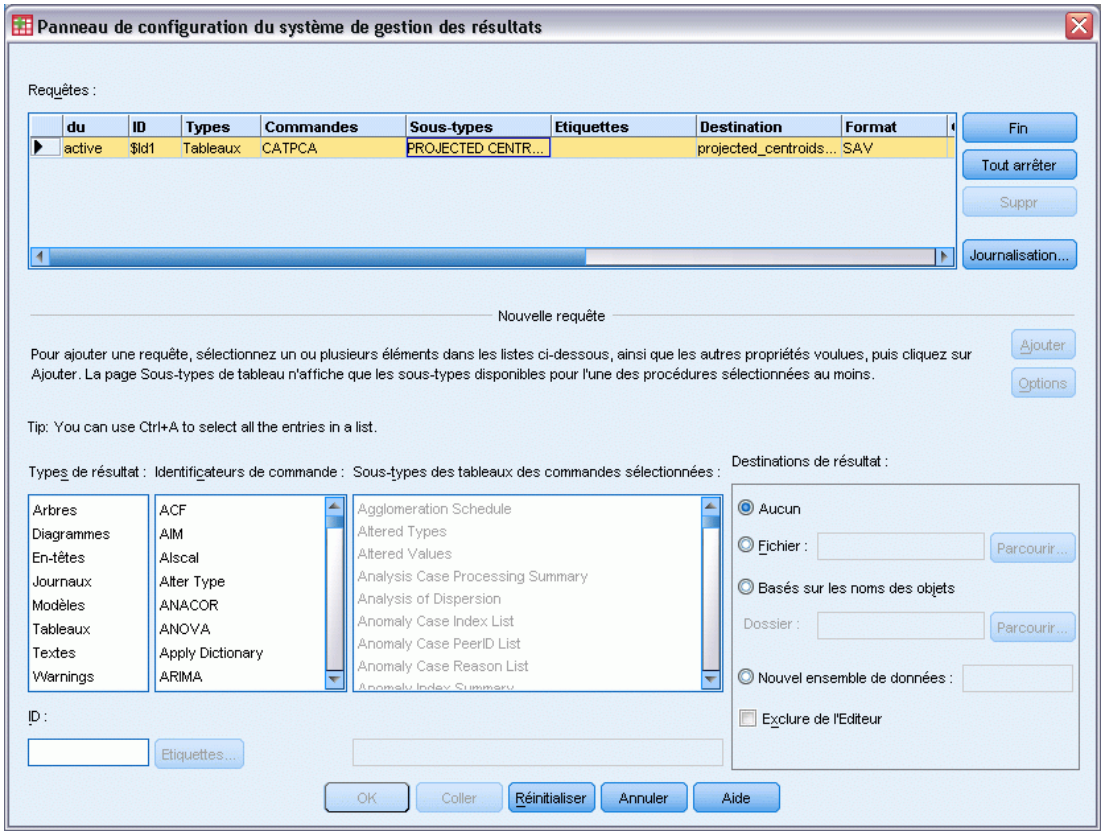


- Indiquez que vous souhaitez obtenir les diagrammes de modalité pour la variable *moment/diagnostic*.
- Indiquez que vous souhaitez obtenir les diagrammes de transformation pour les variables allant de *poids* à *corps*.
- Projetez les centres de *moment/diagnostic* sur *frénésie alimentaire*, *atts* et *préo*.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Composantes principales qualitatives.

  La procédure aboutit à des scores pour les sujets (de moyenne 0 et de variance unitaire) et à des valeurs affectées aux modalités qui maximisent la corrélation des carrés des moyennes des scores de sujet et les variables transformées. Dans l'analyse actuelle, les valeurs affectées aux modalités ont été contraintes de manière à refléter les informations ordinales.

En dernier lieu, pour écrire les informations du tableau des centres de gravité projetés dans le fichier *projected_centroids.sav*, vous devez mettre fin à la requête OMS. Affichez de nouveau le panneau de configuration du système de gestion des résultats.

Figure 10-31
Panneau de configuration du système de gestion des résultats

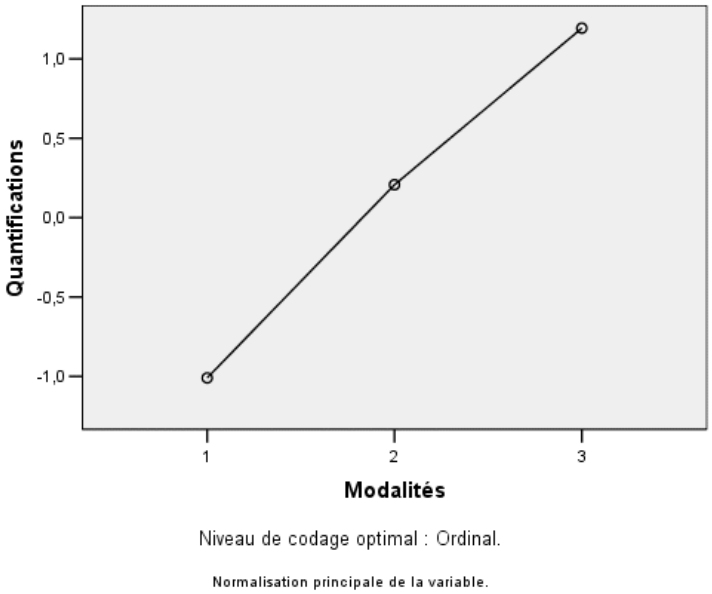


- Cliquez sur Fin.
- Cliquez sur OK, puis de nouveau sur OK pour confirmer.

Diagrammes de transformation

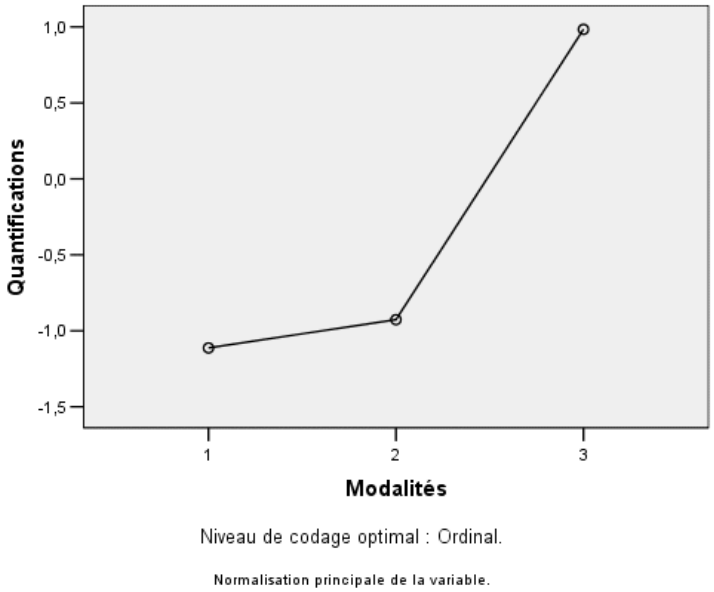
Les diagrammes de transformation affichent le numéro de modalité initial sur les axes horizontaux ; les axes verticaux donnent les quantifications optimales.

Figure 10-32
Diagramme de transformation pour les menstruations



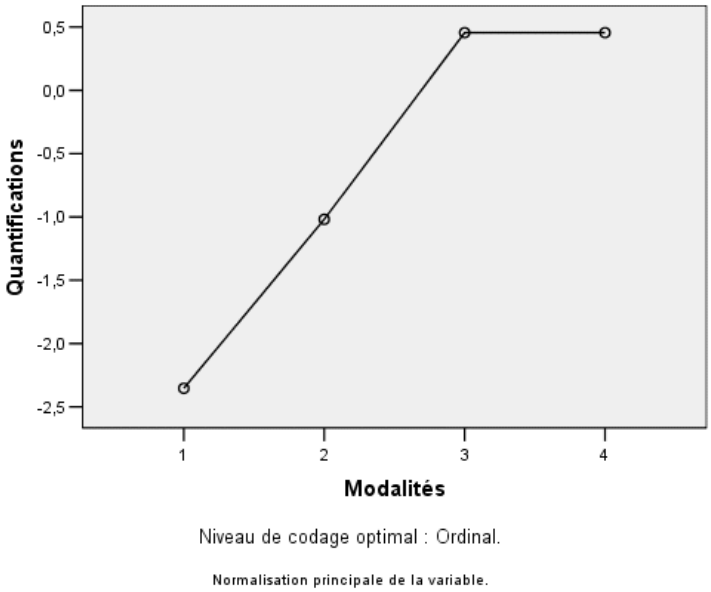
Certaines variables, telles que *Menstruation*, ayant obtenu des transformations presque linéaires, vous pouvez, dans cette analyse, les interpréter comme des données numériques.

Figure 10-33
Diagramme de transformation des antécédents scolaires/professionnels



Les quantifications des autres variables, telles que *Antécédents scolaires/professionnels*, n'ont pas obtenu de transformations linéaires et doivent être interprétées au niveau de codage ordinal. La différence entre les deuxième et troisième modalités est beaucoup plus importante que celle entre les première et deuxième modalités.

Figure 10-34
Diagramme de transformation de la frénésie alimentaire



Une situation intéressante se présente dans les quantifications de la *frénésie alimentaire*. La transformation obtenue est linéaire pour les modalités 1 à 3, mais les valeurs quantifiées pour les modalités 3 et 4 sont égales. Ce résultat montre que les scores 3 et 4 ne font pas de différences entre les patients et suggère que vous pouvez utiliser le niveau de codage numérique dans une solution à deux composantes en recodant les scores 4 en 3.

Récapitulatif des modèles

Figure 10-35
Récapitulatif du modèle

Récapitulatif du modèle

Dimension	Alpha de Cronbach	Variance expliquée	
		Total (valeur propre)	Pourcentage de variance expliquée
1	,874	5,550	34,690
2	,522	1,957	12,234
Total	,925[a]	7,508	46,924

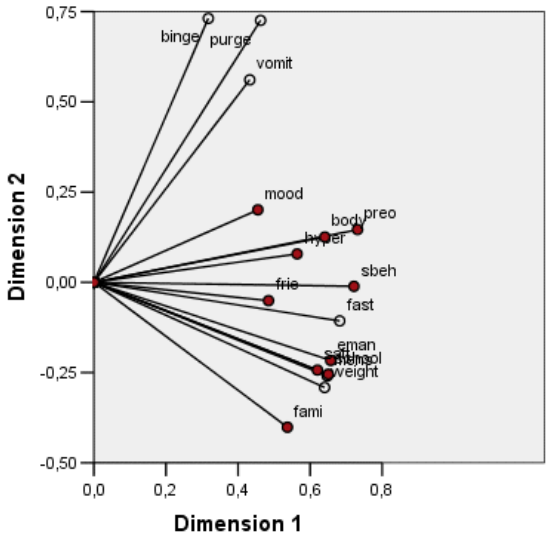
a. La valeur Alpha de Cronbach totale est basée sur la valeur propre totale.

Pour évaluer l'adéquation du modèle par rapport aux données, consultez le récapitulatif du modèle. Environ 47 % de la variance totale est expliquée par le modèle à deux composantes, à raison de 35 % par la première dimension et de 12 % par la deuxième. Par conséquent, presque la moitié de la variabilité au niveau des différents objets est expliquée par le modèle à deux composantes.

Saturations

Pour démarrer l'interprétation des deux dimensions de votre solution, observez les corrélations entre composantes. Toutes les variables possèdent une corrélation entre composantes positives dans la première dimension, ce qui signifie qu'il existe un facteur commun corrélé positivement avec toutes les variables.

Figure 10-36
Diagramme des corrélations entre composantes



Normalisation principale de la variable.

La deuxième dimension sépare les variables. Les variables *Frénésie alimentaire*, *Vomissement* et *Laxatifs* forment un groupe possédant des corrélations entre composantes positives élevées dans la deuxième dimension. Ces symptômes sont généralement considérés comme représentatifs d'un comportement boulimique.

Les variables *Emancipation par rapport à la famille*, *Antécédents scolaires/professionnels*, *Attitude sexuelle*, *Poids corporel* et *Menstruations* forment un autre groupe, dans lequel vous pouvez inclure les variables *Perte de l'appétit (inappétance)* et *Relations familiales* car leurs vecteurs sont proches de la classe principale, et ces variables sont considérées comme étant des symptômes de l'anorexie (inappétance, poids, menstruation) ou de nature psychosociale (émancipation, antécédents scolaires/professionnels, attitude sexuelle, relations familiales). Les vecteurs de ce groupe sont orthogonaux (perpendiculaires) aux vecteurs de la frénésie alimentaire, du vomissement et des laxatifs, ce qui signifie que cet ensemble de variables n'est pas corrélé avec l'ensemble des variables de la boulimie.

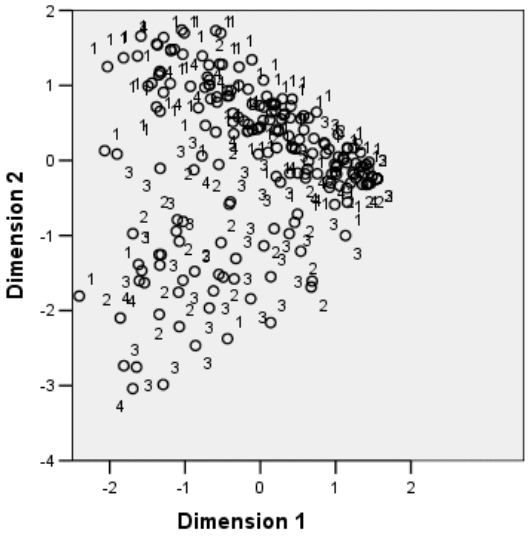
Les variables *Relations amicales*, *Etat mental (humeur)* et *Hyperactivité* ne semblent pas s'adapter correctement à la solution. Vous pouvez le constater dans le diagramme en observant les longueurs de chaque vecteur. La longueur du vecteur d'une variable donnée correspond à son ajustement, et ces variables possèdent les vecteurs les plus courts. Dans le cadre d'une solution à deux composantes, vous retireriez probablement ces variables de l'ébauche d'une symptomatologie des troubles du comportement alimentaire. Toutefois, elles peuvent mieux s'intégrer dans une solution impliquant davantage de dimensions.

Les variables *Comportement sexuel*, *Préoccupation nourriture et poids* et *Perception du corps* forment un autre groupe théorique de symptômes, liés à la perception que le patient a de son corps. Tout en étant corrélées avec les deux groupes de variables orthogonaux, ces variables possèdent des vecteurs assez longs et sont étroitement associées à la première dimension ; par conséquent, elles peuvent fournir certaines informations utiles sur le facteur “commun”.

Coordonnées principales

Le schéma suivant illustre un diagramme des coordonnées principales, dans lequel les sujets sont étiquetés d’après leur modalité de diagnostic.

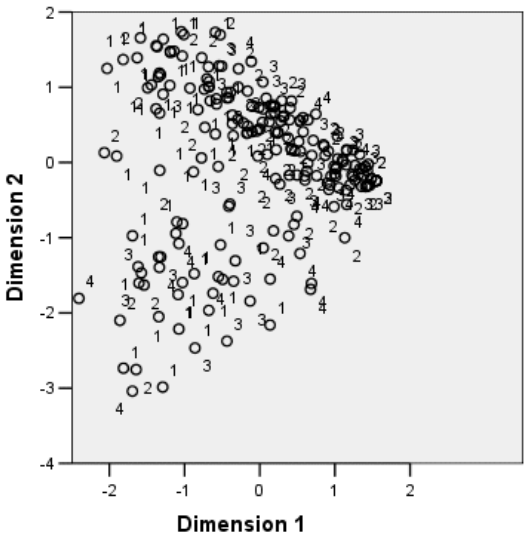
Figure 10-37
Diagramme des coordonnées principales étiqueté en fonction du diagnostic



Ce diagramme ne permet pas d’interpréter la première dimension car les patients ne sont pas séparés par diagnostic le long de celle-ci. Toutefois, il comprend certaines informations sur la deuxième dimension. Les sujets anorexiques (1) et les patients présentant un trouble atypique du comportement alimentaire (4) forment un groupe, situé au-dessus des sujets souffrant d’une forme de boulimie (2 et 3). Par conséquent, la deuxième dimension sépare les patients boulimiques des autres, comme l’a également indiqué la section précédente (les variables du groupe boulimique possèdent des corrélations entre composantes positives élevées dans la deuxième dimension). Cela est cohérent dans la mesure où les saturations des symptômes traditionnellement associés à la boulimie possèdent des valeurs élevées dans la deuxième dimension.

Le schéma suivant illustre un diagramme des coordonnées principales, dans lequel les sujets sont étiquetés d'après le moment de leur diagnostic.

Figure 10-38
Coordonnées principales étiquetées en fonction du moment de l'entrevue



L'étiquetage des coordonnées principales d'après le moment met en évidence que la première dimension possède une relation au moment, car il semble y avoir une progression des moments de diagnostic entre les 1 essentiellement vers la gauche et les autres vers la droite. Vous pouvez lier les points dans le temps au sein de ce diagramme ; pour ce faire, enregistrez les coordonnées principales et créez un diagramme de dispersion en utilisant les scores de la dimension 1 sur l'axe des *x*, les scores de la dimension 2 sur l'axe des *y* et en définissant des marques à partir des numéros de patient.

La comparaison du diagramme des coordonnées principales étiqueté en fonction du moment à celui étiqueté d'après le diagnostic peut vous donner une idée des objets inhabituels. Par exemple, dans le diagramme étiqueté en fonction du moment, il existe un patient dont le diagnostic au moment 4 figure à gauche de tous les autres points du diagramme. Cela est peu courant car, d'après la tendance générale des points, les moments les plus récents figurent plus à droite. Il est intéressant de constater que ce point, dont le moment semble mal positionné, possède également un diagnostic inhabituel, en ce sens que le patient est un anorexique dont les scores le placent dans le groupe des boulimiques. Le tableau des coordonnées principales indique qu'il s'agit du patient 43, chez qui a été diagnostiquée une anorexie mentale, et dont les coordonnées principales sont indiquées dans le tableau ci-après.

Table 10-4
Coordonnées principales du patient 43

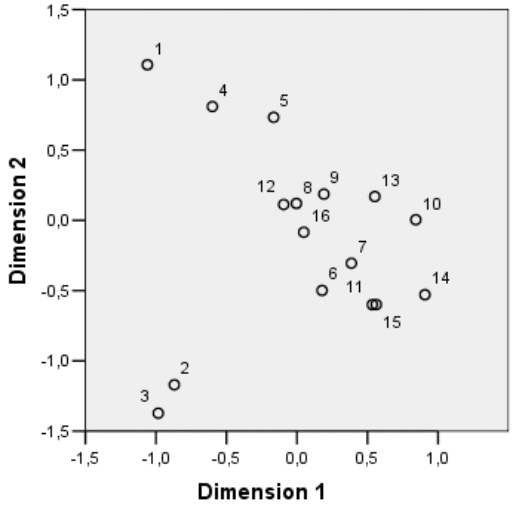
Heure	Dimension 1	Dimension 2
1	–2.031	1.250
2	–2.067	0.131
3	–1.575	–1.467
4	–2.405	–1.807

Les scores du patient au moment 1 sont prototypiques des anorexiques : le score négatif élevé dans la dimension 1 correspond à une mauvaise image du corps et le score positif dans la dimension 2 correspond à des symptômes d'anorexie ou à un comportement psychosocial perturbé. Toutefois, à la différence de la majorité des patients, la progression est faible ou nulle dans la dimension 1. Dans la dimension 2, il semble y avoir une certaine progression vers la "normale" (autour de 0, entre un comportement anorexique et boulimique), mais ensuite le patient présente des symptômes de boulimie.

Examen de la structure de l'évolution de la maladie

Pour que vous puissiez mieux comprendre les liens unissant les deux dimensions aux quatre modalités de diagnostic et aux quatre points dans le temps, une variable supplémentaire *Interaction moment/diagnostic* a été créée par une classification croisée des quatre modalités de *Diagnostics des patients* et des quatre modalités de *Moment de l'entrevue*. Par conséquent, la variable *Interaction moment/diagnostic* possède 16 modalités, dont la première représente les patients atteints d'anorexie mentale à leur première visite. La cinquième modalité représente les patients atteints d'anorexie mentale au point 2 dans le temps, et ainsi de suite jusqu'à la seizième modalité, qui représente les patients souffrant d'un trouble atypique du comportement alimentaire au point 4 dans le temps. L'utilisation de la variable supplémentaire *Interaction moment/diagnostic* permet d'étudier l'évolution dans le temps des maladies affectant les différents groupes. La variable possède un niveau de codage nominal multiple et le schéma ci-après illustre les points de modalité.

Figure 10-39
Points de modalité de l'interaction moment/diagnostic

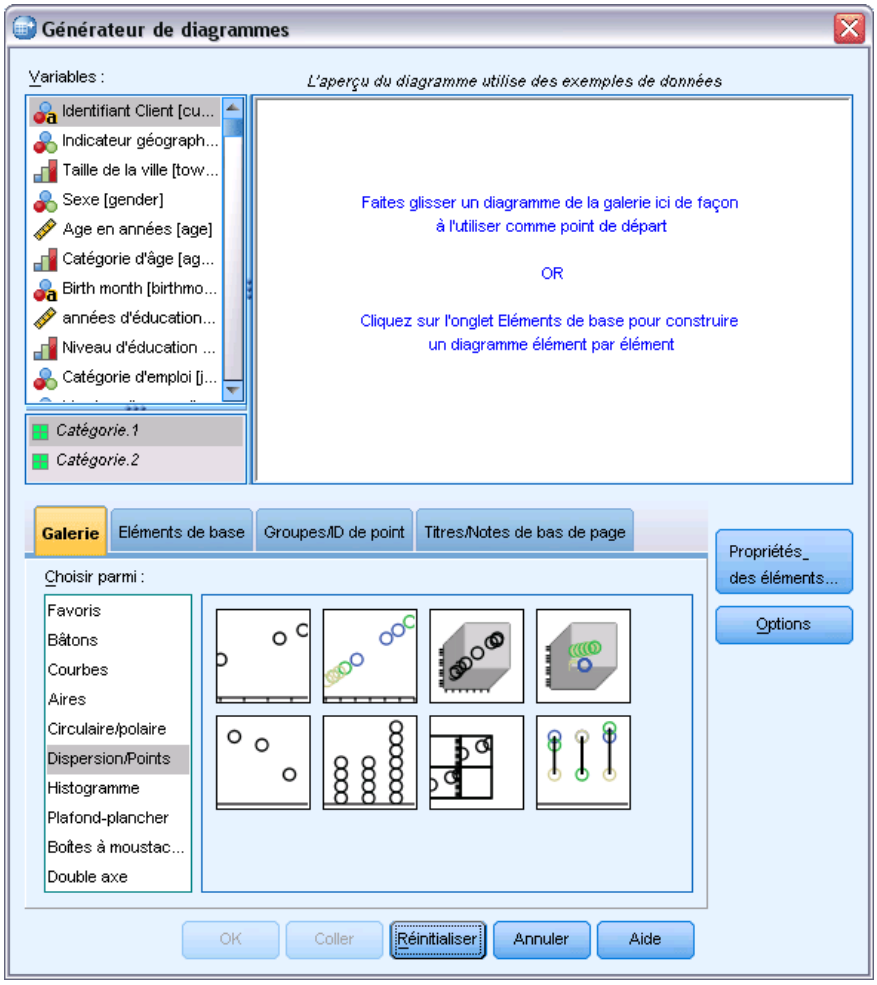


Une partie de la structure apparaît dans ce diagramme : les modalités de diagnostic au point 1 dans le temps séparent nettement l'anorexie mentale et le trouble atypique du comportement alimentaire de l'anorexie mentale avec boulimie mentale et de la boulimie mentale après anorexie mentale dans la deuxième dimension. Au-delà, il est un peu plus difficile de discerner les modèles.

Toutefois, vous pouvez faciliter la lisibilité des modèles en créant un diagramme de dispersion basé sur les quantifications. Pour ce faire, dans les menus, choisissez :

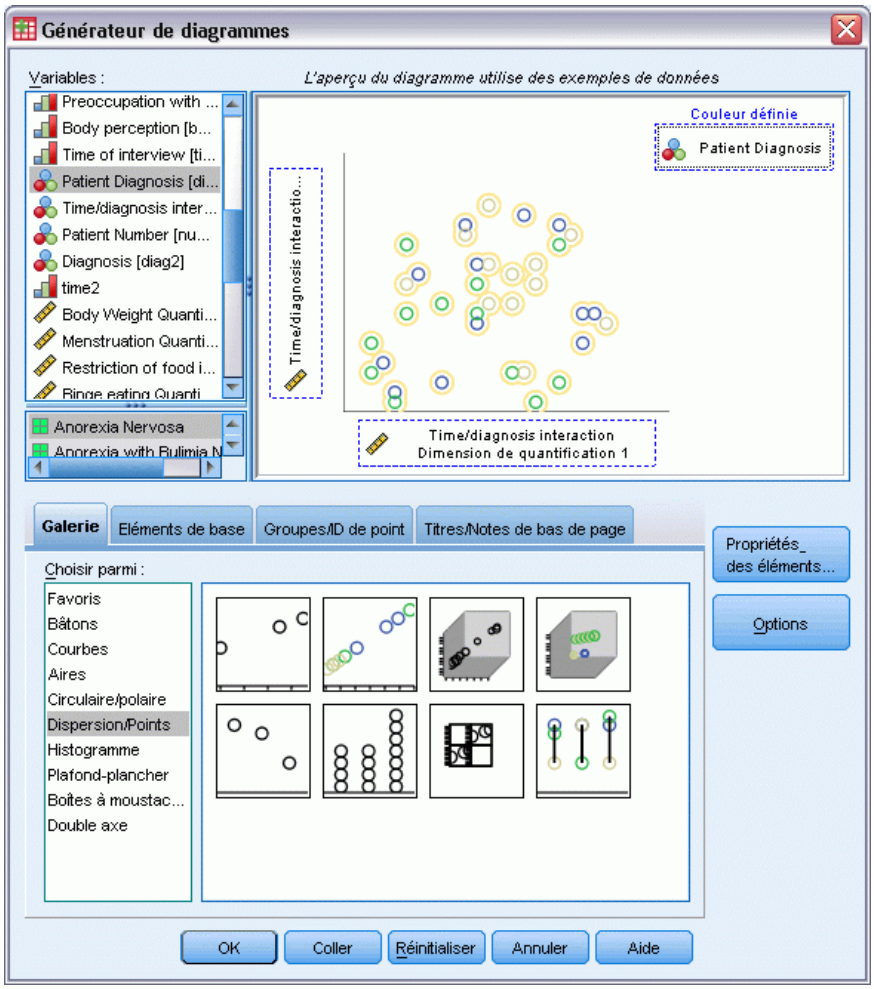
Graphes > Générateur de diagrammes...

Figure 10-40
Galerie Dispersion/Points



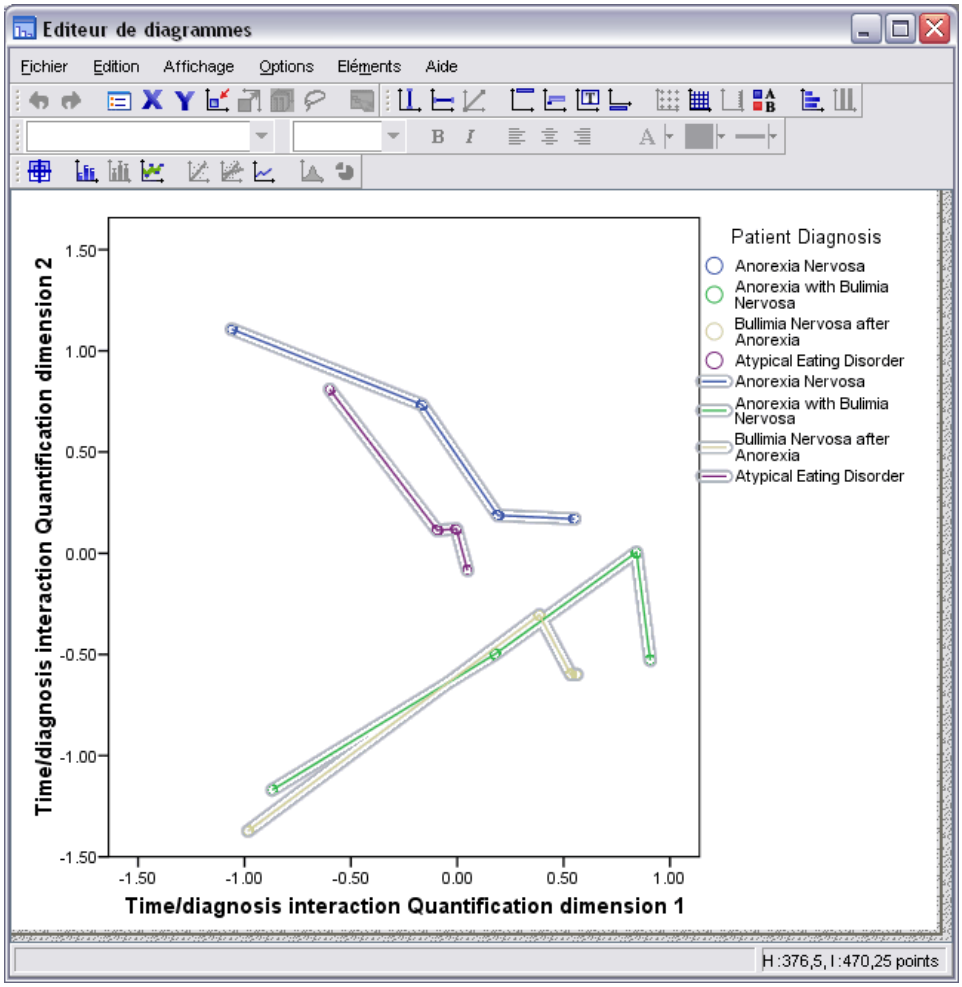
- Sélectionnez la galerie Dispersion/Points et choisissez Diagramme de dispersion regroupé.

Figure 10-41
Générateur de diagrammes



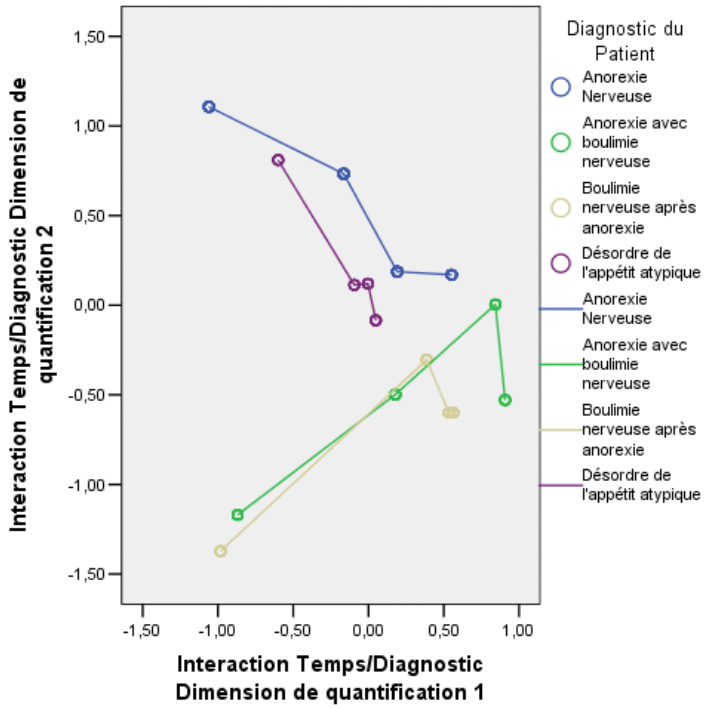
- Sélectionnez *Interaction moment/diagnostic Quantification dimension 2* comme variable de l'axe *y* et *Interaction moment/diagnostic Quantification dimension 1* comme variable de l'axe *x*.
- Pour la définition des couleurs, choisissez l'option *Diagnostics des patients*.
- Cliquez sur OK.

Figure 10-42
Structures de l'évolution des maladies



- Ensuite, pour relier les points, double-cliquez sur le graphique, puis cliquez sur l'onglet Ajouter une courbe d'interpolation dans l'éditeur de diagrammes.
- Fermez l'éditeur de diagrammes.

Figure 10-43
Structures de l'évolution des maladies



Une fois que vous avez relié les points de chaque modalité de diagnostic dans le temps, les motifs suggèrent que la première dimension est associée au moment et la deuxième au diagnostic, comme vous l'avez précédemment déterminé à partir des diagrammes des coordonnées principales.

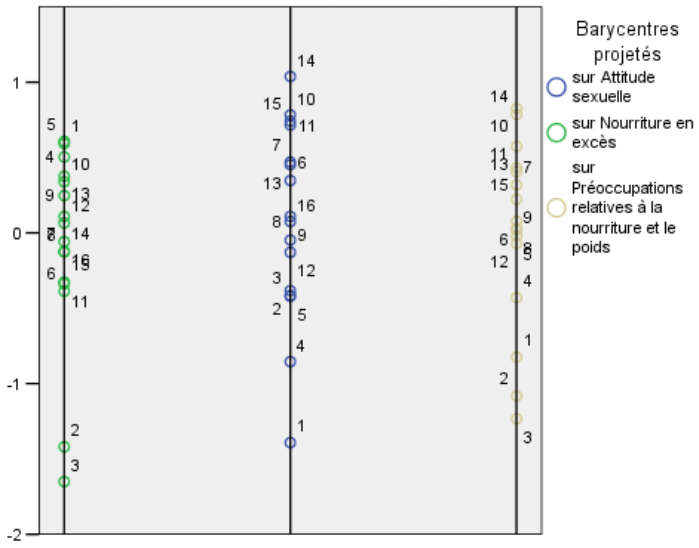
Toutefois, ce diagramme indique aussi que, sur la durée, les maladies ont tendance à se ressembler. En outre, pour tous les groupes, la progression est la plus forte entre les points 1 et 2 dans le temps ; les patients anorexiques présentent un peu plus de progression entre les points 2 et 3, mais les autres groupes affichent peu de progression.

Développement différentiel de variables spécifiques

Une variable de chaque groupe de symptômes identifié par les corrélations entre composantes a été sélectionnée comme étant "représentative" du groupe. La frénésie alimentaire a été sélectionnée dans le groupe boulimique, l'attitude sexuelle dans le groupe anorexique/psychosocial et la préoccupation du corps dans le troisième groupe.

Pour que vous puissiez examiner les éventuelles évolutions différentielles des maladies, les projections de *Interaction moment/diagnostic* sur *Frénésie alimentaire*, *Attitude sexuelle* et *Préoccupation nourriture et poids* ont été calculées et représentées dans le schéma ci-après.

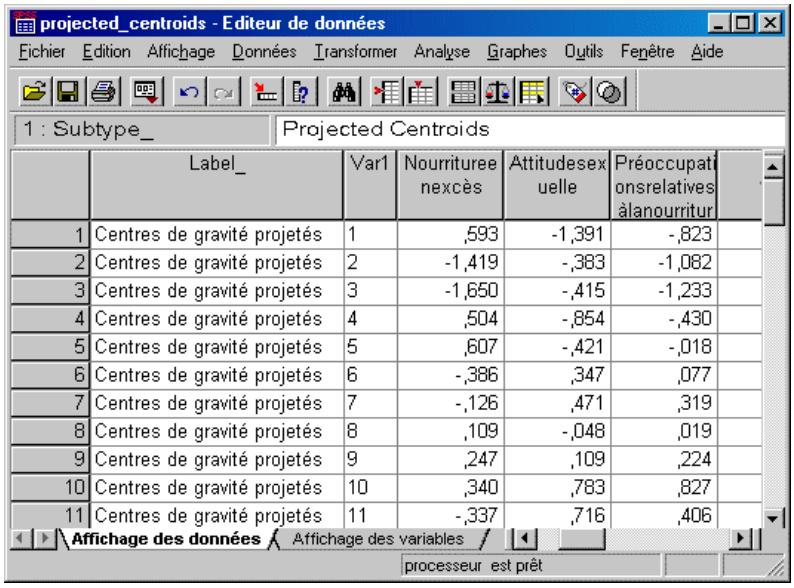
Figure 10-44
Centres de gravité projetés de Interaction moment/diagnostic sur Frénésie alimentaire, Attitude sexuelle et Préoccupation nourriture et poids



Ce diagramme indique qu'au premier point dans le temps, la frénésie alimentaire symptomatique sépare les patients boulimiques (2 et 3) des autres patients (1 et 4), que l'attitude sexuelle sépare les patients anorexiques et atypiques (1 et 4) des autres patients (2 et 3), et que la préoccupation du corps ne sépare pas véritablement les patients. Dans de nombreuses applications, ce diagramme suffirait pour décrire la relation entre les symptômes et le diagnostic mais, en raison du caractère multiple des points dans le temps, l'image perd de sa netteté.

Pour visualiser ces projections sur la durée, vous devez être en mesure de représenter le contenu du tableau des centres de gravité projetés. Cette opération est possible grâce à la requête OMS ayant enregistré ces informations dans le fichier *projected_centroids.sav*.

Figure 10-45
Projected_centroids.sav

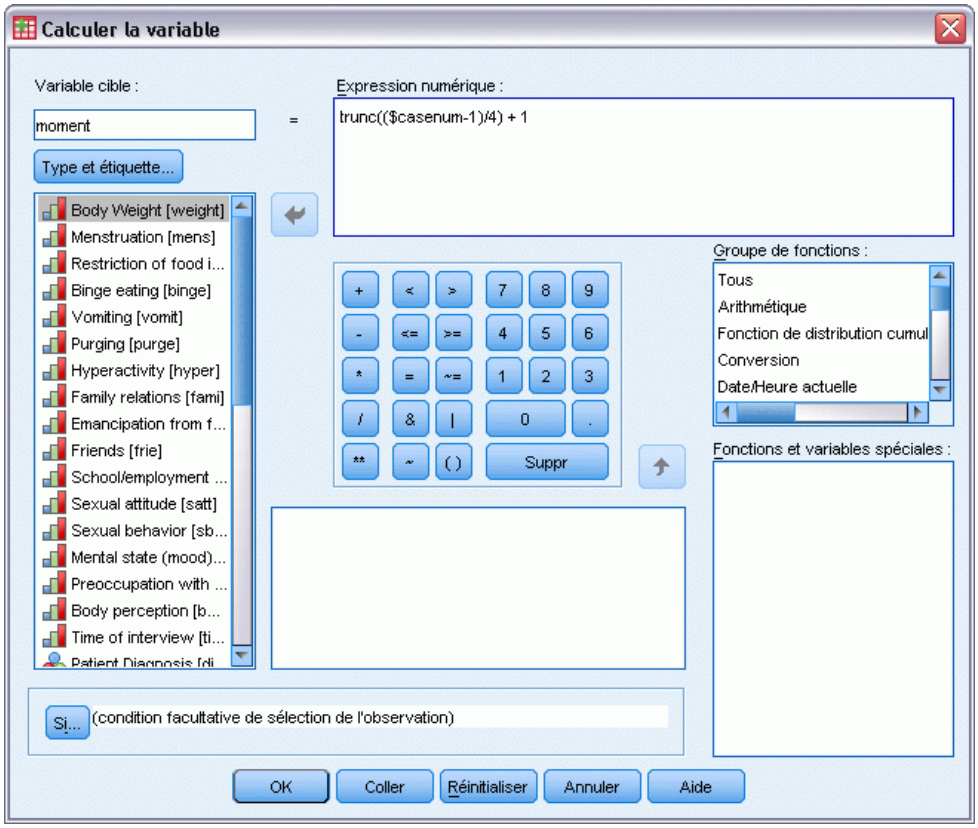


	Label_	Var1	Nourriturees nexcès	Attitudesexuelle	Préoccupationsrelativesàlanourritur
1	Centres de gravité projetés	1	,593	-1,391	-,823
2	Centres de gravité projetés	2	-1,419	-,383	-1,082
3	Centres de gravité projetés	3	-1,650	-,415	-1,233
4	Centres de gravité projetés	4	,504	-,854	-,430
5	Centres de gravité projetés	5	,607	-,421	-,018
6	Centres de gravité projetés	6	-,386	,347	,077
7	Centres de gravité projetés	7	-,126	,471	,319
8	Centres de gravité projetés	8	,109	-,048	,019
9	Centres de gravité projetés	9	,247	,109	,224
10	Centres de gravité projetés	10	,340	,783	,827
11	Centres de gravité projetés	11	-,337	,716	,406

Les variables *FrénésieAlimentaire*, *AttitudeSexuelle* et *PréoccupationAlimentationPoids* contiennent les valeurs des barycentres projetés sur chacun des symptômes d'intérêt. Le numéro d'observation (1 à 16) correspond à l'interaction moment/diagnostic. Vous devrez calculer de nouvelles variables permettant de distinguer les valeurs des moments de celles des diagnostics.

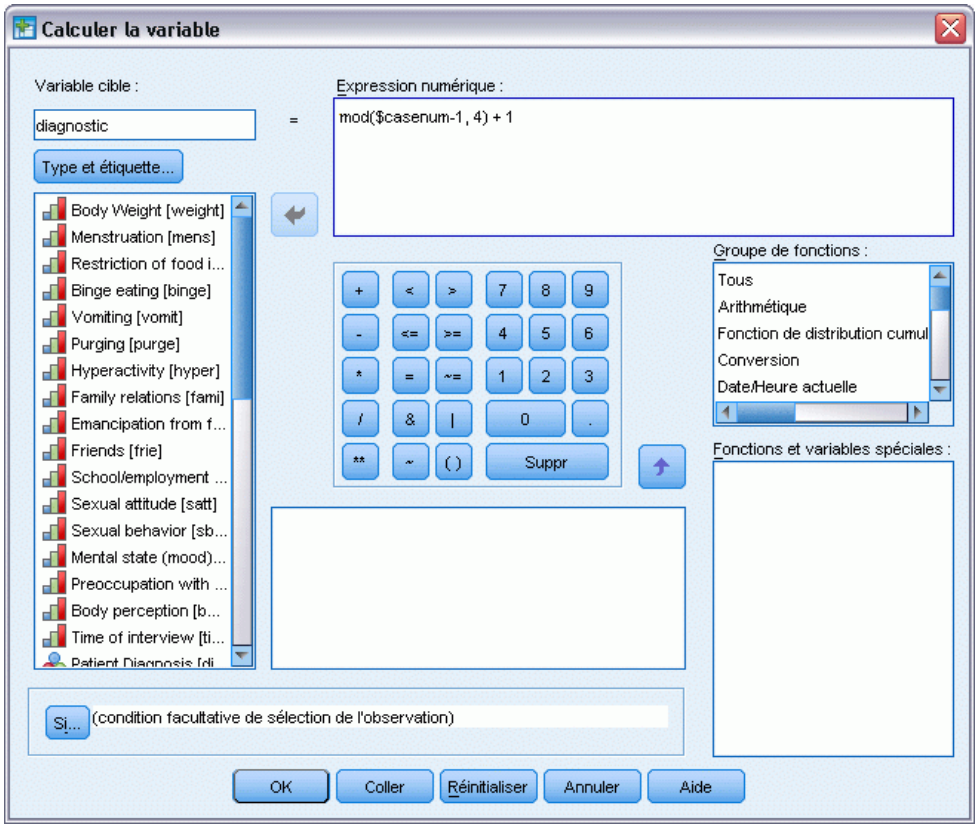
▶ A partir des menus, sélectionnez :
Transformer > Calculer la variable...

Figure 10-46
Boîte de dialogue Calculer la variable



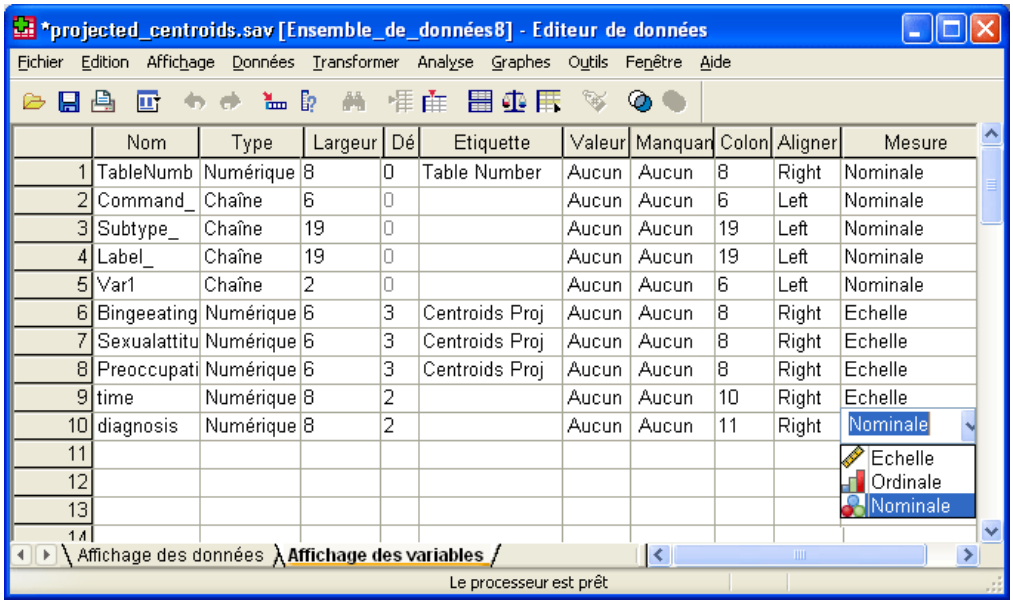
- Tapez *moment* comme variable de destination.
- Tapez trunc(($casenum-1)/4) + 1 comme expression numérique.
- Cliquez sur OK.

Figure 10-47
Boîte de dialogue Calculer la variable



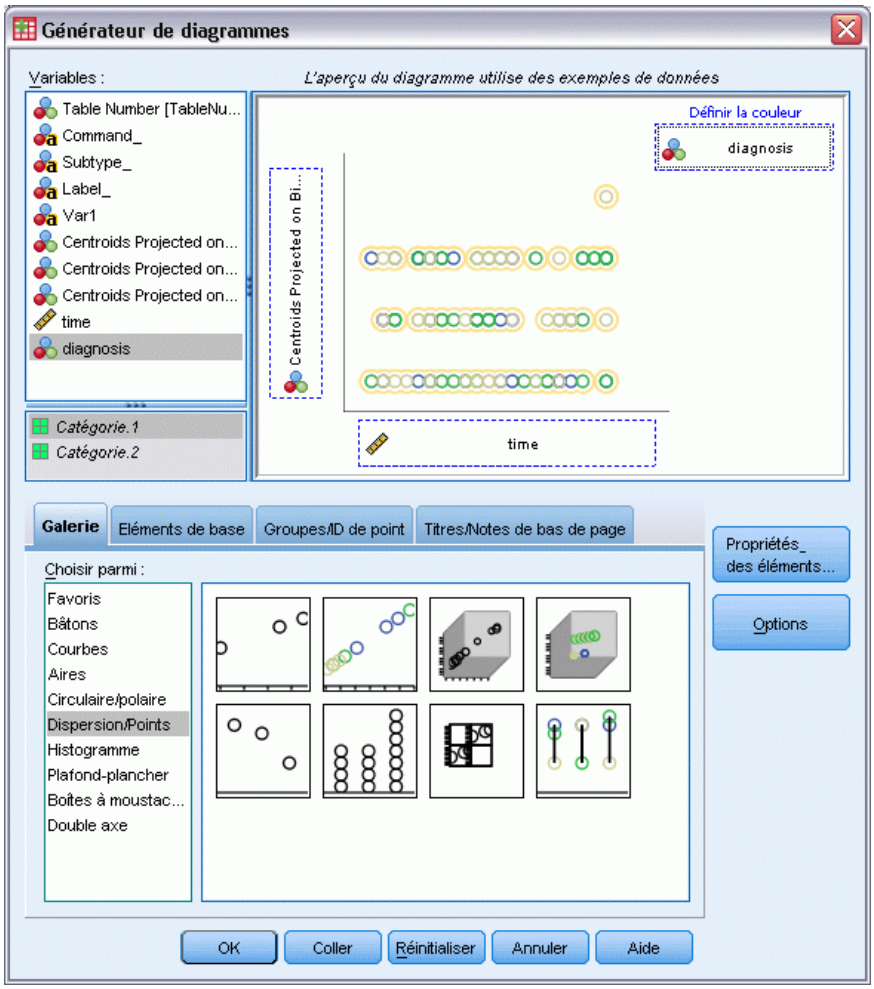
- Rappelez la boîte de dialogue Calculer la variable.
- Tapez *diagnostic* comme variable de destination.
- Tapez mod($casenum-1, 4) + 1 comme expression numérique.
- Cliquez sur OK.

Figure 10-48
Projected_centroids.sav



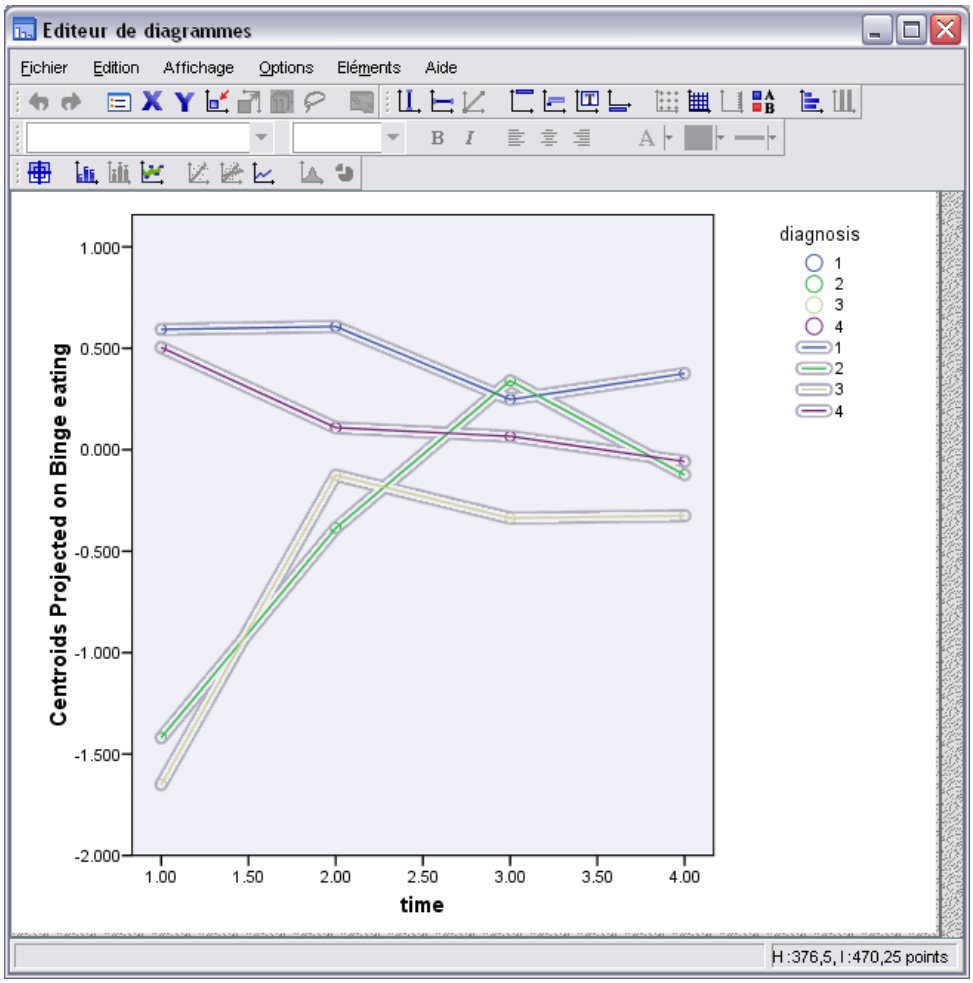
Dans la vue des variables, changez la mesure du *diagnostic* de Positionnement en Nominal.

Figure 10-49
Générateur de diagrammes



- En dernier lieu, pour visualiser dans le temps les barycentres des moments de diagnostic projetés sur la frénésie alimentaire, affichez de nouveau le Générateur de diagrammes, puis cliquez sur le bouton Réinitialiser pour effacer les sélections antérieures.
- Sélectionnez la galerie Dispersion/Points et choisissez Diagramme de dispersion regroupé.
- Sélectionnez l'option *Barycentres projetés sur Frénésie alimentaire* comme variable de l'axe *y* et l'option *moment* comme variable de l'axe *x*.
- Pour la définition des couleurs, choisissez l'option *diagnostic*.
- Cliquez sur OK.

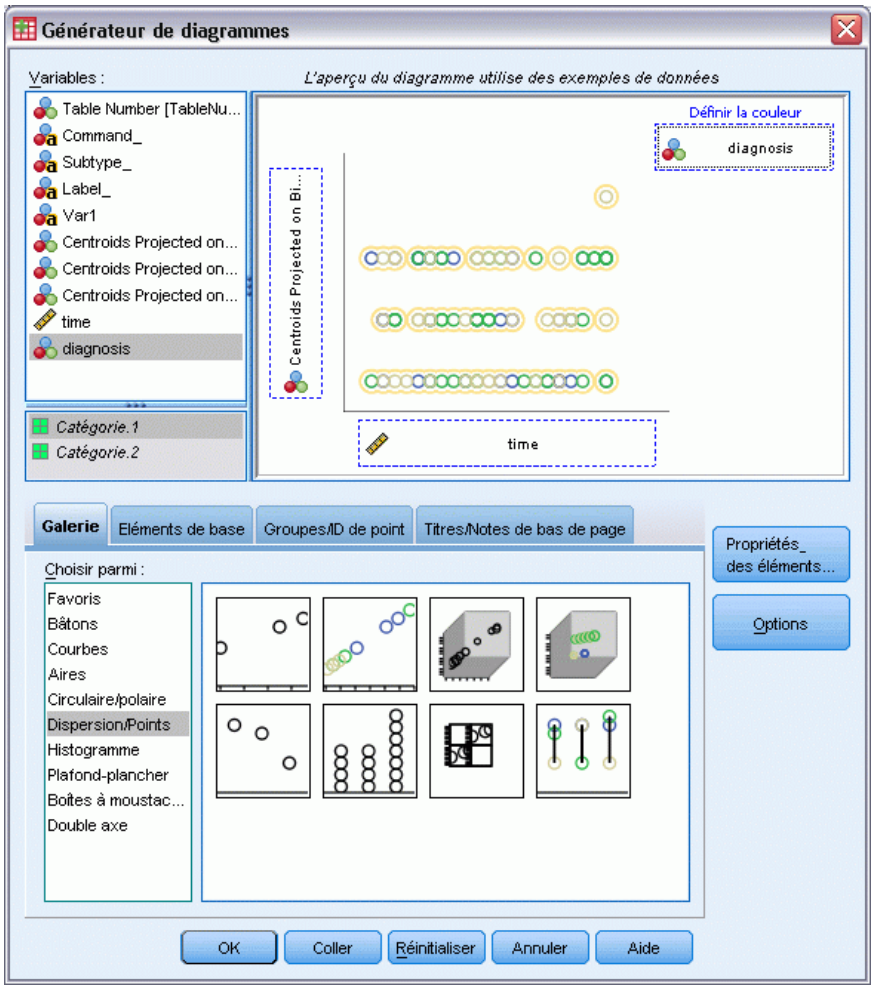
Figure 10-50
Projection dans le temps des barycentres des moments de diagnostic sur la frénésie alimentaire



- Ensuite, pour relier les points, double-cliquez sur le graphique, puis cliquez sur l'onglet Ajouter une courbe d'interpolation dans l'éditeur de diagrammes.
- Fermez l'éditeur de diagrammes.

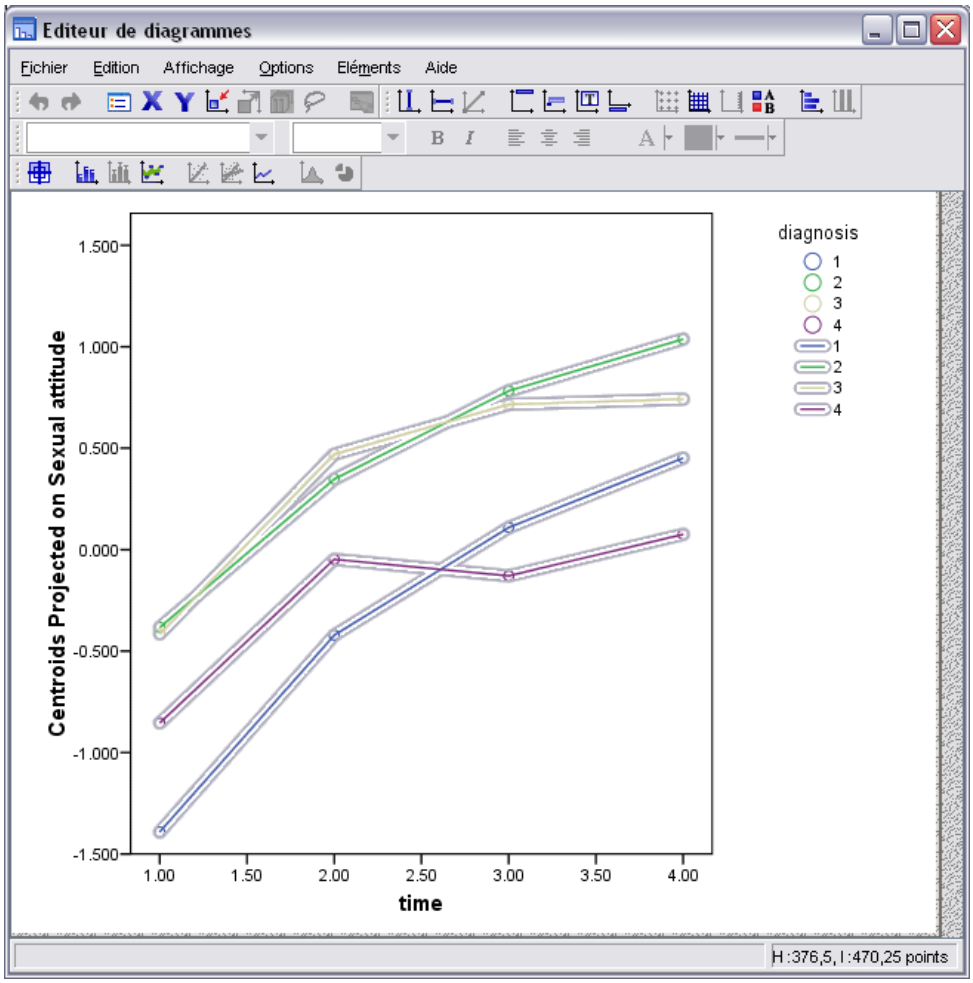
Concernant la frénésie alimentaire, il est manifeste que les groupes anorexiques présentent des valeurs initiales différentes de celles des groupes boulimiques. Cette différence s'estompe au fil du temps, car les groupes anorexiques évoluent très peu tandis que les groupes boulimiques affichent une progression.

Figure 10-51
Générateur de diagrammes



- Affichez de nouveau le Générateur de diagrammes.
- Désélectionnez l'option *Barycentres projetés sur Frénésie alimentaire* comme variable de l'axe *y* et sélectionnez l'option *Barycentres projetés sur Attitude sexuelle* à la place.
- Cliquez sur OK.

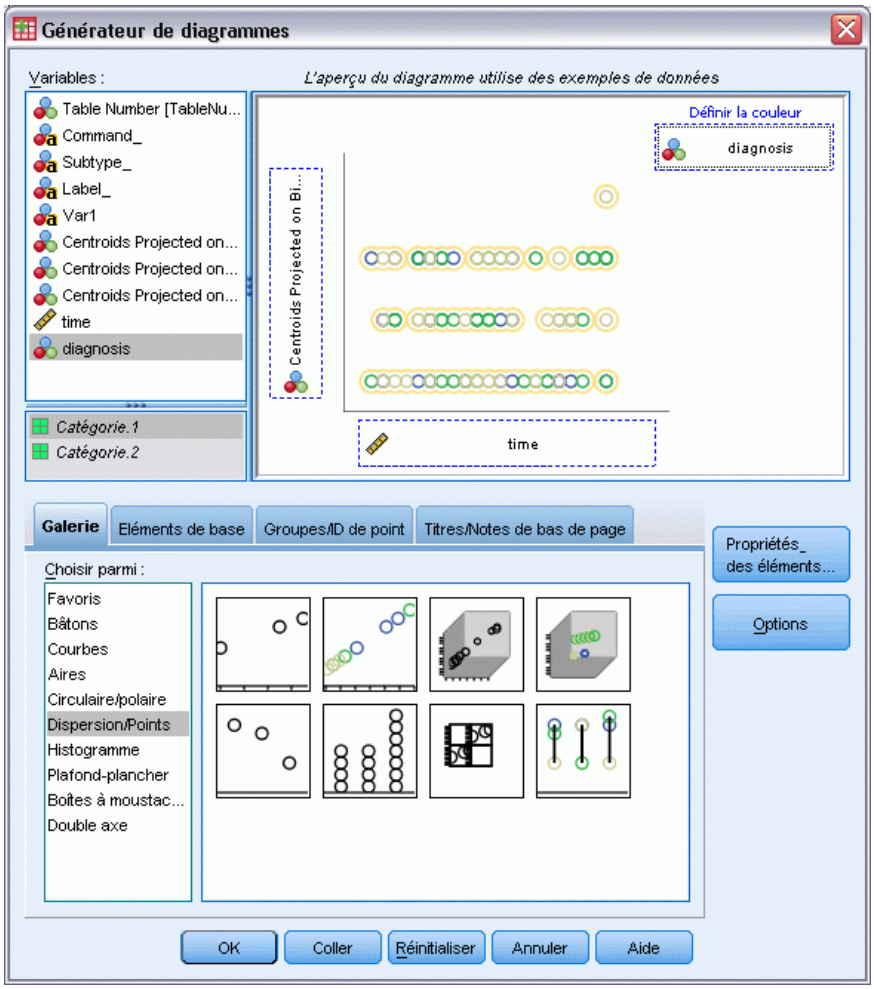
Figure 10-52
Projection dans le temps des barycentres des moments de diagnostic sur l'attitude sexuelle



- Ensuite, pour relier les points, double-cliquez sur le graphique, puis cliquez sur l'onglet Ajouter une courbe d'interpolation dans l'éditeur de diagrammes.
- Fermez l'éditeur de diagrammes.

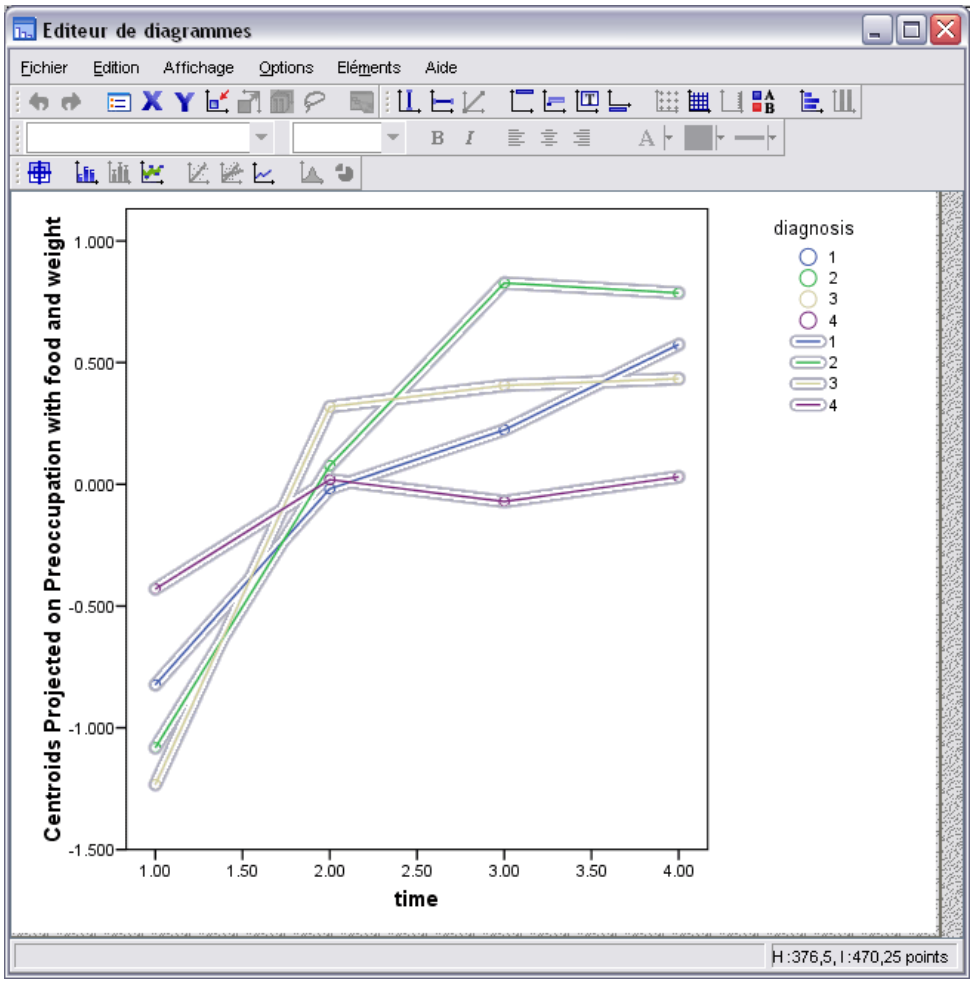
En ce qui concerne l'attitude sexuelle, les quatre trajectoires sont plus ou moins parallèles sur la durée et tous les groupes présentent une progression. Toutefois, les scores des groupes boulimiques sont plus élevés (meilleurs) que ceux du groupe anorexique.

Figure 10-53
Générateur de diagrammes



- Affichez de nouveau le Générateur de diagrammes.
- Désélectionnez l'option *Barycentres projetés sur Attitude sexuelle* comme variable de l'axe *y* et sélectionnez l'option *Barycentres projetés sur Préoccupation nourriture et poids* à la place.
- Cliquez sur OK.

Figure 10-54
Projection dans le temps des barycentres des moments de diagnostic sur la préoccupation du corps



- Ensuite, pour relier les points, double-cliquez sur le graphique, puis cliquez sur l'onglet Ajouter une courbe d'interpolation dans l'éditeur de diagrammes.
- Fermez l'éditeur de diagrammes.

La préoccupation du corps est une variable qui représente les symptômes fondamentaux, partagés par les quatre groupes. En dehors des patients atteints de troubles atypiques du comportement alimentaire, le groupe anorexique et les deux groupes boulimiques présentent des niveaux très similaires au début comme à la fin.

Lectures recommandées

Pour plus d'informations sur l'analyse des composantes principales qualitatives, reportez-vous aux documents suivants :

De Haas, M., J. A. Algera, H. F. J. M. Van Tuijl, et J. J. Meulman. 2000. Macro and micro goal setting: In search of coherence. *Applied Psychology*, 49, .

De Leeuw, J. 1982. Nonlinear principal components analysis. Dans : *COMPSTAT Proceedings in Computational Statistics,* Vienne: Physica Verlag.

Eckart, C., et G. Young. 1936. The approximation of one matrix by another one of lower rank. *Psychometrika*, 1, .

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, 58, .

Gifi, A. 1985. *PRINCALS. Research Report UG-85-02*. Leiden: Department of Data Theory, University of Leiden.

Gower, J. C., et J. J. Meulman. 1993. The treatment of categorical information in physical anthropology. *International Journal of Anthropology*, 8, .

Heiser, W. J., et J. J. Meulman. 1994. Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. Dans : *Correspondence Analysis in the Social Sciences: Recent Developments and Applications,* M. Greenacre, et J. Blasius, éds. New York: Academic Press.

Kruskal, J. B. 1978. Factor analysis and principal components analysis: Bilinear methods. Dans : *International Encyclopedia of Statistics,* W. H. Kruskal, et J. M. Tanur, éds. New York: The Free Press.

Kruskal, J. B., et R. N. Shepard. 1974. A nonmetric variety of linear factor analysis. *Psychometrika*, 39, .

Meulman, J. J. 1993. Principal coordinates analysis with optimal transformations of the variables: Minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology*, 46, .

Meulman, J. J., et P. Verboon. 1993. Points of view analysis revisited: Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika*, 58, .

Meulman, J. J., A. J. Van der Kooij, et A. Babinec. 2000. New features of categorical principal components analysis for complicated data sets, including data mining. Dans : *Classification, Automation and New Media,* W. Gaul, et G. Ritter, éds. Berlin: Springer-Verlag.

Meulman, J. J., A. J. Van der Kooij, et W. J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. Dans : *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, éd. Thousand Oaks, Californie: Sage Publications, Inc..

Theunissen, N. C. M., J. J. Meulman, A. L. Den Ouden, H. M. Koopman, G. H. Verrips, S. P. Verloove-Vanhorick, et J. M. Wit. 2003. Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology*, 4, .

Tucker, L. R. 1960. Intra-individual and inter-individual multidimensionality. Dans : *Psychological Scaling: Theory & Applications,* H. Gulliksen, et S. Messick, éds. New York: John Wiley and Sons.

Vlek, C., et P. J. Stallen. 1981. Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance*, 28, .

Wagenaar, W. A. 1988. *Paradoxes of gambling behaviour*. Londres: Lawrence Erlbaum Associates, Inc.

Young, F. W., Y. Takane, et J. De Leeuw. 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, .

Zeijl, E., Y. te Poel, M. du Bois-Reymond, J. Ravesloot, et J. J. Meulman. 2000. The role of parents and peers in the leisure activities of young adolescents. *Journal of Leisure Research*, 32, .

Chapitre
11

Analyse de corrélation canonique non linéaire

L'analyse de corrélation canonique non linéaire a pour but de déterminer le degré de ressemblance entre plusieurs groupes de variables. Comme dans l'analyse de corrélation canonique linéaire, l'objectif est d'évaluer autant que possible la variance dans les relations entre les groupes dans un espace comportant peu de dimensions. En revanche, contrairement à l'analyse de corrélation canonique linéaire, l'analyse de corrélation canonique non linéaire ne suppose pas qu'un niveau d'intervalle de mesure soit défini ou que les relations soient linéaires. Autre différence importante : l'analyse de corrélation canonique non linéaire établit la similarité qui existe entre les groupes en comparant simultanément des combinaisons linéaires des variables de chaque groupe avec un groupe inconnu, les coordonnées des objets.

Exemple \: Analyse des résultats d'enquête

L'exemple utilisé dans ce chapitre est tiré d'une enquête (Verdegaal, 1985). Les réponses de 15 sujets à 8 variables ont été enregistrées. Les variables, les étiquettes de variable et les étiquettes de valeur (modalités) de l'ensemble de données sont indiquées dans le tableau suivant.

Table 11-1
Données de l'enquête

Nom de variable	Etiquette de variable	Etiquette de valeur
âge	*Age en années*	20–25, 26–30, 31–35, 36–40, 41–45, 46–50, 51–55, 56–60, 61–65, 66–70
situation familiale	*Situation familiale*	Célibataire, Marié, Autre
animal domestique	*Animaux domestiques possédés*	Aucun, Chat(s), Chien(s), Autre que chien ou chat, Plusieurs animaux domestiques
presse	*Journal lu le plus souvent*	Aucun, Telegraaf, Volkskrant, NRC, Autre
musique	*Musique préférée*	Classique, New wave, Pop, Variété, N'aime pas la musique
habitat	*Voisinage préféré*	Ville, Village, Campagne
maths	*Résultat du test mathématique*	0–5, 6–10, 11–15
Langage	*Résultat du test linguistique*	0–5, 6–10, 11–15, 16–20

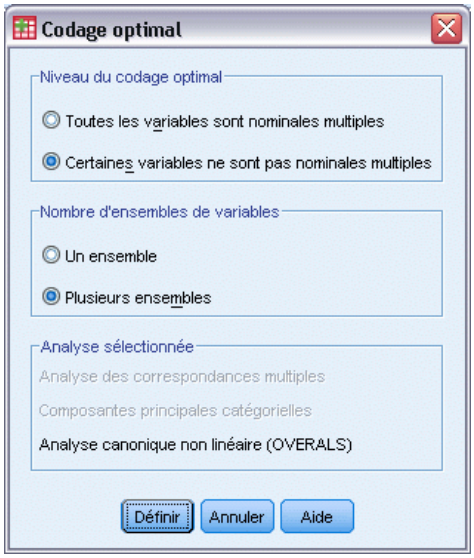
Cet ensemble de données est disponible dans le fichier *verd1985.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*. Les variables qui nous intéressent ici sont les six premières ; elles sont réparties en trois groupes. Le groupe 1 comprend l'*âge* et la *situation familiale*, le groupe 2 les *animaux domestiques* et la

presse, et le groupe 3 la *musique* et l'*habitat*. A la variable *animal domestique* est appliqué un codage nominal multiple et à *âge*, un codage ordinal ; toutes les autres variables ont un codage nominal simple. Pour cette analyse, il est nécessaire d'utiliser une configuration initiale aléatoire. Par défaut, la configuration initiale est numérique. Toutefois, lorsque certaines variables sont traitées comme des valeurs nominales simples sans possibilité de tri, il est préférable d'utiliser une configuration initiale aléatoire. C'est le cas de la plupart des variables dans cette enquête.

Examen des données

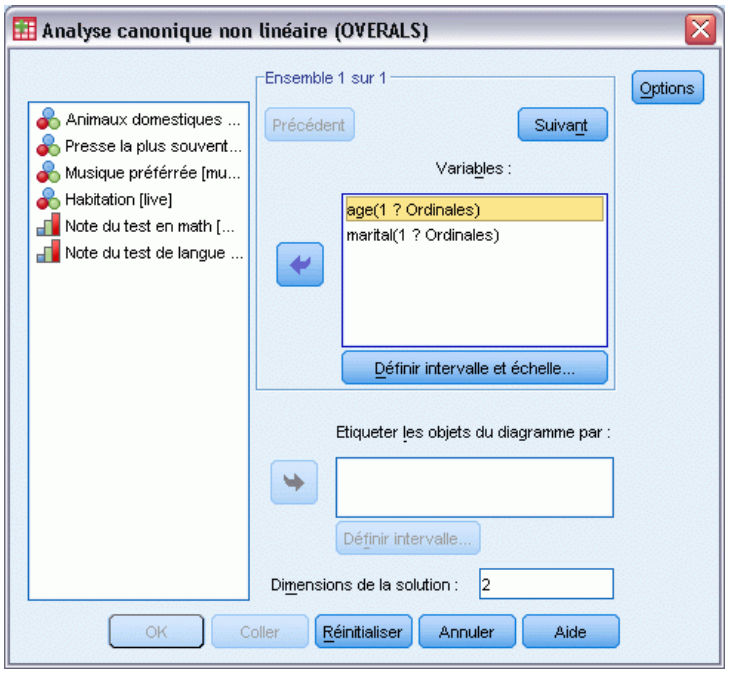
- Pour obtenir une analyse de corrélation canonique non linéaire pour cet ensemble de données, sélectionnez les options suivantes dans les menus :
  Analyse > Réduction des dimensions > Codage optimal

Figure 11-1
Boîte de dialogue Niveau du codage optimal



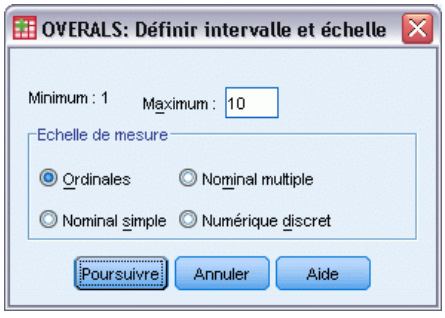
- Sélectionnez l'option Certaines variables non nominales multiples dans le groupe Niveau du codage optimal.
- Sélectionnez Plusieurs dans le groupe Nombre de groupes de variables.
- Cliquez sur Définir.

Figure 11-2
Boîte de dialogue Analyse de corrélation canonique non linéaire



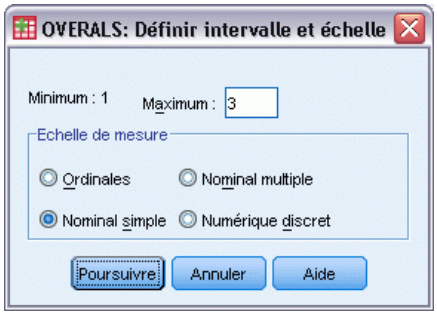
- Sélectionnez *Age en années* et *Situation familiale* comme variables du premier groupe.
- Sélectionnez *âge* et cliquez sur Définir intervalle et échelle.

Figure 11-3
Boîte de dialogue Définir intervalle et échelle



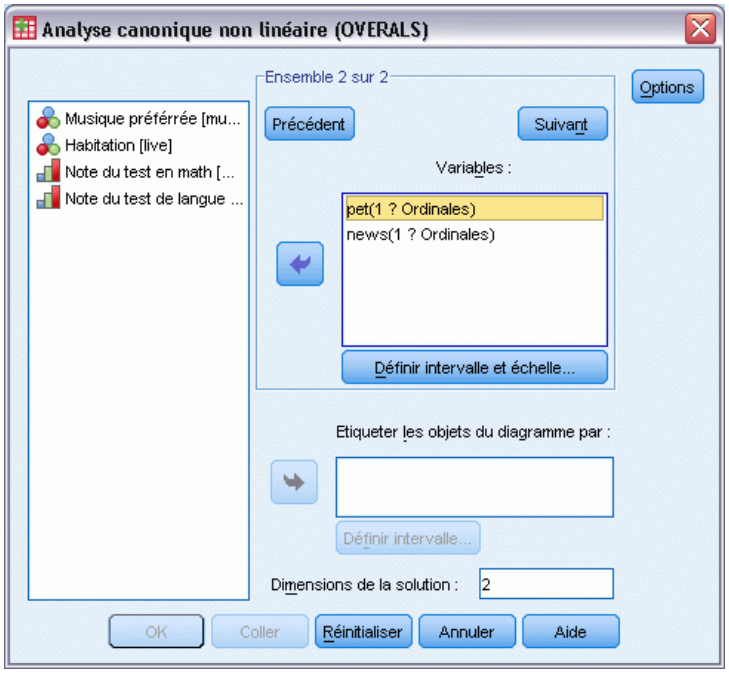
- Entrez 10 comme valeur maximale pour cette variable.
- Cliquez sur Poursuivre.
- Sélectionnez *situatio*, puis cliquez sur Définir intervalle et échelle dans la boîte de dialogue Analyse de corrélation canonique non linéaire.

Figure 11-4
Boîte de dialogue Définir intervalle et échelle



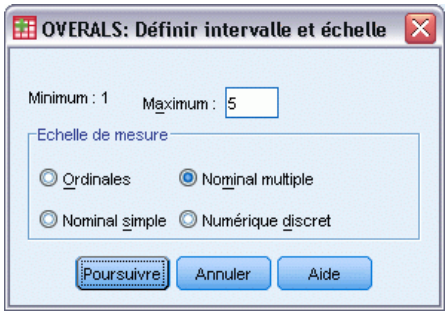
- Entrez 3 comme valeur maximale pour cette variable.
- Sélectionnez l'option Nominal simple comme échelle de mesure.
- Cliquez sur Poursuivre.
- Dans la boîte de dialogue Analyse de corrélation canonique non linéaire, cliquez sur Suivant pour définir le groupe de variables suivant.

Figure 11-5
Boîte de dialogue Analyse de corrélation canonique non linéaire



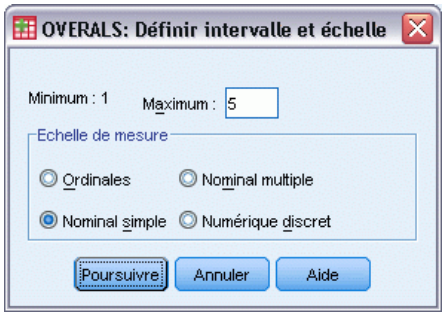
- Sélectionnez *Animaux domestiques possédés* et *Journal lu le plus souvent* comme variables du deuxième groupe.
- Sélectionnez *animal domestique* et cliquez sur Définir intervalle et échelle.

Figure 11-6
Boîte de dialogue Définir intervalle et échelle



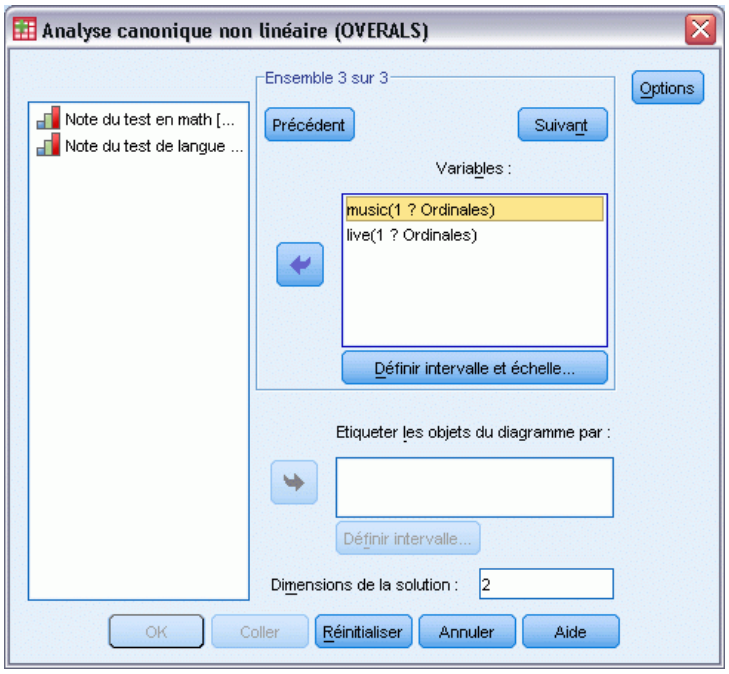
- Entrez 5 comme valeur maximale pour cette variable.
- Sélectionnez Variables nominales multiples comme échelle de mesure.
- Cliquez sur Poursuivre.
- Dans la boîte de dialogue Analyse de corrélation canonique non linéaire, sélectionnez *informations*, puis cliquez sur Définir intervalle et échelle.

Figure 11-7
Boîte de dialogue Définir intervalle et échelle



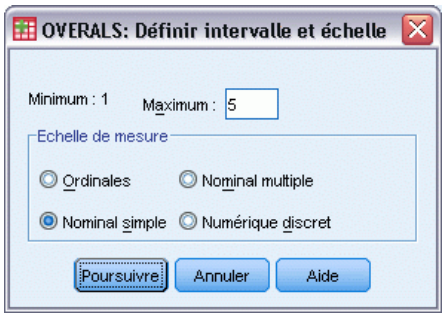
- Entrez 5 comme valeur maximale pour cette variable.
- Sélectionnez l'option Nominal simple comme échelle de mesure.
- Cliquez sur Poursuivre.
- Dans la boîte de dialogue Analyse de corrélation canonique non linéaire, cliquez sur Suivant pour définir le dernier groupe de variables.

Figure 11-8
Boîte de dialogue Analyse de corrélation canonique non linéaire



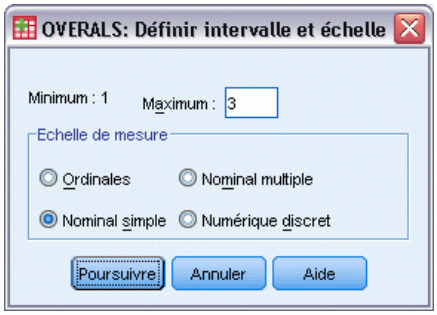
- Sélectionnez *Musique favorite* et *Préférence de voisinage* comme variables du troisième groupe.
- Sélectionnez *musique*, puis cliquez sur Définir intervalle et échelle.

Figure 11-9
Boîte de dialogue Définir intervalle et échelle



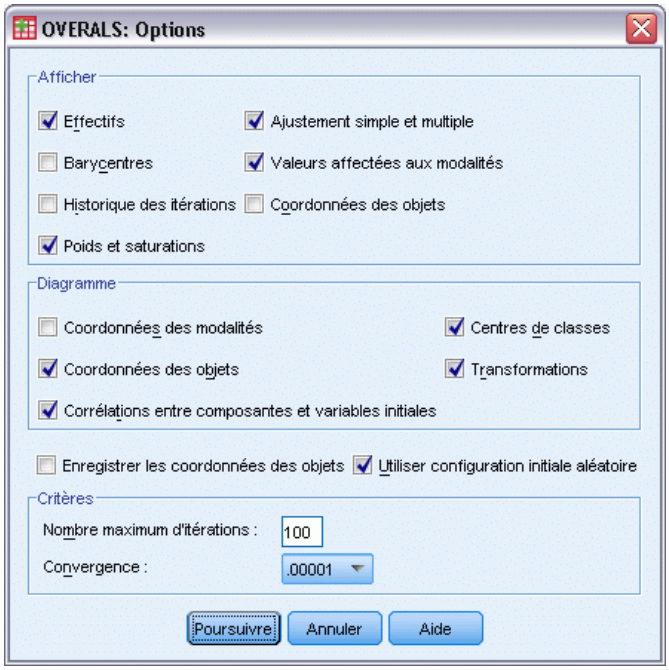
- Entrez 5 comme valeur maximale pour cette variable.
- Sélectionnez l'option Nominal simple comme échelle de mesure.
- Cliquez sur Poursuivre.
- Dans la boîte de dialogue Analyse de corrélation canonique non linéaire, sélectionnez *habitat*, puis cliquez sur Définir intervalle et échelle.

Figure 11-10
Boîte de dialogue Définir intervalle et échelle



- Entrez 3 comme valeur maximale pour cette variable.
- Sélectionnez l'option Nominal simple comme échelle de mesure.
- Cliquez sur Poursuivre.
- Dans la boîte de dialogue Analyse canonique non linéaire, cliquez sur Options.

Figure 11-11
Options



- Désélectionnez la case Barycentres et sélectionnez l'option Poids et saturations dans le groupe Affichage.
- Sélectionnez les options Centres de classes et Transformations dans le groupe Diagramme.
- Sélectionnez l'option Utiliser configuration initiale aléatoire.
- Cliquez sur Poursuivre.

▶ Dans la boîte de dialogue Analyse canonique non linéaire, cliquez sur OK.

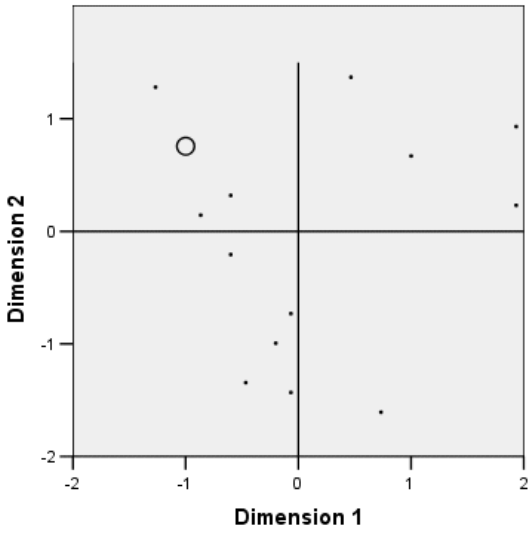
Après la liste des variables incluant leurs niveaux de codage optimal, l'analyse de corrélation canonique nominale avec codage optimal génère un tableau illustrant les effectifs des objets dans les modalités. Ce tableau s'avère essentiel en cas de données manquantes ; en effet, les modalités quasiment vides ont plus de chance d'influencer la solution. Cet exemple ne comporte aucune donnée manquante.

Une autre vérification préliminaire consiste à étudier le diagramme de coordonnées des objets pour les valeurs éloignées. Les valeurs éloignées ont des quantifications si différentes des autres objets qu'elles se situent à la limite du diagramme, dominant ainsi une ou plusieurs dimensions.

Vous pouvez gérer les éventuelles valeurs éloignées de deux manières. Vous pouvez simplement les retirer des données et exécuter à nouveau l'analyse de corrélation canonique non linéaire. Ou vous pouvez essayer de recoder les réponses extrêmes des objets éloignés en fusionnant certaines modalités.

Comme l'illustre le diagramme de coordonnées des objets, les données de l'enquête ne comportent aucune valeur éloignée.

Figure 11-12
Coordonnées des objets



Observations pondérées par le nombre d'objets.

Similarités entre les groupes

Il existe plusieurs manières de mesurer l'association entre les groupes dans une analyse de corrélation canonique non linéaire (chacune d'elles étant exposée en détail dans un tableau ou un groupe de tableaux distinct).

Récapitulatif de l'analyse

Les valeurs d'ajustement et de perte vous renseignent sur l'adéquation entre l'analyse de corrélation canonique non linéaire et les données quantifiées de manière optimale, en ce qui concerne l'association entre les groupes. Le tableau récapitulatif de l'analyse affiche les valeurs d'ajustement, les valeurs de perte et les valeurs propres de cet exemple d'enquête.

Figure 11-13
Récapitulatif de l'analyse

		Dimension		
		1	2	Somme
Perte	Groupe 1	,240	,183	,423
	Groupe 2	,184	,408	,593
	Groupe 3	,171	,205	,376
	Moyenne	,199	,265	,464
Valeur propre		,801	,735	
Ajustement LDN				1,536

La perte est répartie entre les dimensions et les groupes. Pour chaque dimension et groupe, la perte représente la proportion de variation des coordonnées d'objet qui ne peut pas être représentée par la combinaison pondérée des variables du groupe. La perte moyenne est intitulée « moyenne ». Dans cet exemple, la perte moyenne sur tous les groupes est de 0,464. La perte est plus importante pour la deuxième dimension que pour la première.

La valeur propre de chaque dimension est égale à 1 moins la perte moyenne de la dimension ; la valeur propre indique la quantité de la relation révélée par chaque dimension. Les valeurs propres s'ajoutent à l'ajustement total. Pour les données relatives à Verdegaal, 0,801/1,536 = 52 % de l'ajustement réel sont attribuables à la première dimension.

La valeur d'ajustement maximale est égale au nombre de dimensions. Si elle est obtenue, elle indique une relation parfaite. La valeur de perte moyenne sur tous les groupes et dimensions indique l'écart entre la valeur d'ajustement maximale et l'ajustement réel. La valeur d'ajustement plus la valeur de perte moyenne sont égales au nombre de dimensions. Une similarité parfaite est très rare et repose généralement sur des aspects insignifiants des données.

On trouve également parmi les outils statistiques très utilisés, avec deux groupes de variables, la corrélation canonique. La corrélation canonique étant liée à la valeur propre et ne fournissant par conséquent aucune information supplémentaire, elle n'est pas incluse dans les résultats de l'analyse de corrélation canonique non linéaire. Dans le cas de deux groupes de variables, on obtient la corrélation canonique par dimension à l'aide de la formule suivante :

$$\rho_d = 2 \times E_d - 1$$

d correspondant au nombre de dimensions et *E* à la valeur propre.

Il est possible d'étendre la corrélation canonique à plus de deux groupes ; pour ce faire, utilisez la formule suivante :

$$\rho_d = ((K \times E_d) - 1)/(K - 1)$$

d correspondant au nombre de dimensions, *K* au nombre de groupes et *E* à la valeur propre. Dans notre exemple,

$$\rho_1 = ((3 \times 0.801) - 1)/2 = 0.702$$

et

$$\rho_2 = ((3 \times 0.735) - 1)/2 = 0.603$$

Poids et saturations :

Il existe également une autre mesure d'association : la corrélation multiple entre les combinaisons linéaires de chaque groupe et les coordonnées des objets. Si un groupe ne contient aucune variable nominale multiple, vous pouvez calculer cette mesure en multipliant les pondérations et corrélations entre composantes de chaque variable du groupe, en ajoutant ces produits et en calculant la racine carrée de la somme obtenue.

Figure 11-14
Poids

Groupe		Dimension 1	Dimension 2
1	Age in years	,680	,789
	Marital status	,296	-1,016
2	Newspaper read most often	-,845	-,361
3	Music preferred	,631	-,749
	Neighborhood preference	-,484	-,780

a. Variables nominales multiples non incluses dans le tableau.

Figure 11-15
Corrélations entre composantes et variables initiales

Groupe				Dimension 1	Dimension 2
1	Age in years [a,b]			,834	,259
	Marital status [c,b]			,651	-,604
2	Pets owned [d,e]	Dimension	1	,397	-,431
			2	-,277	,680
	Newspaper read most often [c,b]			-,667	-,391
3	Music preferred [c,b]			,786	-,500
	Neighborhood preference [c,b]			-,687	-,540

a. Niveau de codage optimal : Ordinal

b. Projections des variables quantifiées uniques dans l'espace des objets

c. Niveau de codage optimal : Nominal simple

d. Niveau de codage optimal : Nominal multiple

e. Projections des variables quantifiées multiples dans l'espace des objets

Ces chiffres donnent les pondérations et les corrélations entre composantes des variables de cet exemple. La corrélation multiple (*R*) est comme suit pour la première somme pondérée des variables codées de façon optimale (*Age en années* et *Situation familiale*) avec la première dimension des coordonnées des objets :

$$\begin{aligned} R &= \sqrt{(0.701 \times 0.841 + (-0.273 \times -0.631))} \\ &= \sqrt{(0.5895 + 0.1723)} \\ &= 0.873 \end{aligned}$$

Pour chaque dimension, 1 –perte = R^2. Par exemple, dans le tableau récapitulatif de l'analyse, 1 – 0,238 = 0,762, soit 0,873 au carré (à une erreur d'arrondi près). Par conséquent, les valeurs de perte faibles indiquent de fortes corrélations multiples entre les sommes pondérées des variables codées de façon optimale et les dimensions. Les pondérations ne sont pas uniques pour les variables nominales multiples. Pour les variables nominales multiples, utilisez la formule 1 – perte par groupe.

Partitionnement des ajustements et des pertes

La perte de chaque groupe est répartie de différentes manières par l'analyse de corrélation canonique non linéaire. Le tableau d'ajustement présente les tableaux d'ajustement multiple, d'ajustement unique et de perte unique produits par l'analyse de corrélation canonique non linéaire pour cet exemple d'enquête. Remarque : l'ajustement multiple moins l'ajustement unique est égal à la perte unique.

Figure 11-16
Partitionnement des ajustements et des pertes

		Ajustement multiple			Ajustement unique			Perte unique		
		Dimension			Dimension			Dimension		
Groupe		1	2	Somme	1	2	Somme	1	2	Somme
1	Age in years[a]	,494	,676	1,170	,462	,622	1,085	,032	,054	,085
	Marital status[b]	,089	1,033	1,122	,088	1,033	1,120	,001	,000	,001
2	Pets owned[c]	,402	,439	,841						
	Newspaper read most often[b]	,724	,187	,911	,714	,130	,844	,010	,057	,067
3	Music preferred[b]	,421	,577	,998	,398	,561	,960	,022	,016	,039
	Neighborhood preference[b]	,234	,609	,843	,234	,608	,843	,000	,000	,000

a. Niveau de codage optimal : Ordinal

b. Niveau de codage optimal : Nominal simple

c. Niveau de codage optimal : Nominal multiple

La perte unique indique la perte résultant de la limitation des variables à un seul groupe de quantifications (c'est-à-dire, nominal simple, ordinal ou nominal). Si la perte unique est élevée, il est préférable de traiter les variables comme des variables nominales multiples. Dans cet exemple, toutefois, les ajustements unique et multiple sont presque égaux, ce qui signifie que les coordonnées multiples sont presque toutes situées sur une ligne droite, dans la direction indiquée par les pondérations.

L'ajustement multiple est égal à la variance des coordonnées de modalité multiples pour chaque variable. Ces mesures sont analogues aux mesures de discrimination trouvées dans l'analyse d'homogénéité. Vous pouvez consulter le tableau d'ajustement multiple pour connaître

les variables les plus discriminantes. Par exemple, reportez-vous au tableau d'ajustement multiple pour examiner les variables *Situation familiale* et *Journal lu le plus souvent*. Les valeurs d'ajustement, additionnées sur les deux dimensions, sont 1,122 pour *Situation familiale* et 0,911 pour *Journal lu le plus souvent*. Ces informations nous indiquent que la situation familiale d'une personne a un plus grand pouvoir discriminant que le journal auquel elle est abonnée.

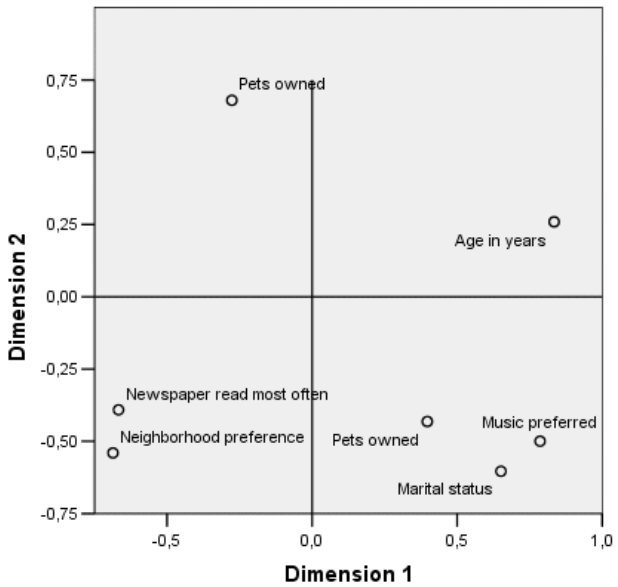
L'ajustement unique correspond à la pondération au carré de chaque variable ; il est égal à la variance des coordonnées de modalité simples. Ainsi, les pondérations sont égales aux écarts-types des coordonnées de modalité simples. En étudiant la manière dont l'ajustement unique est réparti entre les dimensions, on constate que la variable *Journal lu le plus souvent* est discriminante principalement sur la première dimension et on constate que la variable *Situation familiale* est discriminante essentiellement sur la deuxième dimension. Autrement dit, les différentes modalités de *Journal lu le plus souvent* sont plus éloignées dans la première dimension que dans la deuxième, contrairement à celles de *Situation familiale*. En revanche, la variable *Age en années* a un pouvoir discriminant à la fois dans la première et la deuxième dimension ; la dispersion des modalités est donc identique sur les deux dimensions.

Saturations

Le schéma ci-dessous représente le diagramme de corrélations entre composantes des données de l'enquête. Lorsqu'il ne manque aucune donnée, les corrélations entre composantes sont équivalentes aux corrélations de Pearson entre les variables quantifiées et les coordonnées des objets.

La distance depuis l'origine de chaque point de variable est proche de l'importance de cette variable. Les variables canoniques ne sont pas reportées ; elles peuvent toutefois être représentées par des lignes horizontales et verticales tracées via l'origine.

Figure 11-17
Corrélations entre composantes et variables initiales



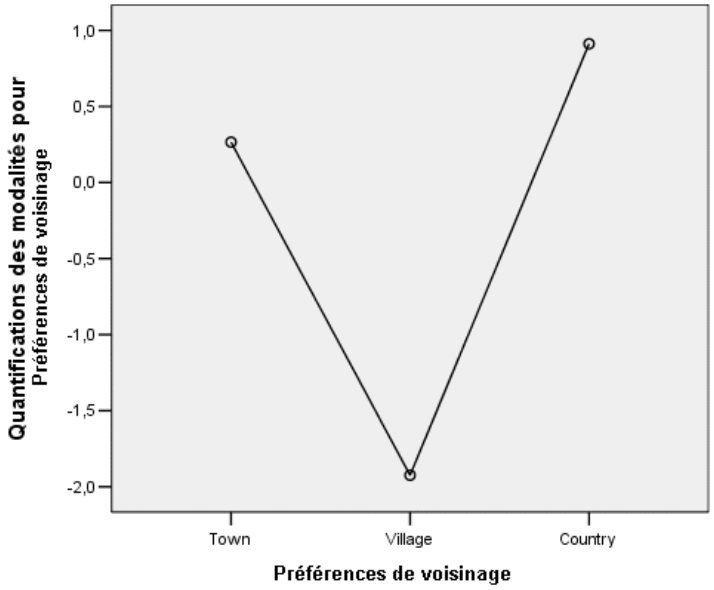
Les relations entre les variables sont apparentes. Deux directions ne coïncident pas avec les axes horizontal et vertical. L'une des directions est déterminée par les variables *Age en années*, *Journal lu le plus souvent* et *Préférence de voisinage*. L'autre est définie par les variables *Situation familiale*, *Musique favorite* et *Animaux domestiques possédés*. La variable *Animaux domestiques possédés* est une variable nominale multiple et est donc représentée par deux points. Chaque quantification est interprétée comme une variable unique.

Diagrammes de transformation

Les différents niveaux auxquels chaque variable peut être codée imposent des restrictions dans les quantifications. Les diagrammes de transformation illustrent la relation entre les quantifications et les modalités d'origine résultant du niveau de codage optimal sélectionné.

Le diagramme de transformation de la variable *Préférence de voisinage*, qui a été traitée comme variable nominale, affiche une forme en U, dans laquelle la modalité centrale reçoit la plus petite quantification et les modalités extrêmes, des valeurs identiques. Cette configuration indique une relation quadratique entre la variable d'origine et la variable transformée. L'utilisation d'un autre niveau de codage optimal n'est pas recommandée pour la variable *Préférence de voisinage*.

Figure 11-18
Diagramme de transformation de la variable Préférence de voisinage (nominale)



Les quantifications de *Journal lu le plus souvent*, en revanche, marquent une croissance entre les trois modalités dotées d'observations. La première modalité reçoit la plus faible quantification, la deuxième une valeur plus élevée et la troisième reçoit la valeur la plus élevée. Bien que la variable soit codée comme nominale, l'ordre des modalités est récupéré dans les quantifications.

Figure 11-19
Diagramme de transformation de la variable Journal lu le plus souvent (nominale)

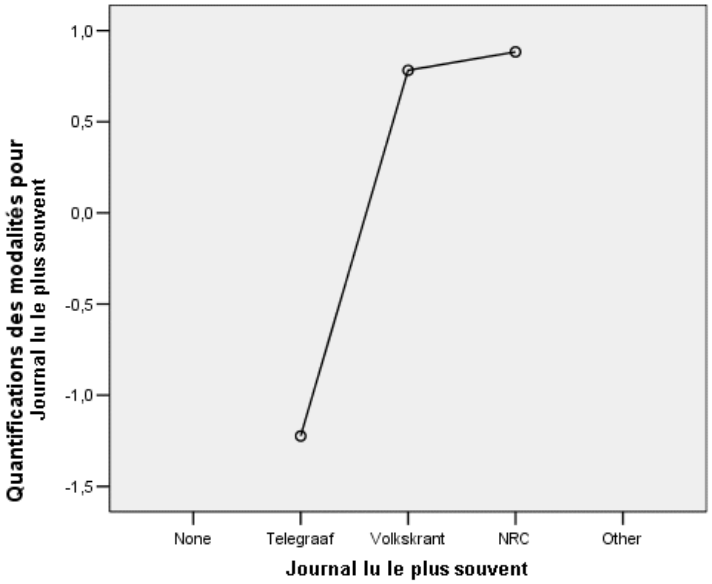
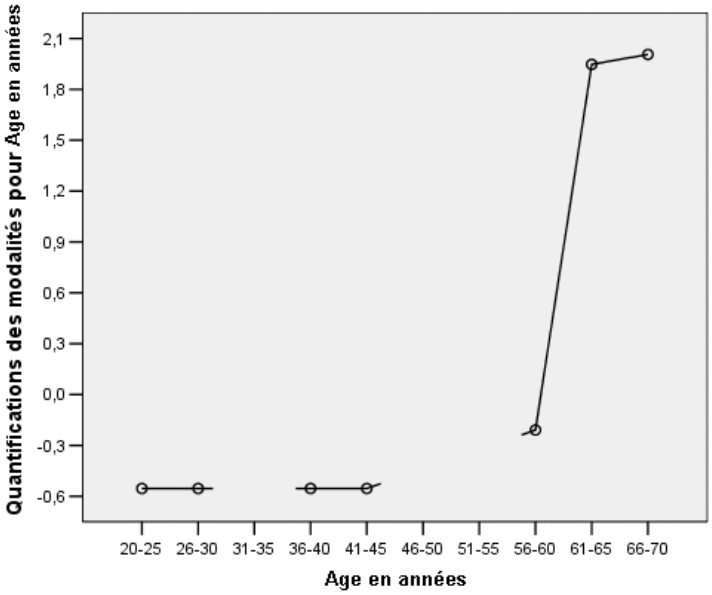


Figure 11-20
Diagramme de transformation de Age en années (ordinale)



Le diagramme de transformation de la variable *Age en années* affiche une courbe en S. Les quatre plus jeunes modalités observées reçoivent toutes la même quantification négative, tandis que les deux modalités les plus vieilles reçoivent les mêmes valeurs positives. Par conséquent, il est possible de fusionner tous les groupes les plus jeunes dans une même modalité (les moins de 50 ans) et de fusionner les deux modalités les plus âgées en une seule. Toutefois, l'égalité parfaite des quantifications des groupes les plus jeunes indique qu'il n'est peut-être pas souhaitable de restreindre l'ordre des quantifications à celui des modalités d'origine. Puisque les quantifications des groupes 26–30, 36–40 et 41–45 ne peuvent pas être inférieures à la quantification du groupe 20–25, ces valeurs sont alignées sur la même valeur de borne. En autorisant ces valeurs à être inférieures à la quantification du groupe le plus jeune (c'est-à-dire, en considérant l'âge comme étant nominal), il est possible d'améliorer l'ajustement. Par conséquent, considérer l'âge comme une variable ordinale ne semble pas approprié dans ce cas. En outre, en considérant l'âge comme une variable numérique, et en conservant donc les distances entre les modalités, il est possible de réduire considérablement l'ajustement.

Coordonnées de modalités simples et coordonnées de modalités multiples

Pour chaque variable considérée comme nominale simple, ordinale ou numérique, les quantifications, les coordonnées de modalité simples et multiples sont déterminées. Ces statistiques sont présentées pour la variable *Age en années*.

Figure 11-21
Coordonnées pour Age en années

	Effectif marginal	Quantification	Coordonnées des modalités uniques		Coordonnées des modalités multiples	
			Dimension		Dimension	
			1	2	1	2
20-25	3	-,554	-,377	-,437	-,192	-,139
26-30	5	-,554	-,377	-,437	-,404	-,623
31-35	0	,000				
36-40	1	-,554	-,377	-,437	-,318	-,733
41-45	1	-,554	-,377	-,437	-,356	-,534
46-50	0	,000				
51-55	0	,000				
56-60	2	-,209	-,142	-,165	-,435	,087
61-65	1	1,947	1,324	1,536	1,710	1,204
66-70	2	2,006	1,364	1,583	1,215	1,711
Manquant	0					

a. Niveau de codage optimal : Ordinal

Les modalités pour lesquelles aucune observation n'est enregistrée reçoivent une quantification de 0. Pour la variable *Age en années*, cela inclut les modalités 31–35, 46–50 et 51–55. Ces modalités ne doivent pas nécessairement être ordonnées avec les autres modalités et n'affectent aucun calcul.

Pour les variables nominales multiples, chaque modalité reçoit une quantification différente sur chaque dimension. Pour tous les autres types de transformation, une modalité ne dispose que d'une quantification, quel que soit le nombre de dimensions de la solution. Chaque ensemble de coordonnées de modalités simples représente l'emplacement de la modalité sur une ligne dans l'espace objet. Pour une modalité donnée, les coordonnées sont égales à la quantification multipliée par les pondérations de dimension de la variable. Par exemple, dans le tableau de la variable *Age en années*, les coordonnées de modalités simples pour la modalité 56-60 (-0,142,

-0,165) correspondent à la quantification (-0,209) multipliée par les pondérations de dimension (0,680, 0,789).

Les coordonnées de modalité multiples des variables considérées comme nominales simples, ordinales ou numériques, représentent les coordonnées des modalités de l'espace objet, avant que les contraintes linéaires ou ordinales soient appliquées. Ces valeurs sont des réducteurs de perte non contraints. Pour les variables nominales multiples, ces coordonnées représentent les quantifications des modalités.

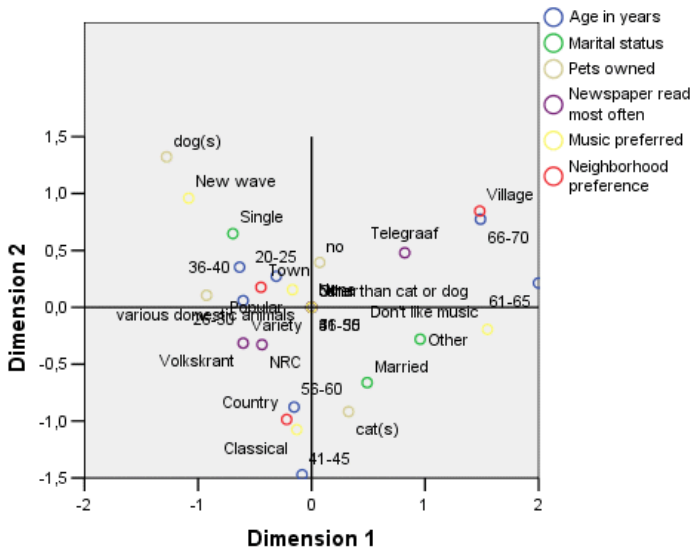
Les effets que peut avoir l'application de contraintes aux relations entre les modalités et leurs quantifications sont révélés par la comparaison des coordonnées de modalités simples avec des coordonnées de modalités multiples. Dans la première dimension, les coordonnées de modalité multiples de la variable *Age en années* diminuent jusqu'à la modalité 2 et restent plus ou moins au même niveau jusqu'à la modalité 9, où se produit une brusque augmentation. Une configuration semblable est mise en évidence pour la seconde dimension. Ces relations sont retirées des coordonnées de modalité simples, auxquelles est appliquée une contrainte ordinale. Dans les deux dimensions, les coordonnées sont alors non décroissantes. Compte tenu de la structure différente des deux groupes de coordonnées, un traitement nominal semble plus approprié.

Barycentres et barycentres projetés

Le diagramme des barycentres étiquetés par des variables doit être interprété de la même manière que le diagramme de quantification des modalités d'une analyse d'homogénéité ou que les coordonnées de modalité multiples d'une analyse non linéaire des composantes principales. Un diagramme de ce type illustre le pouvoir discriminant des variables pour les groupes d'objets (les barycentres sont situés au niveau du centre de gravité des objets).

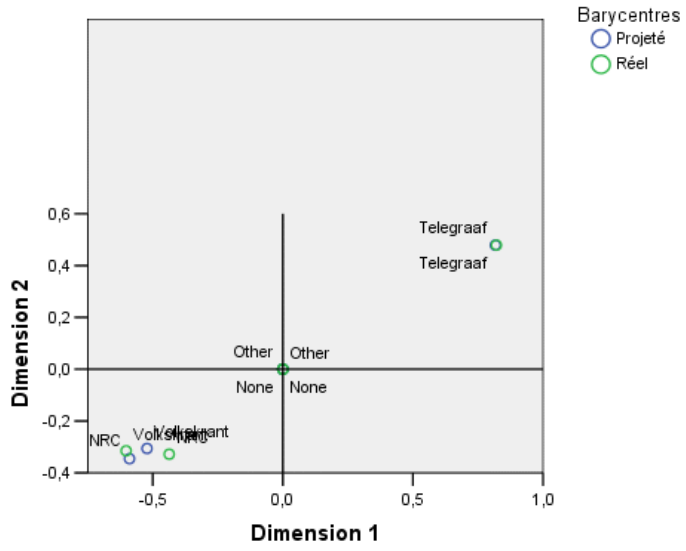
Les modalités de la variable *Age en années* ne sont pas séparées de manière très distincte. Les modalités correspondant aux plus jeunes âges sont regroupées à gauche du diagramme. Comme suggéré précédemment, un niveau de codage ordinal risque d'être trop strict pour la variable *Age en années*.

Figure 11-22
Barycentres étiquetés par des variables



Lorsque vous demandez des diagrammes de représentation des barycentres, des diagrammes de barycentres et de barycentres projetés distincts pour chaque variable étiquetée par des étiquettes de valeurs sont également créés. Les barycentres projetés sont situés sur une ligne de l'espace objet.

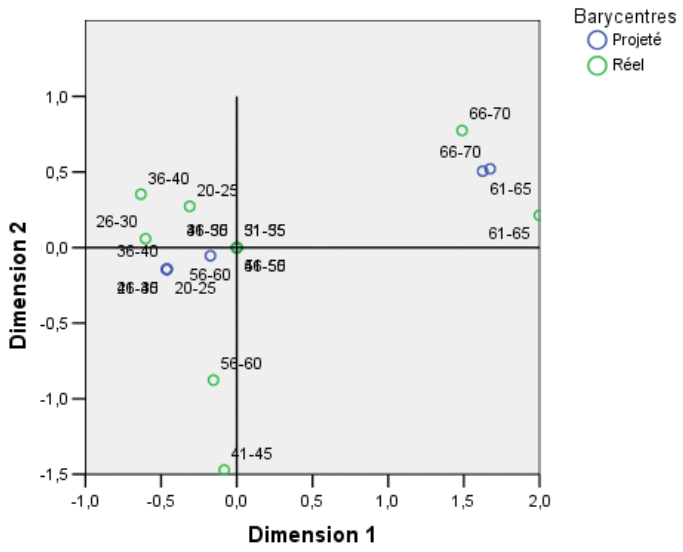
Figure 11-23
Barycentres et barycentres projetés de Journal lu le plus souvent



Les barycentres réels sont projetés sur des vecteurs définis par les corrélations entre composantes. Ces vecteurs ont été ajoutés aux diagrammes de représentation des barycentres afin de faciliter la distinction entre barycentres projetés et barycentres réels. Les barycentres projetés se situent dans l'un des quatre quadrants formés par le tracé de deux lignes de référence perpendiculaires passant par l'origine. L'interprétation de la direction des variables nominales simples, ordinales

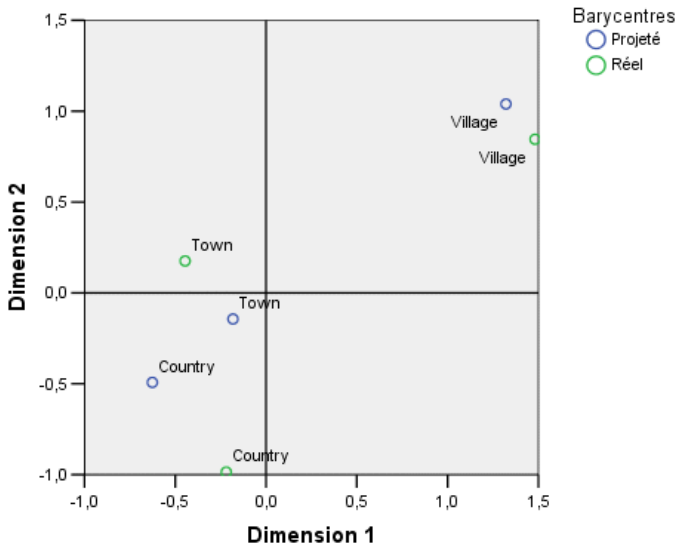
ou numériques est obtenue grâce à la position des barycentres projetés. Par exemple, la variable *Journal lu le plus souvent* est indiquée comme étant nominale simple. Les barycentres projetés mettent en opposition *Volkskrant* et *NRC* d'un côté et *Telegraaf* de l'autre.

Figure 11-24
Barycentres et barycentres projetés de Age en années



Le problème qui se pose pour la variable *Age en années* est mis en évidence par les barycentres projetés. Traiter *Age en années* comme une variable ordinale implique que l'ordre des groupes d'âge soit conservé. Pour respecter cette restriction, tous les groupes d'âge en dessous de 45 sont projetés sur le même point. Sur la direction définie par les variables *Age en années*, *Journal lu le plus souvent* et *Préférence de voisinage*, il n'existe aucune séparation entre les groupes les plus jeunes. Ce constat suggère que l'on traite la variable comme étant nominale.

Figure 11-25
Barycentres et barycentres projetés de Préférence de voisinage



Pour comprendre les relations existant entre les variables, il convient de déterminer les modalités (valeurs) spécifiques des classes de modalités, dans les diagrammes de représentation des barycentres. Les relations existant entre les variables *Age en années*, *Journal lu le plus souvent* et *Préférence de voisinage* peuvent être définies grâce aux angles supérieur droit et inférieur gauche des diagrammes. Dans l'angle supérieur droit, les groupes d'âge correspondent aux répondants les plus âgés ; ces derniers lisent le Telegraaf et préfèrent vivre dans un village. Dans l'angle inférieur gauche de chaque diagramme, il apparaît que les répondants les plus jeunes jusqu'aux répondants d'âge moyen lisent Volkskrant ou NRC, et veulent vivre à la campagne ou en ville. Il est néanmoins difficile de différencier les groupes les plus jeunes.

Le même type d'interprétation peut être appliqué à l'autre direction (*Musique favorite*, *Situation familiale* et *Animaux domestiques possédés*), en étudiant cette fois les angles supérieur gauche et inférieur droit des diagrammes de représentation des barycentres. Dans l'angle supérieur gauche, il apparaît que les personnes célibataires ont souvent des chiens et aiment la musique New wave. Les personnes mariées et la modalité des autres situations familiales ont des chats ; le premier groupe préfère la musique classique et le dernier n'aime pas la musique.

Autre analyse

Compte tenu des résultats de l'analyse, considérer la variable *Age en années* comme étant ordinale ne semble pas approprié. Bien que *Age en années* soit mesuré à un niveau ordinal, ses relations avec les autres variables ne sont pas monotones. Pour étudier les effets d'un changement du niveau de codage optimal en niveau de codage nominal simple, relancez l'analyse.

Pour lancer l'analyse

- ▶ Rappelez la boîte de dialogue Analyse canonique non linéaire et déplacez-vous jusqu'au premier groupe.
- ▶ Sélectionnez *âge* et cliquez sur Définir intervalle et échelle.
- ▶ Dans la boîte de dialogue Définir intervalle et échelle, sélectionnez Nominale simple comme intervalle de codage.
- ▶ Cliquez sur Poursuivre.
- ▶ Dans la boîte de dialogue Analyse canonique non linéaire, cliquez sur OK.

Les valeurs propres d'une solution à deux dimensions sont respectivement 0,806 et 0,757, avec un ajustement total de 1,564.

Figure 11-26
Valeurs propres d'une solution à deux dimensions

		Dimension		
		1	2	Somme
Perte	Groupe 1	,249	,115	,363
	Groupe 2	,176	,408	,584
	Groupe 3	,157	,205	,363
	Moyenne	,194	,243	,436
Valeur propre		,806	,757	
Ajustement LDN				1,564

Les tableaux d'ajustement multiple et d'ajustement unique montrent que la variable *Age en années* a toujours un fort pouvoir discriminant, comme l'illustre la somme des valeurs d'ajustement multiple. Toutefois, contrairement aux précédents résultats, l'examen des valeurs d'ajustement unique révèle que ce pouvoir discriminant concerne principalement la deuxième dimension.

Figure 11-27
Partitionnement des ajustements et des pertes

		Ajustement multiple			Ajustement unique			Perte unique		
		Dimension			Dimension			Dimension		
Groupe		1	2	Somme	1	2	Somme	1	2	Somme
1	Age in years[a]	,246	1,197	1,443	,195	1,188	1,384	,051	,008	,059
	Marital status[a]	,273	1,136	1,409	,272	1,135	1,407	,001	,000	,002
2	Pets owned[b]	,530	,392	,921						
	Newspaper read most often[a]	,639	,185	,824	,631	,149	,780	,008	,036	,044
3	Music preferred[a]	,604	,438	1,041	,603	,437	1,040	,000	,001	,001
	Neighborhood preference[a]	,075	,822	,897	,075	,822	,897	,000	,000	,000

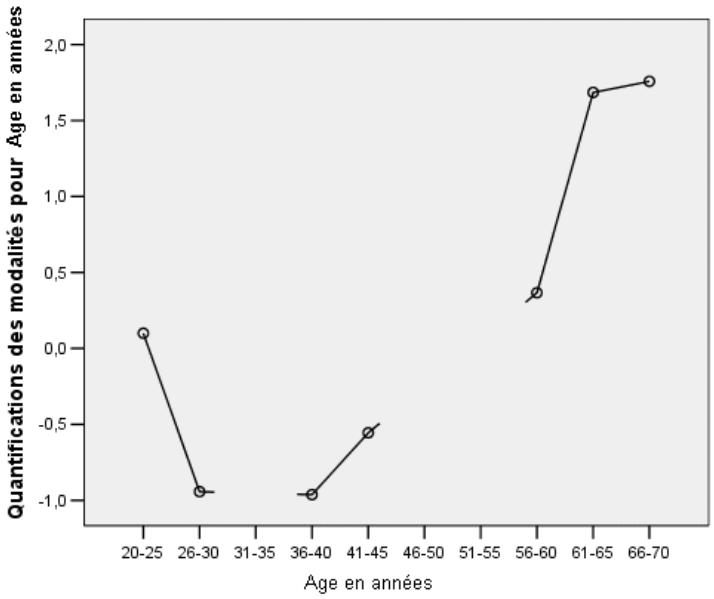
a. Niveau de codage optimal : Nominal simple

b. Niveau de codage optimal : Nominal multiple

Reportez-vous au diagramme de transformation pour la variable *Age en années*. Les quantifications d'une variable nominale n'ont pas de restriction ; par conséquent, la tendance non décroissante affichée lorsque la variable *Age en années* était traitée de manière ordinale n'est plus présente. Il y a une diminution jusqu'à 40 ans et une augmentation au-delà de 40 ans, qui correspondent à une relation en U (quadratique). Les deux modalités les plus âgées reçoivent

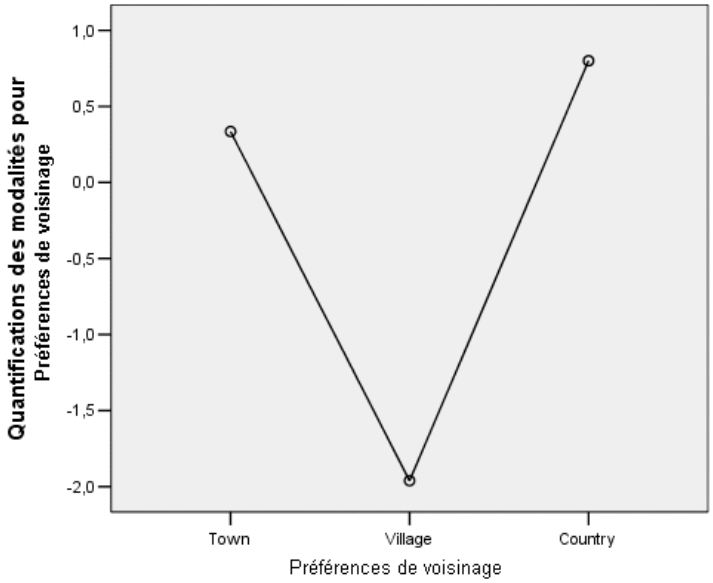
toujours les mêmes scores, et les analyses suivantes risquent de nécessiter une combinaison de ces modalités.

Figure 11-28
Diagramme de transformation de Age en années (nominale)



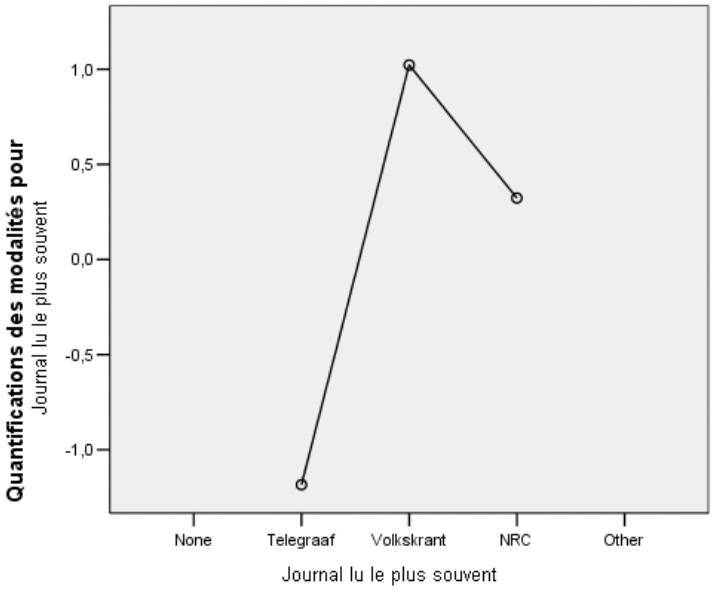
Le diagramme de transformation de la variable *Préférence de voisinage* est affiché ici. Considérer *Age en années* comme une variable nominale n'affecte en aucun cas les quantifications de la variable *Préférence de voisinage*. La modalité centrale reçoit la plus petite quantification, et les modalités extrêmes reçoivent des valeurs positives élevées.

Figure 11-29
Diagramme de transformation de la variable Préférence de voisinage (âge, nominale)



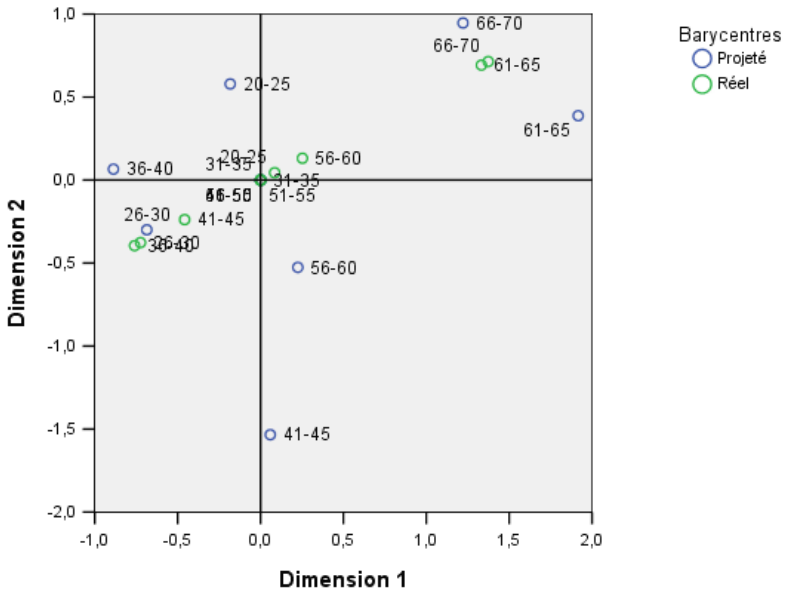
On remarque un changement dans le diagramme de transformation de la variable *Journal lu le plus souvent*. On pouvait noter auparavant une augmentation dans les quantifications, ce qui pouvait suggérer un traitement ordinal de cette variable. Toutefois, en traitant *Age en années* comme une variable nominale, on élimine cette tendance des quantifications liées à la presse.

Figure 11-30
Diagramme de transformation de la variable Journal lu le plus souvent (âge, nominale)



Il s'agit du diagramme de représentation des barycentres de la variable *Age en années*. Remarque : les modalités n'apparaissent pas toutes dans l'ordre chronologique sur la ligne joignant les barycentres projetés. Le groupe 20–25 est situé au centre plutôt qu'à la fin. La répartition des modalités s'avère nettement meilleure que dans l'exemple de traitement ordinal présenté ci-dessus.

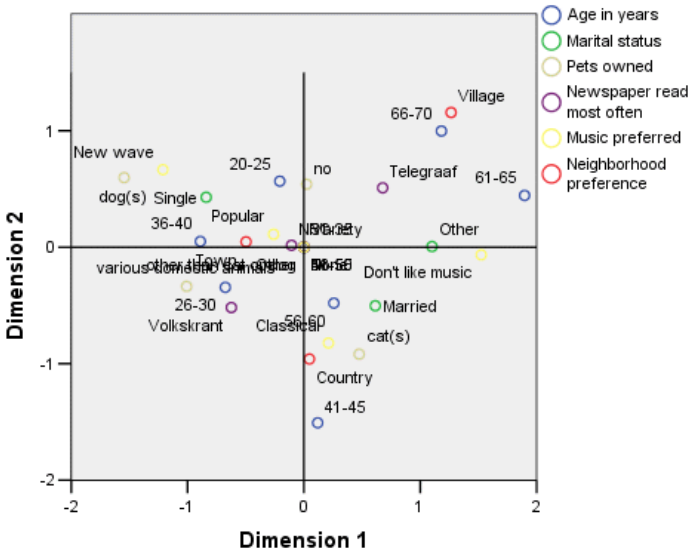
Figure 11-31
Barycentres et barycentres projetés de Age en années (nominale)



Il est à présent possible de fournir une interprétation des groupes les plus jeunes, à partir du diagramme de représentation des barycentres. Les modalités *Volkskrant* et *NRC* sont plus éloignées que dans l'analyse précédente, ce qui permet de fournir une interprétation distincte pour chacune d'elles. Les groupes dont les personnes sont âgées entre 26 et 45 ans lisent Volkskrant et préfèrent vivre à la campagne. Les groupes d'âge 20–25 et 56–60 lisent NRC ; le premier groupe préfère vivre en ville et le deuxième à la campagne. Les groupes les plus âgés lisent le Telegraaf et préfèrent vivre dans un village.

L'interprétation de l'autre direction (*Musique favorite*, *Situation familiale* et *Animaux domestiques possédés*) reste quasiment inchangée par rapport à la précédente analyse. La seule différence nette est que les personnes ayant répondu *Autre* pour la situation familiale ont soit un chat, soit aucun animal domestique.

Figure 11-32
Barycentres étiquetés par des variables (âge, nominal)



Suggestions d'ordre général

Une fois les résultats initiaux étudiés, vous pouvez affiner votre analyse en modifiant certains paramètres de l'analyse de corrélation canonique non linéaire. Voici quelques conseils pour structurer votre analyse :

- Créez autant de groupes que possible. Placez une variable importante, que vous souhaitez évaluer, toute seule dans un groupe distinct.
- Regroupez ensemble les variables indépendantes. En présence de nombreuses variables indépendantes, essayez de les répartir dans différents groupes.
- Placez une variable nominale multiple toute seule dans un groupe distinct.
- Si des variables présentent une forte corrélation entre elles et que vous ne souhaitez pas que cette relation influence la solution, placez-les ensemble dans le même groupe.

Lectures recommandées

Pour plus d'informations sur l'analyse de corrélation canonique non linéaire, reportez-vous aux documents suivants :

Carroll, J. D. 1968. Generalization of canonical correlation analysis to three or more sets of variables. Dans : *Proceedings of the 76th Annual Convention of the American Psychological Association, 3,* Washington, D.C.: American Psychological Association.

De Leeuw, J. 1984. *Canonical analysis of categorical data*, 2nd éd. Leiden: DSWO Press.

Horst, P. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17, .

Horst, P. 1961. Relations among m sets of measures. *Psychometrika*, 26, .

Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika*, 58, .

Van der Burg, E. 1988. *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO Press.

Van der Burg, E., et J. De Leeuw. 1983. Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, .

Van der Burg, E., J. De Leeuw, et R. Verdegaal. 1988. Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 53, .

Verboon, P., et R. A. Van der Lans. 1994. Robust canonical discriminant analysis. *Psychometrika*, 59, .

Chapitre
12

Analyse des correspondances

Un **tableau des correspondances** est tout tableau à deux entrées dont les cellules contiennent une certaine mesure de correspondance entre les lignes et les colonnes. La mesure de correspondance peut être toute indication de la similarité, du rapport, de la confusion, de l'association ou de l'interaction entre les variables de ligne et de colonne. Un type très courant de tableau des correspondances est le tableau croisé, dont les cellules contiennent des effectifs.

La procédure de tableaux croisés permet d'obtenir facilement de tels tableaux. Toutefois, un tableau croisé ne fournit pas toujours une image claire de la nature de la relation entre les deux variables. Cela est particulièrement vrai si les variables d'intérêt sont nominales (sans ordre ou rang inhérent) et qu'elles contiennent de nombreuses modalités. Le tableau croisé peut indiquer que les effectifs observés par cellule diffèrent sensiblement des effectifs prévus dans un tableau croisé $10x9$ de *profession* et de *céréale pour le petit déjeuner*, mais il peut être difficile de discerner les groupes professionnels qui présentent des goûts similaires ou ce que sont ces goûts.

L'analyse des correspondances vous permet d'examiner graphiquement la relation entre deux variables nominales dans un espace multidimensionnel. Elle calcule les coordonnées principales des colonnes et des lignes et génère des diagrammes basés sur les scores. Les modalités similaires apparaissent proches les unes des autres dans les diagrammes. Ainsi, il est facile de repérer les modalités similaires d'une variable ou les modalités liées entre les deux variables. En outre, la procédure de l'analyse des correspondances vous permet d'ajuster des points supplémentaires dans l'espace défini par les points actifs.

Si l'ordre des modalités en fonction de leurs scores est indésirable ou paradoxal, vous pouvez imposer des restrictions d'ordre en contraignant les scores de certaines modalités à être égaux. Par exemple, nous pouvons imaginer que la variable *consommation de tabac* ayant pour modalités *non-fumeur*, *léger fumeur*, *fumeur moyen* et *gros fumeur* possède des scores correspondant à cet ordre. Toutefois, si l'analyse classe les modalités dans l'ordre suivant : *non-fumeur*, *léger fumeur*, *gros fumeur* et *fumeur moyen*, le fait de contraindre les scores de *gros fumeur* et *fumeur moyen* à être égaux protège l'ordre des modalités dans leurs scores.

L'interprétation de l'analyse des correspondances en matière de distances dépend de la méthode de normalisation utilisée. La procédure d'analyse des correspondances permet d'analyser les différences entre les modalités d'une variable ou celles entre les variables. Selon la normalisation par défaut, elle analyse les différences entre les variables de ligne et de colonne.

L'algorithme d'analyse des correspondances autorise de nombreux types d'analyse. Le centrage des lignes et des colonnes et l'utilisation de distances Khi-deux relèvent de l'analyse de correspondance standard. Toutefois, l'utilisation d'autres options de centrage combinées avec des distances euclidiennes permet de varier la représentation d'une matrice dans un espace de petite dimension.

Trois exemples seront présentés. Le premier utilise un tableau des correspondances relativement réduit et illustre les concepts inhérents à l'analyse des correspondances. Le deuxième exemple illustre une application marketing. Le dernier exemple utilise un tableau de distances dans une approche de positionnement multidimensionnel.

Normalisation

La normalisation permet de répartir l'inertie sur les coordonnées principales des colonnes et des lignes. Certains aspects de la solution d'analyse des correspondances, tels que les valeurs singulières, l'inertie par dimension et les contributions, ne changent pas d'une normalisation à l'autre. Les coordonnées principales des colonnes et des lignes et leurs variances sont affectées. L'analyse des correspondances peut répartir l'inertie de plusieurs façons. Les trois façons les plus courantes sont la répartition sur les coordonnées principales des lignes uniquement, la répartition sur les coordonnées principales des colonnes uniquement ou la répartition symétrique sur, à la fois, les coordonnées principales des lignes et les coordonnées principales des colonnes.

Principale en ligne : Dans la normalisation principale en ligne, les distances euclidiennes entre les points des lignes se rapprochent des distances Khi-deux entre les lignes du tableau des correspondances. Les scores des lignes correspondent à la moyenne pondérée des scores des colonnes. Les coordonnées principales des colonnes sont standardisées de manière à avoir une somme pondérée des carrés des distances par rapport au centre égale à 1. Dans la mesure où cette méthode maximise les distances entre les modalités de ligne, vous devez utiliser la normalisation principale en ligne si vous avez essentiellement l'intention d'observer les différences entre les modalités de la variable de ligne.

Principale en colonne : Par ailleurs, vous pouvez approximer les distances Khi-deux entre les colonnes du tableau des correspondances. Dans ce cas, les coordonnées principales des colonnes doivent correspondre à la moyenne pondérée des coordonnées principales des lignes. Les coordonnées principales des lignes sont standardisées de manière à avoir une somme pondérée des carrés des distances par rapport au centre égale à 1. Cette méthode maximise les distances entre les modalités de colonnes et vous devez l'utiliser si vous avez essentiellement l'intention d'observer les différences entre les modalités de la variable de colonne.

Symétrique : En outre, vous pouvez traiter les lignes et les colonnes de manière symétrique. Cette normalisation répartit l'inertie de façon égale sur les coordonnées des lignes et des colonnes. Dans ce cas, ni les distances entre les points des lignes ni celles entre les points des colonnes ne sont des approximations de distances Khi-deux. Utilisez cette méthode si vous envisagez essentiellement d'examiner les différences ou les similitudes entre les deux variables. Généralement, cette méthode est à privilégier pour réaliser des diagrammes doubles.

Principale. Une quatrième option, la normalisation principale, permet de répartir l'inertie deux fois dans la solution —une fois sur les coordonnées des lignes et une fois sur celles des colonnes. Vous devez utiliser cette méthode si vous souhaitez examiner les distances entre les points des lignes et celles entre les points des colonnes séparément, sans vouloir analyser la relation entre les points lignes et colonnes. Les diagrammes doubles n'étant pas appropriés pour cette option de normalisation, ils ne sont pas disponibles si vous avez spécifié la méthode de normalisation principale.

Exemple : Perceptions des marques de café

L'exemple précédent repose sur un petit tableau de données hypothétiques. Les applications réelles impliquent souvent des tableaux beaucoup plus volumineux. Dans cet exemple, vous utiliserez des données relatives aux images perçues de six marques de café frappé (Kennedy, Riquier, et Sharp,

1996). Cet ensemble de données est disponible dans le fichier *coffee.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*.

Pour chacun des 23 attributs d'image de café frappé, les personnes sollicitées ont sélectionné toutes les marques décrites par l'attribut. Les six marques sont appelées *AA*, *BB*, *CC*, *DD*, *EE* et *FF* à des fins de confidentialité.

Table 12-1
Attributs du café frappé

Attribut image	Etiquette	Attribut image	Etiquette
bon remède contre la gueule de bois	*remède*	produit non allégé	*fait grossir*
produit allégé/faible en calories	*allégé*	plaît aux hommes	*hommes*
marque ciblant les enfants	*enfants*	marque sud-australienne	*Australie du sud*
marque de la classe ouvrière	*classe ouvrière*	marque traditionnelle/démodée	*traditionnel*
produit fort en goût/léger en goût	*léger en goût*	marque de luxe	*luxe*
marque impopulaire	*impopulaire*	marque bio	*bio*
marque pour personnes obèses/laides	*laid*	produit fortement caféiné	*caféine*
très frais	*frais*	nouvelle marque	*nouveau*
marque pour jeunes cadres dynamiques	*jeunes cadres dynamiques*	marque pour personnes séduisantes	*séduisant*
produit nourrissant	*nourrissant*	marque fiable	*fiable*
marque pour femmes	*femmes*	marque populaire	*populaire*
marque secondaire	*secondaire*		

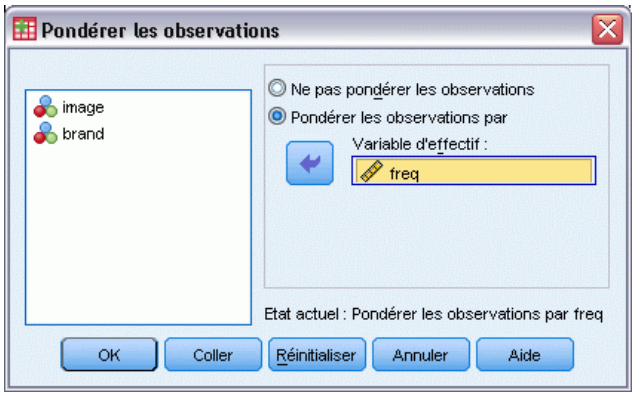
Dans un premier temps, vous allez vous concentrer sur les liens unissant les attributs et sur ceux unissant les marques. L'utilisation de la normalisation principale répartit l'inertie totale une fois sur les lignes et une fois sur les colonnes. Bien que cela empêche l'interprétation des diagrammes doubles, vous pouvez examiner les distances entre les modalités de chaque variable.

Exécution de l'analyse

- La configuration des données implique que les observations soient pondérées par la variable *freq*. Pour ce faire, dans les menus, choisissez :

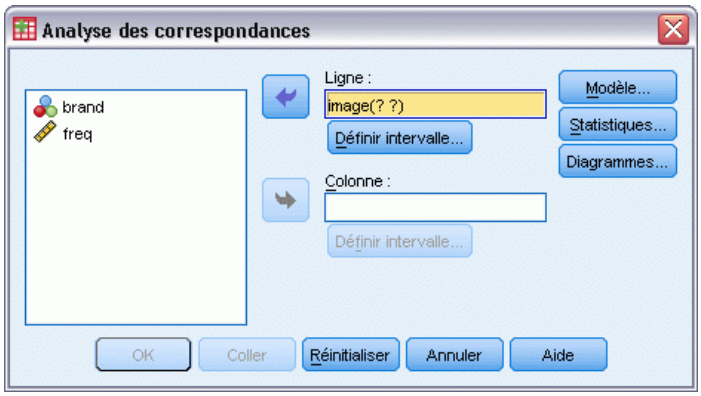
  Données > Pondérer les observations

Figure 12-1
Boîte de dialogue Pondérer les observations



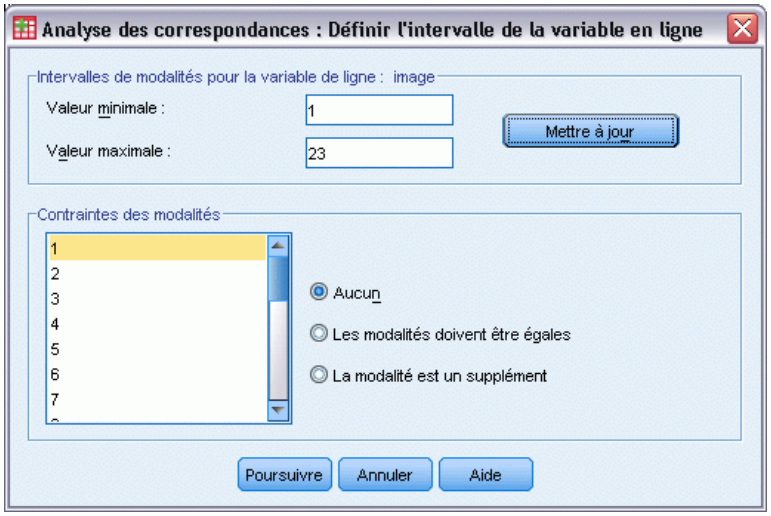
- Pondérez les observations par la variable *freq*.
- Cliquez sur OK.
- Pour obtenir une solution initiale dans cinq dimensions en recourant à la normalisation principale, choisissez dans les menus :
  Analyse > Réduction des dimensions > Analyse des correspondances...

Figure 12-2
Boîte de dialogue Analyse des correspondances



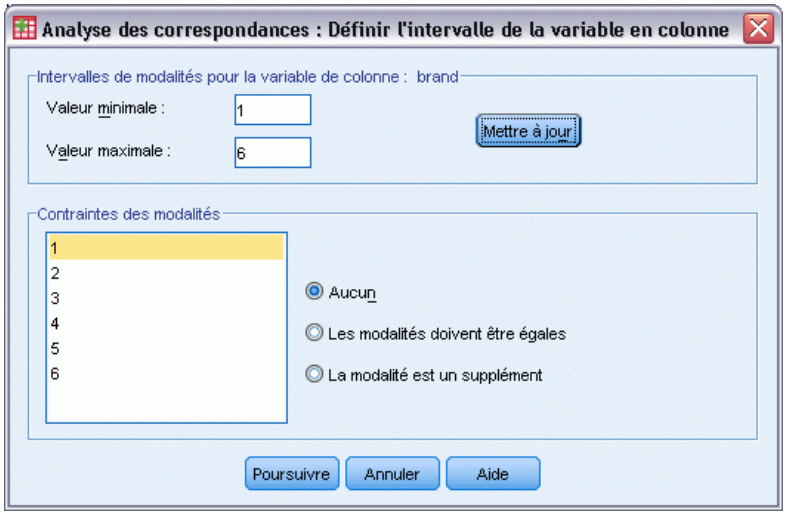
- Sélectionnez l'option *image* comme variable de ligne.
- Cliquez sur Définir intervalle.

Figure 12-3
Boîte de dialogue Définir l'intervalle de la variable en ligne



- Tapez 1 comme valeur minimale.
- Tapez 23 comme valeur maximale.
- Cliquez sur Mettre à jour.
- Cliquez sur Poursuivre.
- Sélectionnez l'option *marque* comme variable de colonne.
- Cliquez sur l'option Définir intervalle dans la boîte de dialogue Analyse des correspondances.

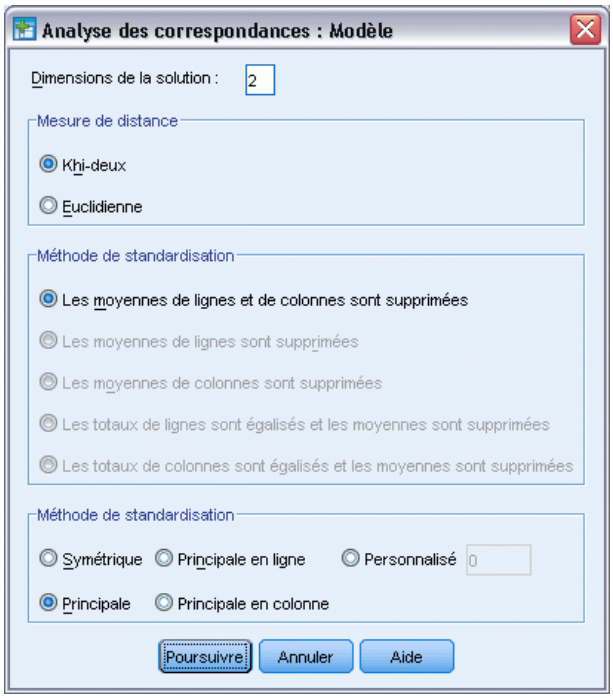
Figure 12-4
Boîte de dialogue Définir l'intervalle de la variable en colonne



- Tapez 1 comme valeur minimale.
- Tapez 6 comme valeur maximale.

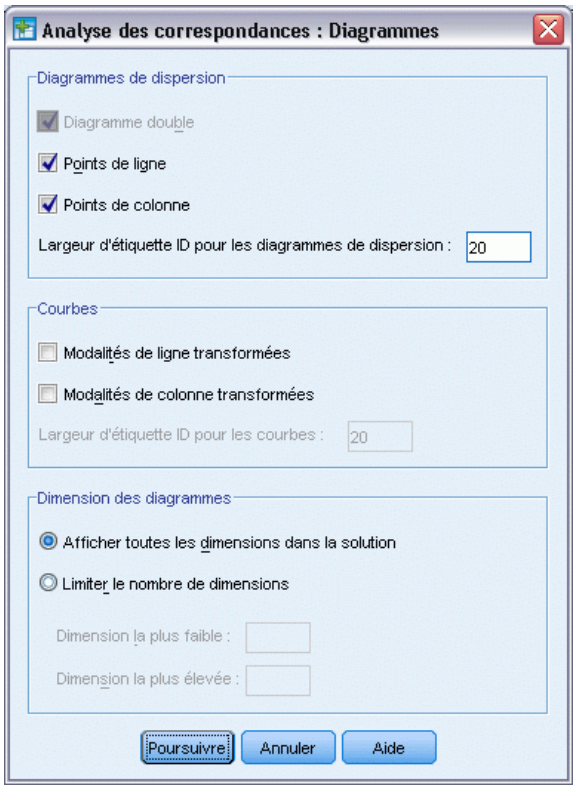
- Cliquez sur Mettre à jour.
- Cliquez sur Poursuivre.
- Cliquez sur l'option Modèle dans la boîte de dialogue Analyse des correspondances.

Figure 12-5
Boîte de dialogue Modèle



- Sélectionnez l'option Principale comme méthode de normalisation.
- Cliquez sur Poursuivre.
- Cliquez sur le bouton Diagrammes dans la boîte de dialogue Analyse des correspondances.

Figure 12-6
Boîte de dialogue Diagrammes



- Sélectionnez les options Points lignes et Points colonnes dans le groupe Diagrammes de dispersion.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Analyse des correspondances.

Nombre de dimensions

L'inertie par dimension indique la décomposition de l'inertie totale le long de chaque dimension. Deux dimensions représentent 83 % de l'inertie totale. L'ajout d'une troisième dimension augmente l'inertie prise en compte de 8,6 % uniquement. Par conséquent, vous optez pour l'utilisation d'une représentation bidimensionnelle.

Figure 12-7
Inertie par dimension

Dimension	Valeur singulière	Inertie	Khi-deux	Sig.	Proportion d'inertie		Valeur singulière de confiance	
					Expliqué	Cumulé	Ecart-type	Corrélation 2
1	,711	,506			,629	,629	,009	,132
2	,399	,159			,198	,827	,014	
3	,263	,069			,086	,913		
4	,234	,055			,068	,982		
5	,121	,015			,018	1,000		
Total		,804	3746,968	,000[a]	1,000	1,000		

a. 110 degrés de liberté

Contributions

Les caractéristiques des points des lignes montrent les contributions des points des lignes à l'inertie des dimensions et les contributions des dimensions à l'inertie des points des lignes. Si tous les points contribuent de façon égale à l'inertie, les contributions ont pour valeur 0,043. Les points *bio* et *allégé* contribuent de façon substantielle à l'inertie de la première dimension. Les points *hommes* et *fiable* sont les éléments qui contribuent le plus à l'inertie de la deuxième dimension. Les deux points *laid* et *frais* contribuent très peu aux deux dimensions.

Figure 12-8
Contributions des attributs

image	Masse	Score dans la dimension		Inertie	Contribution				
					De point à inertie de dimension		De dimension à inertie de point		
		1	2		1	2	1	2	Total
fattening	,080	-,514	-,265	,033	,042	,035	,652	,173	,825
men	,051	-,852	,825	,072	,073	,219	,512	,480	,992
South Australian	,057	-,303	-,350	,046	,010	,044	,114	,152	,266
traditional	,040	-,703	-,532	,043	,039	,071	,454	,260	,715
premium	,042	-,444	-,582	,028	,016	,090	,296	,509	,805
healthy	,053	1,200	,174	,081	,152	,010	,953	,020	,973
caffeine	,047	-,452	,124	,014	,019	,005	,702	,053	,755
new	,047	,960	,147	,048	,086	,006	,893	,021	,914
attractive	,041	,657	-,056	,019	,035	,001	,911	,007	,918
tough	,039	-,850	1,002	,070	,056	,246	,404	,560	,964
popular	,060	-,697	-,042	,038	,058	,001	,771	,003	,774
cure	,026	-,389	,266	,009	,008	,011	,446	,209	,655
low fat	,052	1,305	,196	,094	,175	,013	,941	,021	,962
children	,024	-,352	-,513	,017	,006	,041	,179	,380	,559
working	,045	-,785	,477	,040	,055	,064	,693	,255	,948
sweet	,038	-,519	-,683	,048	,020	,112	,212	,368	,580
unpopular	,024	,489	,186	,010	,011	,005	,585	,085	,670
ugly	,030	,006	-,109	,003	,000	,002	,000	,131	,131
fresh	,036	-,096	-,100	,002	,001	,002	,196	,214	,410
yuppies	,034	,380	-,301	,012	,010	,019	,392	,246	,637
nutritious	,040	,722	,055	,022	,041	,001	,946	,006	,951
women	,054	,758	-,063	,032	,062	,001	,965	,007	,972
minor	,040	,579	,063	,023	,027	,001	,593	,007	,600
Total actif	1,000			,804	1,000	1,000			

a. Normalisation principale

Deux dimensions contribuent sensiblement à l'inertie de la plupart des points des lignes. Les contributions importantes de la première dimension aux points *bio*, *nouveau*, *séduisant*, *allégé*, *nourrissant* et *femmes* indiquent que ces points sont très bien représentés dans une dimension. Par conséquent, les dimensions plus élevées contribuent peu à l'inertie de ces points, qui figurent très près de l'axe horizontal. La deuxième dimension contribue essentiellement aux points *hommes*, *luxe* et *fiable*. Les deux dimensions contribuent très peu à l'inertie pour les points *Australie du sud* et *laid*, si bien que ceux-ci sont faiblement représentés.

Les caractéristiques des points des colonnes montrent les contributions impliquant les points des colonnes. Les marques *CC* et *DD* contribuent le plus à la première dimension, tandis que les marques *EE* et *FF* expliquent une large part de l'inertie de la deuxième dimension. Les marques *AA* et *BB* contribuent très peu aux deux dimensions.

Figure 12-9
Contributions des marques

		Score dans la dimension			Contribution				
					De point à inertie de dimension		De dimension à inertie de point		
brand	Masse	1	2	Inertie	1	2	1	2	Total
AA	,217	-,659	,046	,127	,187	,003	,744	,004	,748
BB	,131	-,284	-,404	,078	,021	,134	,135	,272	,407
CC	,185	,996	,076	,193	,362	,007	,951	,006	,957
DD	,162	,915	,101	,146	,267	,010	,928	,011	,939
EE	,152	-,651	,706	,153	,127	,477	,420	,494	,914
FF	,153	-,343	-,618	,107	,036	,369	,169	,550	,718
Total actif	1,000			,804	1,000	1,000			

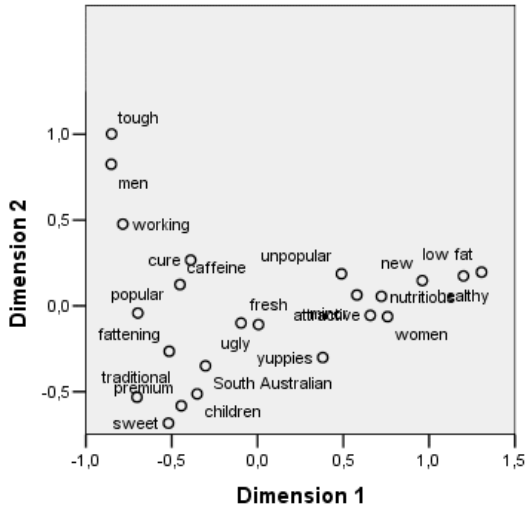
a. Normalisation principale

Dans les deux dimensions, toutes les marques sauf *BB* sont bien représentées. Les marques *CC* et *DD* sont bien représentées dans une dimension. La deuxième dimension représente les parts les plus importantes des marques *EE* et *FF*. La marque *AA* est bien représentée dans la première dimension, mais elle ne contribue pas sensiblement à cette dimension.

Diagrammes

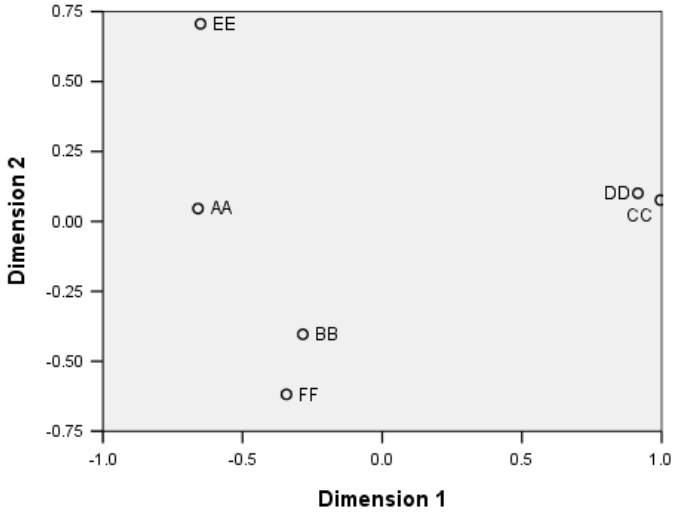
Le diagramme des points des lignes montre que les points *frais* et *laid* sont très proches de l'origine, ce qui indique qu'ils diffèrent peu du profil de ligne moyen. Trois classifications générales émergent. Situés dans la partie supérieure gauche du diagramme, les points *fiable*, *hommes* et *classe ouvrière* sont tous similaires les uns aux autres. La partie inférieure gauche contient les points *léger en goût*, *non allégé*, *enfants* et *luxe*. A l'opposé, les points *bio*, *allégé*, *nourrissant* et *nouveau* sont regroupés sur le côté droit du diagramme.

Figure 12-10
Diagramme d'attributs d'image (normalisation principale)



Dans le diagramme des points des colonnes, toutes les marques étant éloignées de l'origine, aucune d'elles n'est similaire au centre global. Les marques *CC* et *DD* sont regroupées à droite, tandis que les marques *BB* et *FF* sont regroupées dans la moitié inférieure du diagramme. Les marques *AA* et *EE* ne sont similaires à aucune autre marque.

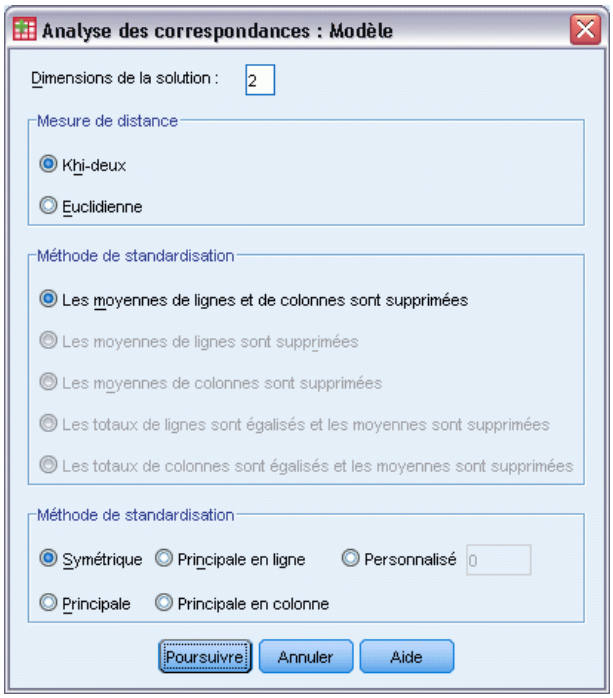
Figure 12-11
Diagramme de marques (normalisation principale)



Normalisation symétrique

Comment les marques sont-elles liées aux attributs d'image ? La normalisation principale ne peut pas traiter ces relations. Pour déterminer les liens entre les variables, utilisez la normalisation symétrique. Au lieu de répartir l'inertie deux fois (comme dans la normalisation principale), la normalisation symétrique la divise de façon égale sur les lignes et sur les colonnes. Les distances entre les modalités d'une variable ne peuvent pas être interprétées, mais celles entre les modalités de différentes variables sont significatives.

Figure 12-12
Boîte de dialogue Modèle

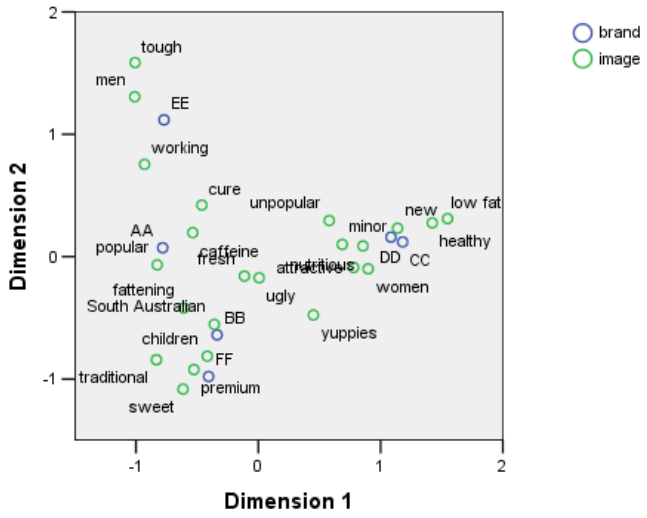


- Pour générer la solution suivante à l'aide de la normalisation symétrique, affichez de nouveau la boîte de dialogue Analyse des correspondances, puis cliquez sur Modèle.
- Sélectionnez l'option Symétrique comme méthode de normalisation.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Analyse des correspondances.

Dans la partie supérieure gauche du diagramme double obtenu, la marque *EE* est la seule marque solide, associée à la classe ouvrière et plaisant aux hommes. La marque *AA* est la plus populaire. En outre, elle est perçue comme étant la plus fortement caféinée. Les marques légères en goût et

non allégées sont *BB* et *FF*. Les marques *CC* et *DD*, tout en étant perçues comme nouvelles et saines, sont les plus impopulaires.

Figure 12-13
Diagramme double des marques et des attributs (normalisation symétrique)



Pour une interprétation plus approfondie, vous pouvez dessiner une ligne passant par l'origine et les deux attributs d'image *hommes*et *jeunes cadres dynamiques*, puis projeter les marques sur cette ligne. Les deux attributs sont opposés l'un à l'autre, ce qui indique que le modèle d'association des marques pour *hommes* est inversé par rapport au modèle pour *jeunes cadres dynamiques*. Autrement dit, les hommes sont le plus fréquemment associés à la marque *EE* et le moins fréquemment à la marque *CC*, tandis que les jeunes cadres dynamiques sont le plus fréquemment associés à la marque *CC* et le moins fréquemment à la marque *EE*.

Lectures recommandées

Pour plus d'informations sur l'analyse des correspondances, reportez-vous aux documents suivants :

Fisher, R. A. 1938. *Statistical methods for research workers*. Edimbourg: Oliver and Boyd.

Fisher, R. A. 1940. The precision of discriminant functions. *Annals of Eugenics*, 10, .

Gilula, Z., et S. J. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association*, 83, .

Chapitre
13

Analyse de correspondance multiple

L'objectif de l'analyse de correspondance multiple, également connue sous le nom d'analyse d'homogénéité, est de rechercher les quantifications optimales dans la mesure où les modalités sont le plus possible séparées les unes des autres. Les objets de la même modalité doivent donc être représentés proches les uns des autres et les objets de modalités différentes doivent être représentés aussi éloignés que possible. Le terme **homogénéité** fait également référence au fait que l'analyse est d'autant plus réussie que les variables sont homogènes, c'est-à-dire lorsqu'elles partitionnent les objets en classes ayant les mêmes modalités ou des modalités similaires.

Exemple : Descriptives du matériel

Pour connaître le fonctionnement de l'analyse de correspondance multiple, reportez-vous aux données de Hartigan(Hartigan, 1975), que vous trouverez dans le fichier *screws.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*. Cet ensemble de données contient des informations sur les descriptives des vis, des boulons, des écrous et des broquettes. Le tableau suivant indique les variables (et leurs étiquettes) et les étiquettes de valeur affectées aux modalités de chaque variable dans l'ensemble de données matérielles de Hartigan.

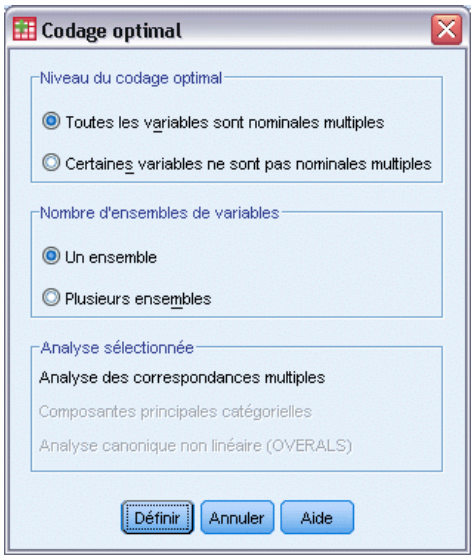
Table 13-1
Ensemble de données matérielles de Hartigan

Nom de variable	l'étiquette Variable	Etiquette de valeur
filetage	*Filetage*	*Yes_Thread, No_Thread*
titre	*Forme de tête*	*Plate, Creuse, Cônique, Arrondie, Cylindrique*
indtête	*Indentation de la tête*	*Aucune, Cruciforme, Fendue*
tige	*Forme tige*	*pointe, plate*
longueur	*Longueur en demi-pouces*	*1/2_in, 1_in, 1_1/2_ in, 2_in, 2_1/2_in*
cuivre	*Cuivre*	*Yes_Br, Not_Br*
objet	*Objet*	*broquette, clou1, clou2, clou3, clou4, clou5, clou6, clou7, clou8, vis1, vis2, vis3, vis4, vis5, boulon1, boulon2, boulon3, boulon4, boulon5, boulon6, broquette1, broquette2, cloub, visb*

Exécution de l'analyse

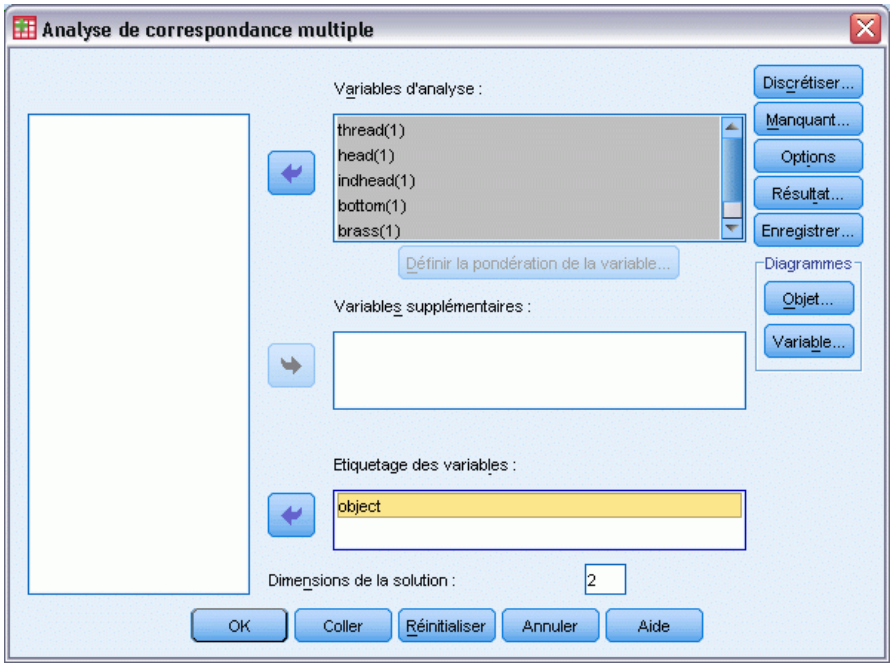
- Pour obtenir une analyse de correspondance multiple, à partir des menus, sélectionnez :
  Analyse > Réduction des dimensions > Codage optimal

Figure 13-1
Boîte de dialogue Niveau du codage optimal



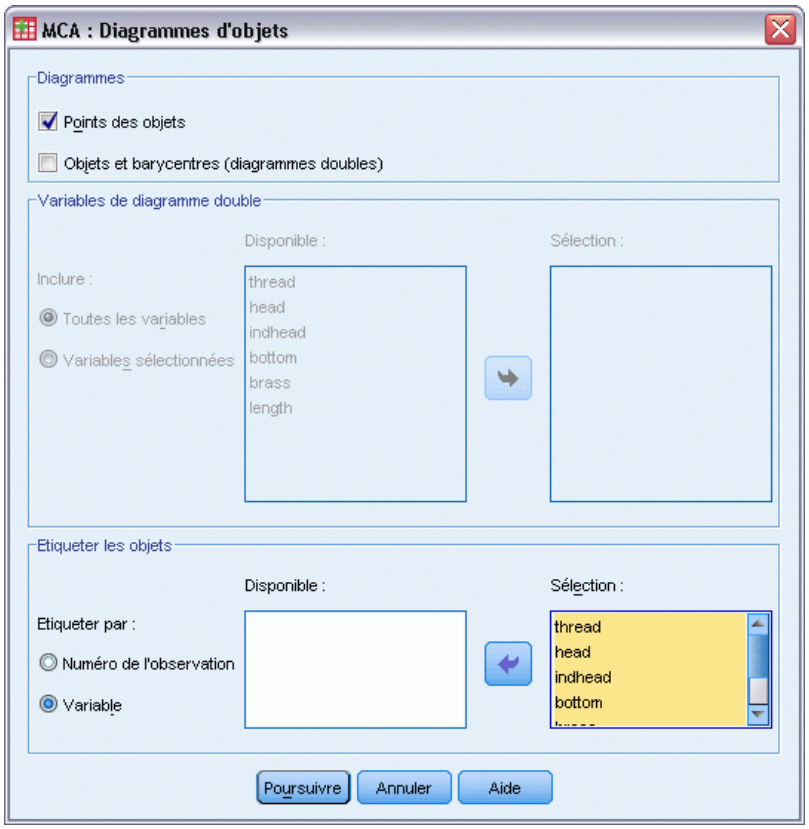
- Assurez-vous que les options Toutes les variables sont nominales multiples et Un groupe sont sélectionnées, puis cliquez sur Définir.

Figure 13-2
Boîte de dialogue Analyse des correspondances multiples



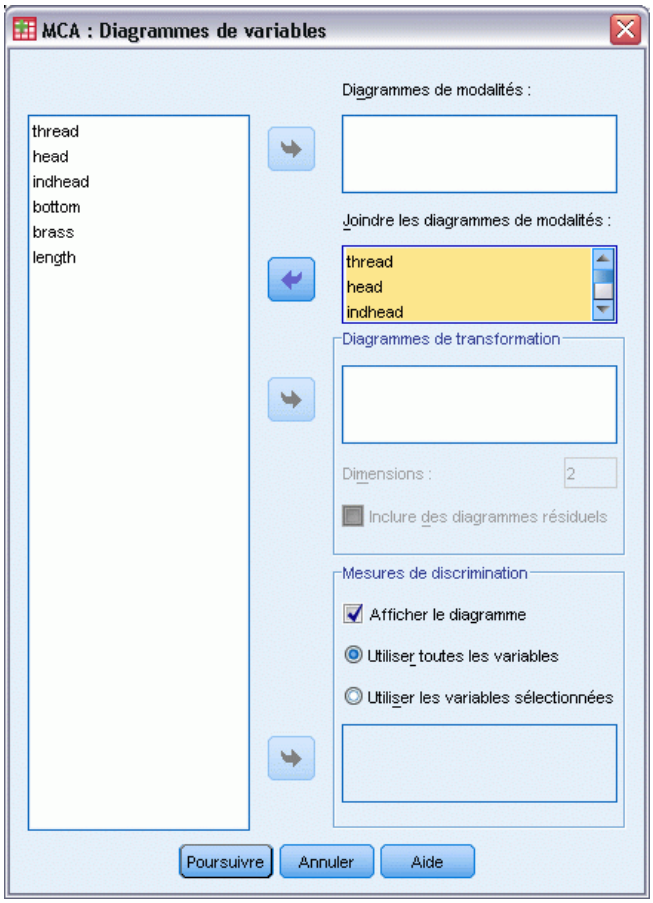
- Sélectionnez *Filetage* comme variable d'analyse via *Longueur en demi-pouces*.
- Sélectionnez *objet* comme variable d'étiquetage.
- Dans le groupe Diagrammes, cliquez sur Objet.

Figure 13-3
Boîte de dialogue Diagrammes d'objets



- Choisissez l'option d'étiquetage des objets Variable.
- Sélectionnez les variables d'étiquetage *filetage* à *objet*.
- Cliquez sur Continuer, puis sur Variable dans le groupe Diagrammes de la boîte de dialogue Analyse de correspondance multiple.

Figure 13-4
Boîte de dialogue Diagrammes de variables



- Appliquez l'opération Joindre les diagrammes de modalités aux options allant de *filetage* jusqu'à *longueur*.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Analyse des correspondances multiples.

Récapitulatif des modèles

L'analyse d'homogénéité peut calculer une solution pour plusieurs dimensions. Le nombre maximal de dimensions est égal soit au nombre de modalités moins le nombre de variables n'ayant aucune donnée manquante, soit au nombre d'observations moins 1, selon le nombre qui est le plus petit. N'utilisez toutefois que rarement le nombre maximal de dimensions. Un nombre de dimensions plus petit est plus facile à interpréter et, après un certain nombre de dimensions, le total de l'association supplémentaire représentée devient négligeable. Une solution à une, deux ou trois dimensions dans une analyse d'homogénéité est chose courante.

Figure 13-5
Récapitulatif du modèle

Dimension	Alpha de Cronbach	Variance expliquée		
		Total (valeur propre)	Inertie	Pourcentage de variance expliquée
1	,878	3,727	,621	62,123
2	,657	2,209	,368	36,809
Total		5,936	,989	
Moyenne	,796[a]	2,968	,495	49,466

a. La valeur Alpha de Cronbach moyenne est basée sur la valeur propre moyenne.

Presque toute la variance des données est représentée par la solution : 62,1 % par la première dimension et 36,8 % par la deuxième.

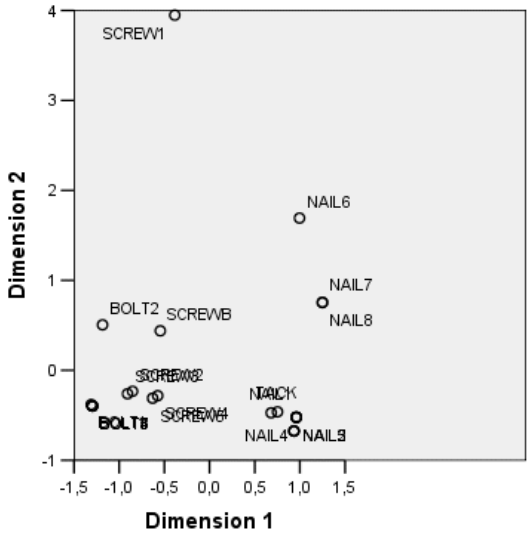
Les deux dimensions offrent une interprétation en matière de distances. Si une variable a un fort pouvoir discriminant, les objets seront proches des modalités auxquelles ils appartiennent. Idéalement, les objets de la même modalité seront proches les uns des autres (ils auront des coordonnées similaires) et les modalités de variables différentes seront proches si elles appartiennent aux mêmes objets (deux objets qui ont des coordonnées similaires pour une variable doivent également être proches l'un de l'autre pour les autres variables de la solution).

Coordonnées principales

Après avoir analysé le récapitulatif des modèles, vérifiez les coordonnées des objets. Vous pouvez indiquer une ou plusieurs variables pour étiqueter le diagramme de coordonnées des objets. Chaque variable d'étiquetage génère un diagramme distinct étiqueté avec les valeurs de la variable. Nous vérifierons le diagramme des coordonnées d'objets étiqueté à l'aide de l'objet de variable. Il s'agit simplement d'une variable d'identification des observations qui n'a été utilisée dans aucun calcul.

La distance séparant un objet de l'origine reflète la variation du modèle de réponse « moyenne ». Ce modèle de réponse moyenne correspond à la modalité la plus fréquente de chaque variable. Les objets dont de nombreuses descriptives correspondent aux modalités les plus fréquentes se trouvent à côté de l'origine. A l'inverse, les objets qui disposent de descriptives uniques sont loin de l'origine.

Figure 13-6
Diagramme de coordonnées des objets étiquetées avec la variable objet



Normalisation principale de la variable.

Si vous observez le diagramme, vous constatez que la première dimension (l'axe horizontal) distingue les vis et boulons (qui ont des filetages) des clous et broquettes (qui n'ont pas de filetage). En effet, les vis et les boulons se trouvent à une extrémité de l'axe horizontal alors que les clous et les broquettes sont à l'autre extrémité. Dans une moindre mesure, la première dimension sépare également les boulons (qui ont un fond plat) de tous les autres objets (qui ont un fond pointu).

La deuxième dimension (l'axe vertical) semble séparer *VIS1* et *CLOU6* de tous les autres objets. *VIS1* et *CLOU6* partagent des valeurs identiques en ce qui concerne la longueur des variables (ce sont les objets les plus longs des données). De plus, *VIS1* est beaucoup plus loin de l'origine que les autres objets, ce qui laisse supposer que, dans l'ensemble, de nombreuses descriptives de cet objet ne sont pas partagées par les autres objets.

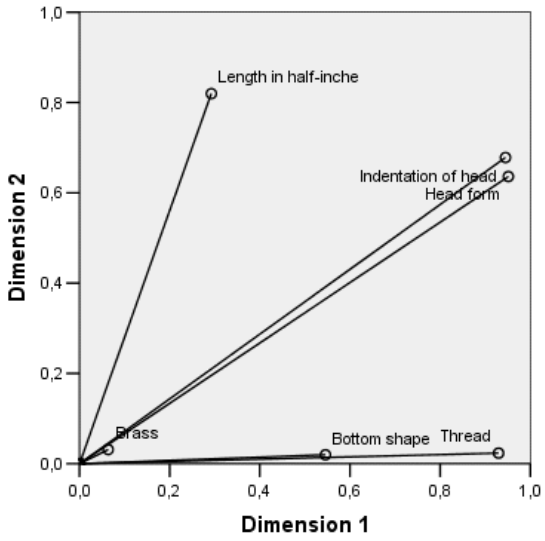
Le diagramme de coordonnées des objets est plus particulièrement utile pour rechercher les valeurs éloignées. La variable *VIS1* peut être considérée comme une valeur éloignée. Nous étudierons ultérieurement ce qu'il advient si vous supprimez cet objet.

Mesures de discrimination

Avant d'étudier le reste des diagrammes de coordonnées des objets, vérifions si les mesures de discrimination sont conformes aux propos précédents. En ce qui concerne les variables, une mesure de discrimination, pouvant être considérée comme une corrélation entre composantes, est calculée pour chaque dimension. Cette mesure est également la variance de la variable quantifiée de cette dimension. La valeur maximale 1 est atteinte si les coordonnées d'objet font partie de groupes mutuellement exclusifs et si toutes les coordonnées d'objets d'une modalité sont identiques. (*Remarque* : La valeur de cette mesure peut être supérieure à 1 si des données sont manquantes.) Des mesures de discrimination importantes correspondent à une répartition étendue parmi les modalités de la variable et indiquent par conséquent un degré de discrimination élevé entre les modalités d'une variable le long de la dimension concernée.

La moyenne des mesures de discrimination d'une dimension est égale au pourcentage de variance indiqué pour cette dimension. Par conséquent, les dimensions sont triées en fonction de la discrimination moyenne. La première dimension dispose de la discrimination moyenne la plus élevée, la deuxième dimension dispose de la deuxième discrimination moyenne la plus élevée, et ainsi de suite pour toutes les dimensions de la solution.

Figure 13-7
Diagramme des mesures de discrimination



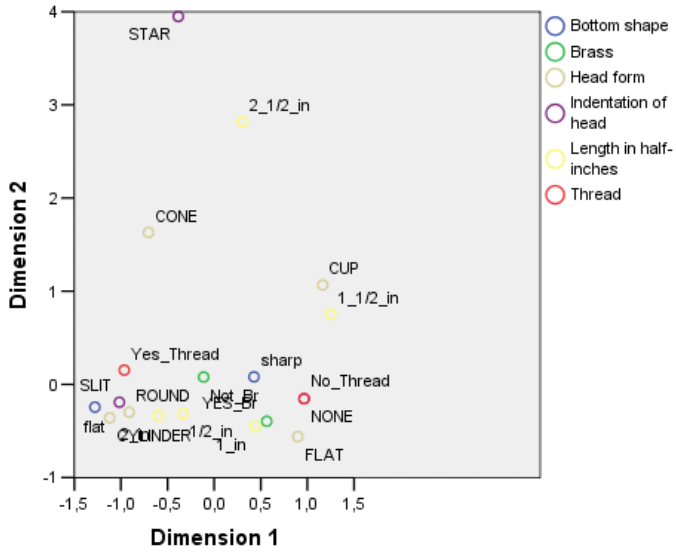
Comme le diagramme de coordonnées des objets, le diagramme des mesures de discrimination indique que la première dimension est liée aux variables *Filetage* et *Forme tige*. Ces variables disposent de mesures de discrimination élevées sur la première dimension et de mesures de discrimination limitées sur la deuxième. Par conséquent, pour ces deux variables, les modalités sont éloignées les unes des autres le long de la première dimension uniquement. La valeur de la variable *Longueur en demi-pouces* est élevée sur la deuxième dimension, mais faible sur la première. La *longueur* est donc l'objet le plus proche de la deuxième dimension. Conformément à l'observation du diagramme de coordonnées des objets, la deuxième dimension semble séparer les objets les plus longs des autres objets. Les valeurs des variables *Indentation de la tête* et *Forme de tête* sont relativement élevées sur les deux dimensions, ce qui indique une discrimination dans les deux premières dimensions. La variable *Cuivre*, très proche de l'origine, ne fait aucune distinction dans les deux premières dimensions. Ceci est logique étant donné que tous les objets peuvent être en cuivre ou dans un autre matériau.

Valeurs affectées aux modalités

Souvenez-vous qu'une mesure de discrimination est la variance de la variable quantifiée le long d'une dimension particulière. Le diagramme des mesures de discrimination contient ces variances et indique ainsi les variables discriminantes le long de la dimension concernée. Cependant, une variance peut correspondre à toutes les modalités modérément éloignées les unes des autres ou à la plupart des modalités proches les unes des autres, avec quelques modalités différant de ce groupe. Le diagramme de discrimination ne peut faire aucune distinction entre ces deux conditions.

Les diagrammes de valeurs affectées aux modalités offrent un autre mode d'affichage de la discrimination des variables qui peut identifier les relations entre les modalités. Dans ce diagramme, les coordonnées des modalités de chaque dimension sont affichées. Vous pouvez donc déterminer les modalités similaires pour chaque variable.

Figure 13-8
Valeurs affectées aux modalités



La variable *Longueur en demi-pouces* compte cinq modalités, dont trois sont regroupées près de la partie supérieure du diagramme. Les deux autres modalités se trouvent dans la moitié inférieure du diagramme, la modalité *2_1/2_in* se trouvant très loin du groupe. La discrimination élevée de longueur le long de la dimension 2 est due à cette modalité qui est très différente des autres modalités de longueur. De la même façon, pour la variable *Forme de tête*, la modalité *CRUCIFORME* est très loin des autres modalités et génère une mesure de discrimination élevée le long de la deuxième dimension. Il est impossible d'illustrer ces modèles dans un diagramme de mesures de discrimination.

La répartition des valeurs affectées aux modalités d'une variable reflète la variance et indique le degré élevé de discrimination de cette variable dans chaque dimension. En ce qui concerne la dimension 1, les modalités de la variable *Filetage* sont éloignées les unes des autres. Cependant, le long de la dimension 2, les modalités de cette variable sont très proches les unes des autres. Par conséquent, le degré de discrimination de la variable *Filetage* est plus élevé dans la dimension 1 que dans la dimension 2. En revanche, les modalités de la variable *Forme de tête* sont éloignées les unes des autres le long des deux dimensions, ce qui laisse supposer que le degré de discrimination de cette variable est élevé dans les deux dimensions.

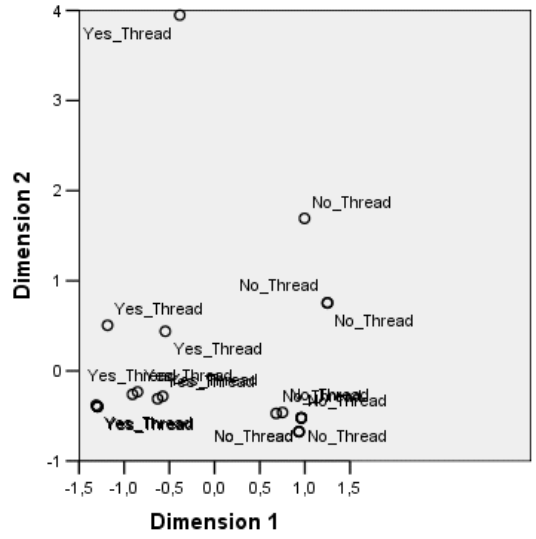
Non seulement le diagramme de valeurs affectées aux modalités détermine le mode de discrimination et les dimensions le long desquelles une variable a un pouvoir discriminant, mais il compare également la discrimination des variables. Une variable ayant des modalités éloignées les unes des autres a un pouvoir discriminant plus élevé qu'une variable comportant des modalités proches les unes des autres. Par exemple, le long de la dimension 1, les deux modalités de la variable *Cuivre* sont plus proches l'une de l'autre que les deux modalités de la variable *Filetage*, ce qui indique que la variable *Filetage* a un pouvoir discriminant plus élevé que la variable *Cuivre*

le long de cette dimension. Cependant, le long de la dimension 2, les distances sont très similaires, ce qui laisse supposer que ces variables ont un pouvoir discriminant identique le long de cette dimension. Le diagramme des mesures de discrimination abordé ci-dessus identifie les mêmes relations à l'aide de variances reflétant la répartition des modalités.

Etude plus détaillée des coordonnées des objets

L'étude des diagrammes de coordonnées des objets étiquetées avec chaque variable offre un meilleur éclairage des données. Idéalement, les objets similaires doivent former des groupes exclusifs, ces groupes devant être éloignés les uns des autres.

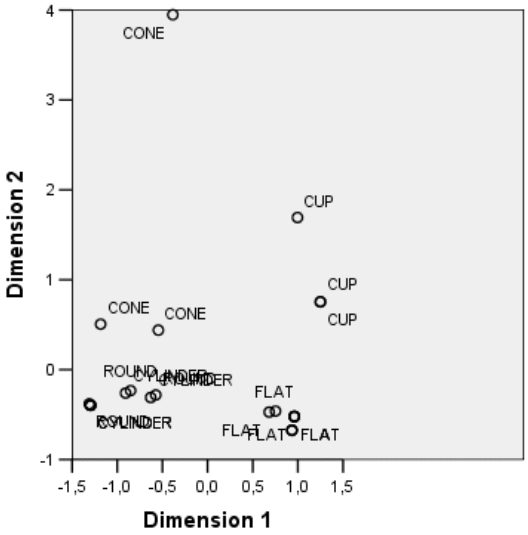
Figure 13-9
Coordonnées des objets étiquetées avec la variable Filetage



Normalisation principale de la variable.

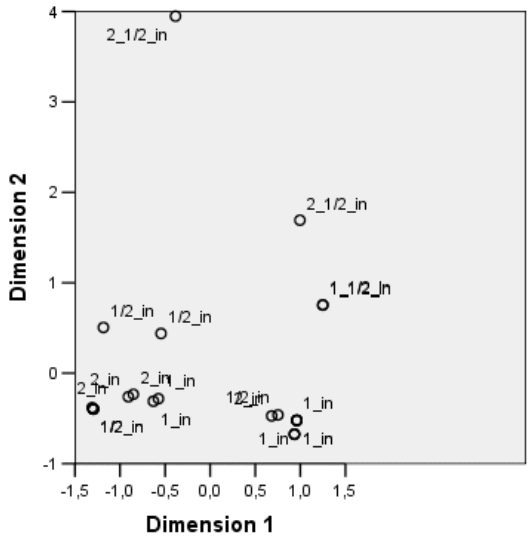
Le diagramme étiqueté avec la variable *Filetage* indique que la première dimension sépare parfaitement *Yes_Thread* et *No_Thread*. Tous les objets comportant des filetages ont des coordonnées d'objet négatives, alors que tous les objets sans filetage ont des coordonnées d'objet positives. Bien que les deux modalités ne forment pas des groupes compacts, la différenciation parfaite entre ces modalités est généralement considérée comme un bon résultat.

Figure 13-10
Coordonnées des objets étiquetées avec la variable Forme de tête



Le diagramme étiqueté avec la variable *Forme de tête* indique que cette variable a un pouvoir discriminant élevé dans les deux dimensions. Les objets *PLATE* sont regroupés dans le coin inférieur droit du diagramme, tandis que les objets *CREUSE* sont regroupés dans le coin supérieur droit. Tous les objets *CONIQUE* se trouvent dans le coin supérieur gauche. Ces objets sont cependant plus éloignés les uns des autres que les autres groupes et ne sont donc pas considérés comme étant homogènes. Enfin, les objets *CYLINDRIQUES* ne peuvent pas être séparés des objets *ARRONDIS*. Tous ces objets se trouvent dans le coin inférieur gauche du diagramme.

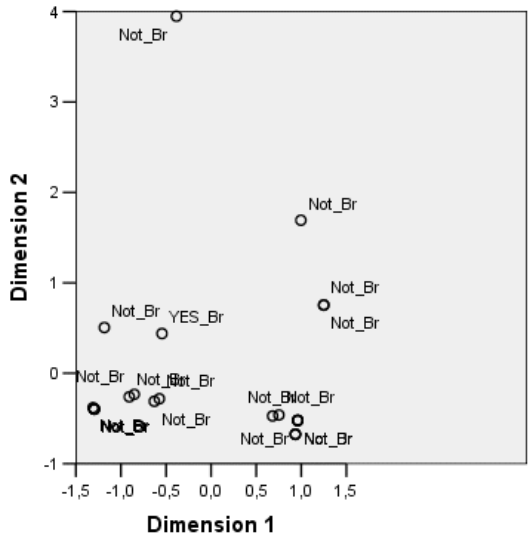
Figure 13-11
Coordonnées des objets étiquetées avec Longueur en demi-pouces



Normalisation principale de la variable.

Le diagramme étiqueté avec la variable *Longueur en demi-pouces* indique que cette variable n'a aucun pouvoir discriminant dans la première dimension. Ses modalités n'indiquent aucun regroupement lorsqu'elles sont projetées sur une ligne horizontale. Cependant, la variable *Longueur en demi-pouces* a un pouvoir discriminant dans la deuxième dimension. Les objets les plus courts correspondent aux coordonnées positives et les objets les plus longs, aux coordonnées négatives.

Figure 13-12
Coordonnées des objets étiquetées avec la variable Cuivre



Normalisation principale de la variable.

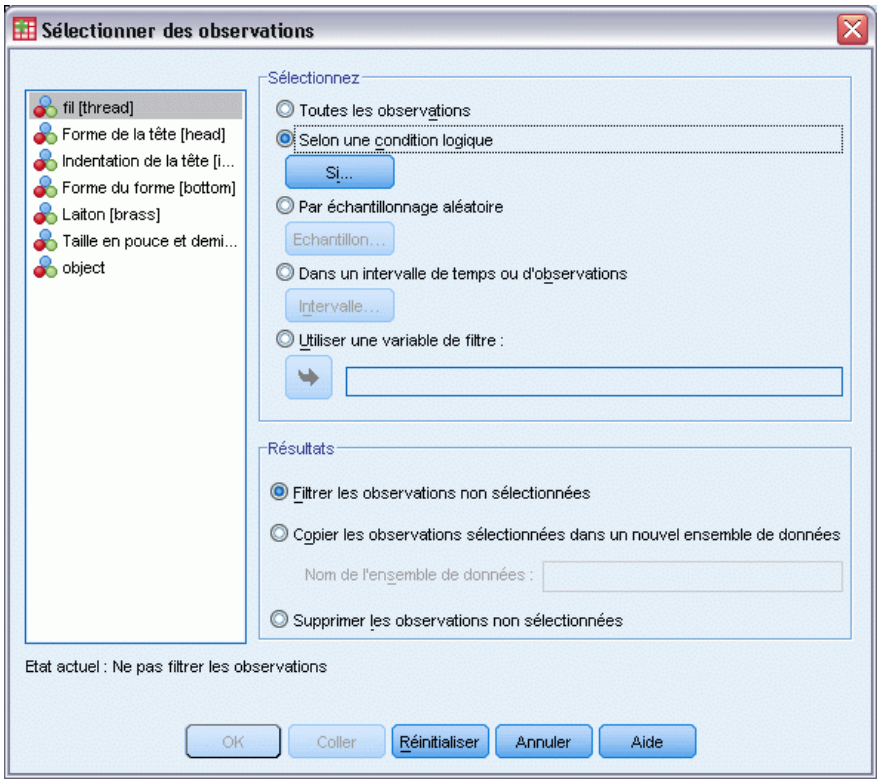
Le diagramme étiqueté avec la variable *Cuivre* indique que cette variable dispose de modalités dont la séparation n'est pas aisée dans la première ou la deuxième dimension. Les coordonnées des objets sont fortement éloignées les unes des autres. Il est impossible de différencier les objets en cuivre des objets qui ne sont pas en cuivre.

Omission des valeurs éloignées

Dans une analyse d'homogénéité, les valeurs éloignées sont des objets qui ont trop de fonctionnalités spécifiques. Comme nous l'avons déjà indiqué, la variable *VIS1* peut être considérée comme une valeur éloignée.

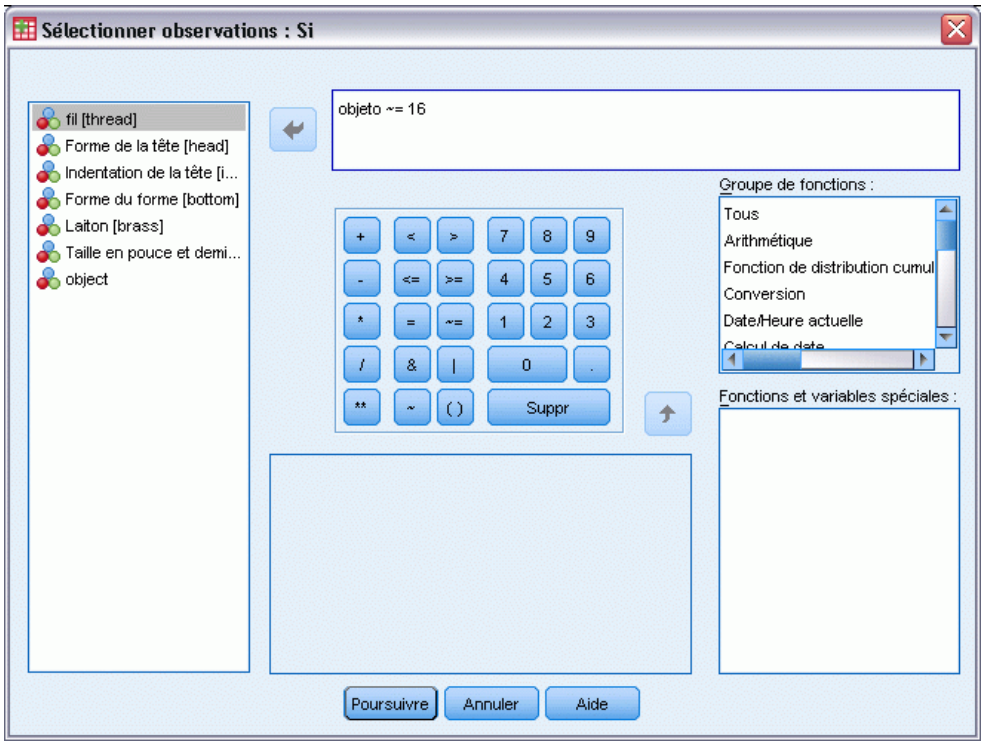
Pour supprimer cet objet et réexécuter l'analyse, à partir des menus, sélectionnez :
Données > Sélectionner des observations

Figure 13-13
Boîte de dialogue Sélectionner des observations



- Sélectionnez Selon une condition logique.
- Cliquez sur Si.

Figure 13-14
Si la boîte de dialogue



- Entrez objet ~= 16 comme condition.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Sélectionner des observations.
- Affichez à nouveau la boîte de dialogue Analyse des correspondances multiples, puis cliquez sur OK.

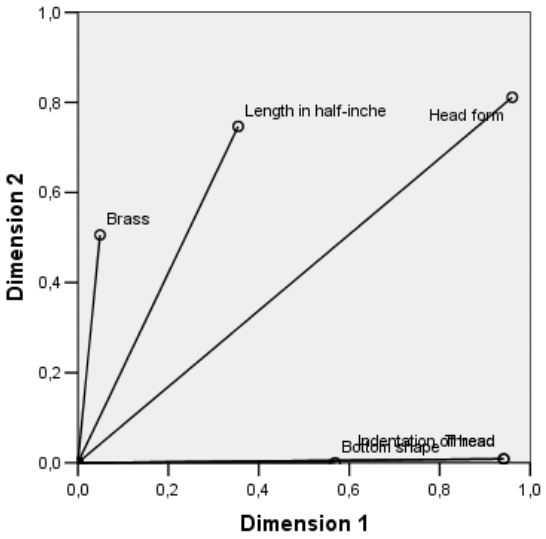
Figure 13-15
Récapitulatif des modèles (valeur éloignée supprimée)

Dimension	Alpha de Cronbach	Variance expliquée		
		Total (valeur propre)	Inertie	Pourcentage de variance expliquée
1	,885	3,815	,636	63,591
2	,623	2,081	,347	34,676
Total		5,896	,983	
Moyenne	,793[a]	2,948	,491	49,133

a. La valeur Alpha de Cronbach moyenne est basée sur la valeur propre moyenne.

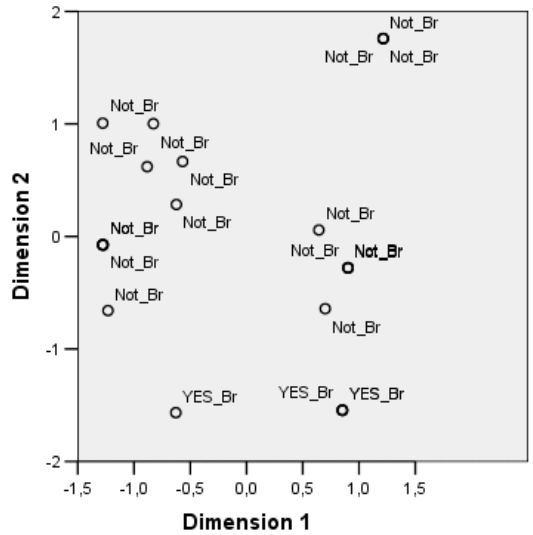
Les valeurs propres changent légèrement. La première dimension représente maintenant une plus grande partie de la variance.

Figure 13-16
Mesures de discrimination



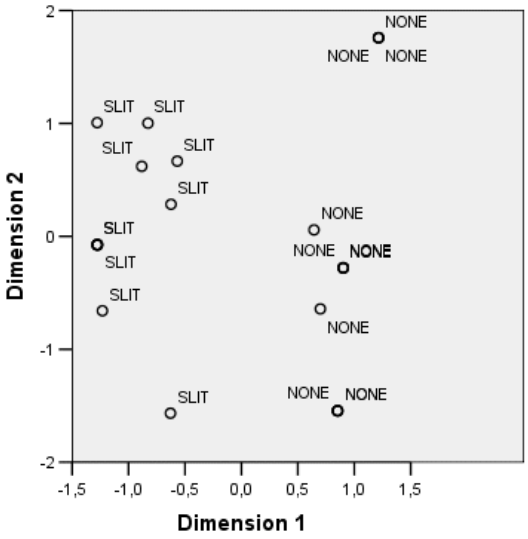
Comme l'indique le diagramme de discrimination, la variable *Indentation de la tête* n'a plus de pouvoir discriminant dans la deuxième dimension, alors que la variable *Cuivre*, qui n'avait aucun pouvoir discriminant, a maintenant un pouvoir discriminant dans la deuxième dimension. La discrimination des autres variables ne change quasiment pas.

Figure 13-17
Coordonnées des objets étiquetées avec la variable Cuivre (valeur éloignée supprimée)



Le diagramme de coordonnées des objets étiquetées avec la variable *Cuivre* indique que les quatre objets en cuivre se situent à proximité de la partie inférieure du diagramme (trois objets se trouvent au même endroit). Par conséquent, la discrimination est élevée le long de la deuxième dimension. Comme pour la variable *Filetage* dans l'analyse précédente, les objets ne forment pas des groupes compacts, mais la différenciation de ces objets par modalité est parfaite.

Figure 13-18
Coordonnées des objets étiquetées avec la variable Indentation de la tête (valeur éloignée supprimée)



Le diagramme de coordonnées des objets étiquetées avec la variable *Indentation de la tête* indique que la première dimension distingue parfaitement les objets non indentés et les objets indentés, comme dans l'analyse précédente. Cependant, par rapport à l'analyse précédente, la deuxième dimension ne peut plus distinguer les deux modalités.

De ce fait, l'omission de *VIS1*, qui est le seul objet ayant une tête en étoile, a une incidence considérable sur l'interprétation de la deuxième dimension. Cette dimension différencie maintenant les objets en fonction des variables *Cuivre*, *Forme de tête* et *Longueur en demi-pouces*.

Lectures recommandées

Pour plus d'informations sur l'analyse de correspondance multiple, reportez-vous aux documents suivants :

Benzécri, J. P. 1992. *Correspondence analysis handbook*. New York: Marcel Dekker.

Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. Dans : *The Prediction of Personal Adjustment,* P. Horst, éd. New York: Social Science Research Council.

Meulman, J. J. 1982. *Homogeneity analysis of incomplete data*. Leiden: DSWO Press.

Meulman, J. J. 1996. Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification*, 13, .

Meulman, J. J., et W. J. Heiser. 1997. Graphical display of interaction in multiway contingency tables by use of homogeneity analysis. Dans : *Visual Display of Categorical Data,* M. Greenacre, et J. Blasius, éds. New York: Academic Press.

Nishisato, S. 1984. Forced classification: A simple application of a quantification method. *Psychometrika*, 49, .

Tenenhaus, M., et F. W. Young. 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, .

Van Rijckevorsel, J. 1987. *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. Leiden: DSWO Press.

Chapitre
14

Positionnement multidimensionnel

Le positionnement multidimensionnel vise à rechercher une représentation d'un ensemble d'objets donné dans un espace de petite dimension. Vous pouvez obtenir cette solution en utilisant des **proximités** entre les objets. La procédure réduit au minimum les carrés des écarts entre l'objet initial, éventuellement transformé, les proximités des objets et leurs distances euclidiennes dans l'espace de petite dimension.

La finalité de l'espace de petite dimension est de mettre en évidence les relations entre les objets. En réduisant la solution à une combinaison linéaire de variables indépendantes, vous pouvez interpréter les dimensions de la solution par rapport à ces variables. L'exemple suivant montre comment représenter 15 termes de parenté différents dans trois dimensions et interpréter l'espace par rapport au sexe, à la génération et au degré de séparation de chacun des termes.

Exemple \: Examen des termes de parenté

Rosenberg et Kim (Rosenberg et Kim, 1975) se sont lancés dans l'analyse de 15 termes de parenté (cousin/cousine, fille, fils, frère, grand-mère, grand-père, mère, neveu, nièce, oncle, père, petite-fille, petit-fils, sœur, tante). Ils ont demandé à quatre groupes d'étudiants (deux groupes de femmes et deux groupes d'hommes) de trier ces termes en fonction des similarités. Deux groupes (un groupe de femmes et un groupe d'hommes) ont été invités à effectuer deux tris, en basant le second sur un autre critère que le premier. Par conséquent, un total de six "sources" a été obtenu, comme le montre le tableau ci-après.

Table 14-1
Structure des sources des données de parenté

Source	sexe	Condition	Taille de l'échantillon
1	Groupe de femmes	Tri unique	85
2	Groupe d'hommes	Tri unique	85
3	Groupe de femmes	Premier tri	80
4	Groupe de femmes	Second tri	80
5	Groupe d'hommes	Premier tri	80
6	Groupe d'hommes	Second tri	80

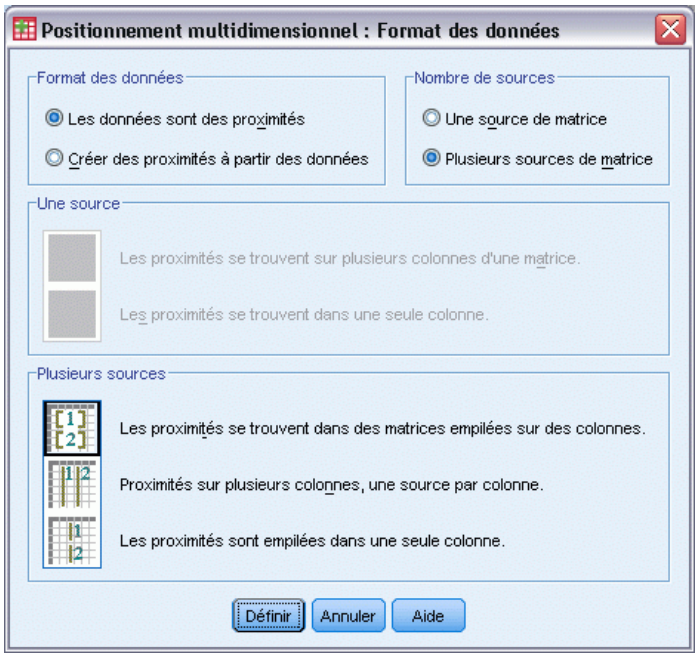
Chaque source correspond à une matrice de proximité 15×15, dont le nombre de cellules est égal au nombre de personnes dans une source moins le nombre de fois où les objets ont été partitionnés dans cette source. Cet ensemble de données est disponible dans le fichier *kinship_dat.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*.

Choix du nombre de dimensions

Il vous appartient de choisir le nombre de dimensions à attribuer à la solution. Le diagramme de valeurs propres peut vous aider à prendre cette décision.

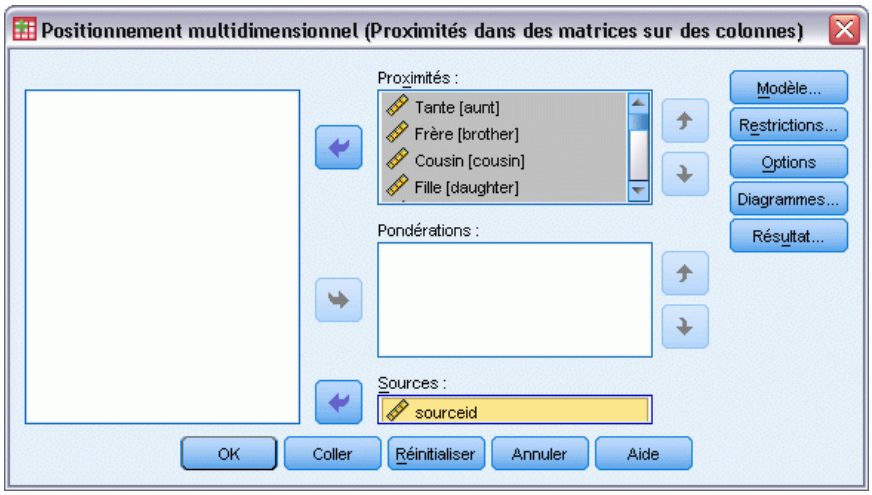
- Pour créer un graphique des valeurs propres, à partir des menus, sélectionnez :
  Analyse > Echelle > Positionnement multidimensionnel (PROXSCAL)

Figure 14-1
Boîte de dialogue Format des données



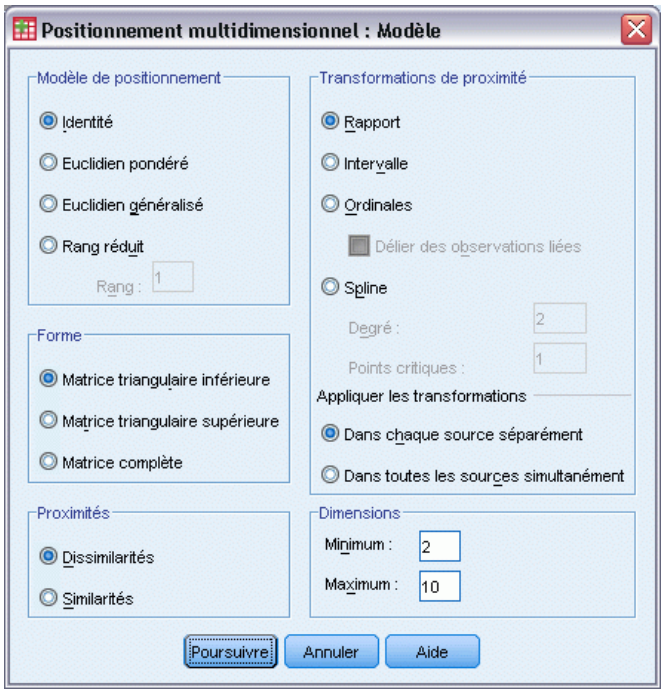
- Sélectionnez l'option Plusieurs sources de matrice dans le groupe Nombre de sources.
- Cliquez sur Définir.

Figure 14-2
Boîte de dialogue Positionnement multidimensionnel



- Sélectionnez les options allant de *Tante* à *Oncle* comme variables de proximités.
- Sélectionnez l'option *idsource* comme variable d'identification de la source.
- Cliquez sur Modèle.

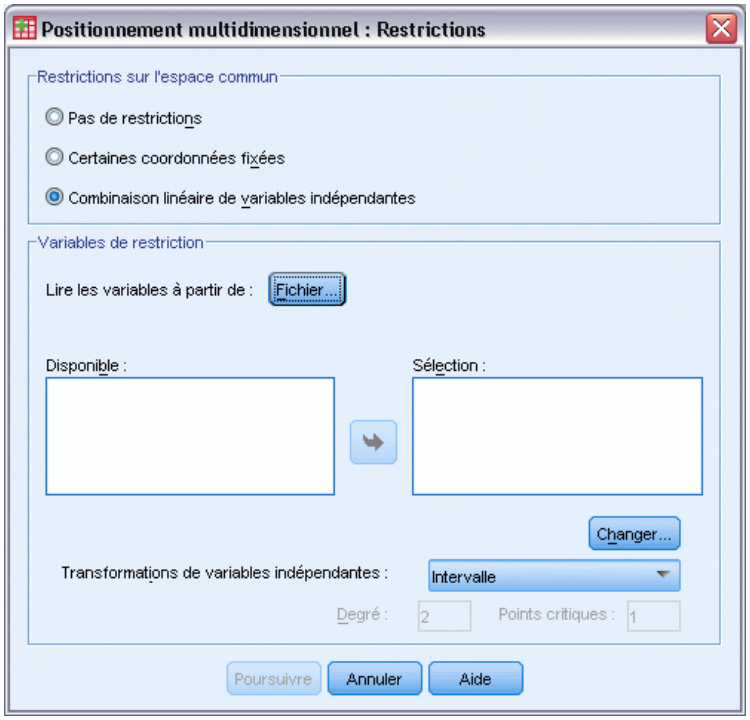
Figure 14-3
Boîte de dialogue Modèle



- Tapez 10 comme nombre maximum de dimensions.
- Cliquez sur Poursuivre.

▶ Cliquez sur Restrictions dans la boîte de dialogue Positionnement multidimensionnel.

Figure 14-4
Boîte de dialogue Restrictions

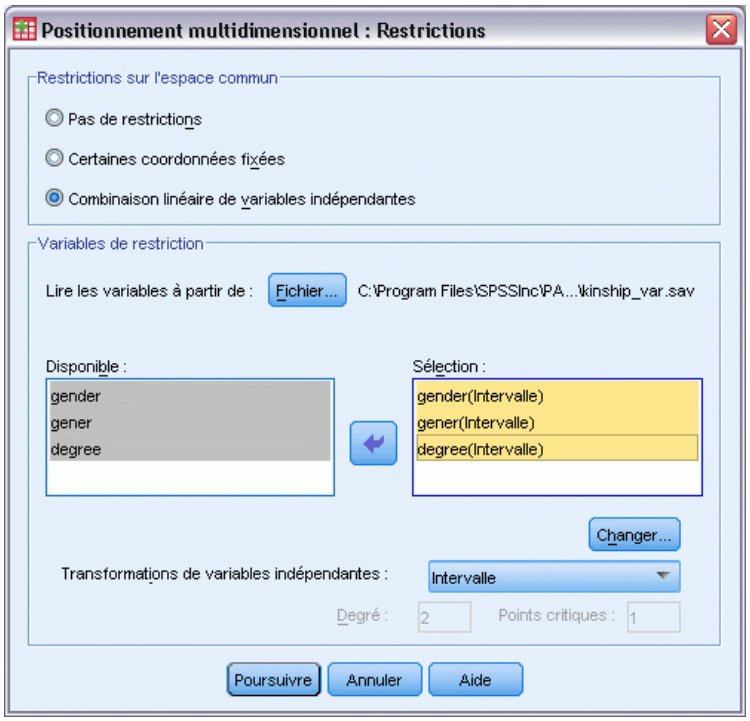


▶ Sélectionnez Combinaison linéaire de variables indépendantes.

▶ Cliquez sur Fichier pour sélectionner la source des variables indépendantes.

▶ Sélectionnez *kinship_var.sav*.

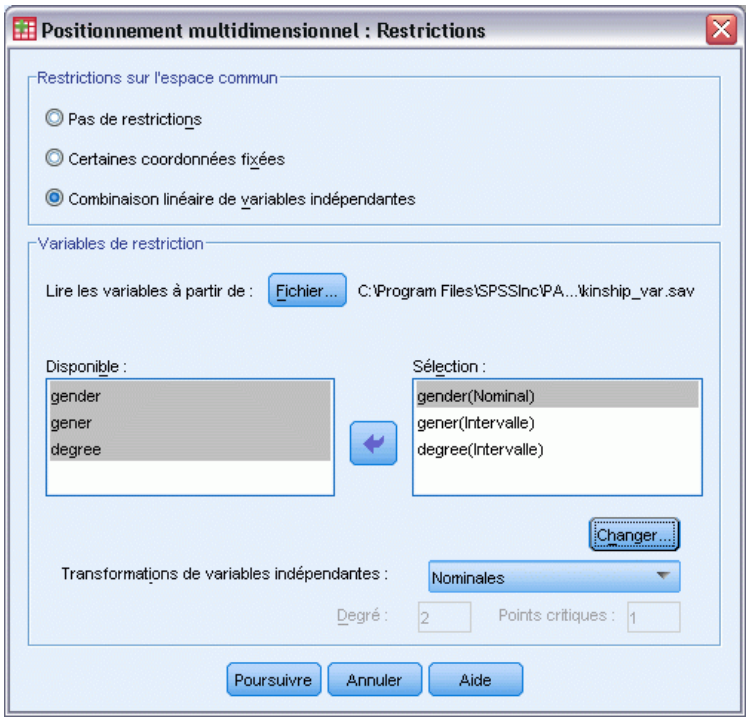
Figure 14-5
Boîte de dialogue Restrictions



▶ Sélectionnez les options *sexe*, *sexe* et *degré* comme variables de restriction.

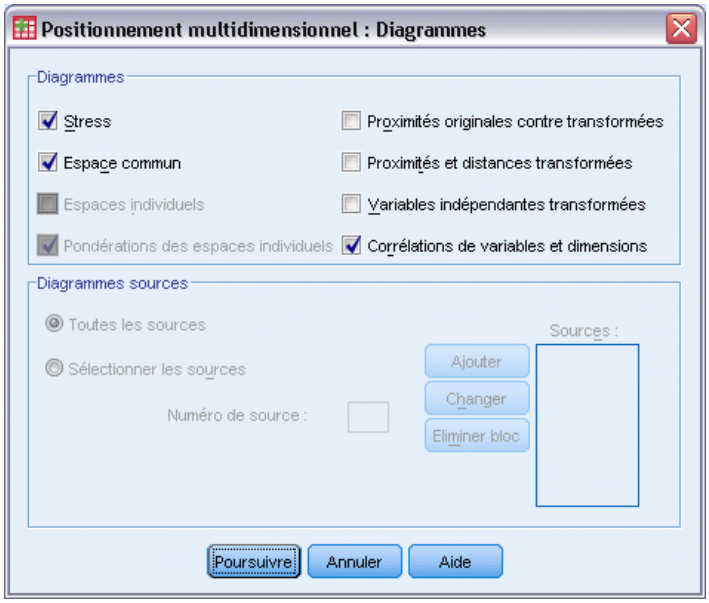
La variable *sexe* possède une valeur manquante définie par l'utilisateur—il s'agit de la valeur 9, pour le lien de parenté cousin. La procédure la traite comme une modalité valide. La transformation linéaire par défaut a donc peu de chance d'être appropriée. Utilisez plutôt une transformation nominale.

Figure 14-6
Boîte de dialogue Restrictions



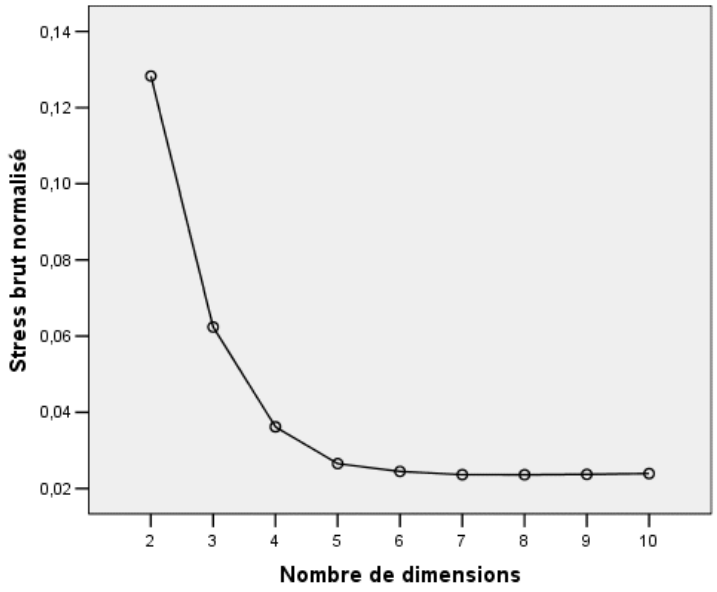
- Sélectionnez *sexe*.
- Sélectionnez l'option Nominal dans la liste déroulante Transformations des variables indépendantes.
- Cliquez sur Changer.
- Cliquez sur Poursuivre.
- Cliquez sur Diagrammes dans la boîte de dialogue Positionnement multidimensionnel.

Figure 14-7
Boîte de dialogue Diagrammes



- Sélectionnez l'option Stress dans le groupe Diagrammes.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Positionnement multidimensionnel.

Figure 14-8
Diagramme des valeurs propres



La procédure commence avec une solution à 10 dimensions et progresse jusqu'à une solution à 2 dimensions. Le graphique des valeurs propres montre le stress brut normalisé de la solution à chaque dimension. Vous pouvez constater d'après le diagramme que l'augmentation du nombre de dimensions de 2 à 3 et de 3 à 4 améliore sensiblement le stress. Au-delà de 4 dimensions, les améliorations sont assez réduites. Vous opterez pour l'analyse des données à l'aide d'une solution à 3 dimensions, dans la mesure où les résultats sont plus faciles à interpréter.

Solution tridimensionnelle

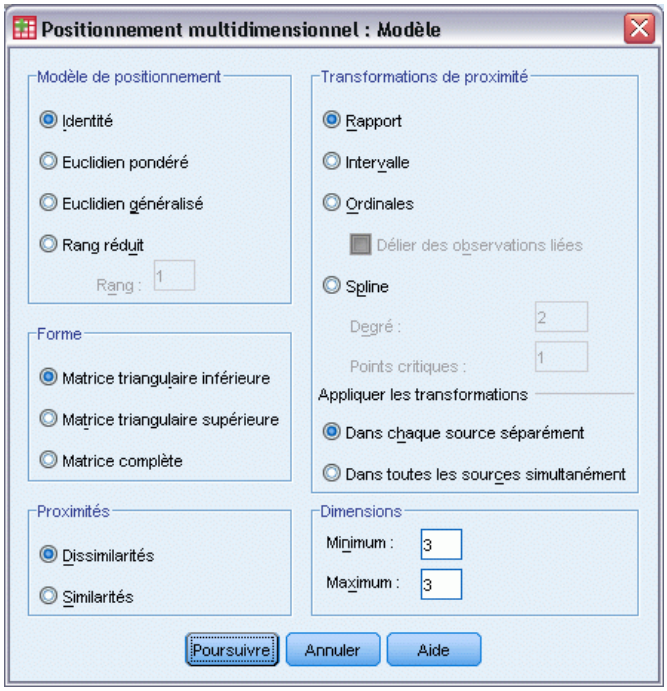
Les variables indépendantes *sexe*, *génér* (génération) et *degré* (degré de séparation) ont été construites en vue de leur utilisation pour interpréter les dimensions de la solution. Les variables indépendantes ont été élaborées comme suit :

sexe	1 = masculin, 2 = féminin, 9 = manquant, pour le lien de parenté cousin.
génér	Nombre de générations par rapport à vous si le terme fait référence à votre famille ; ce nombre est d'autant plus faible que la génération est éloignée. Ainsi, les grands-parents ont la valeur –2, les petits-enfants la valeur 2 et les frères ou soeurs la valeur 0.
degré	Nombre de degrés de séparation le long de votre arbre généalogique. Ainsi, par rapport à vous, vos parents se trouvent un noeud au-dessus, et vos enfants un noeud au-dessous. Pour atteindre vos frères/soeurs, vous devez remonter d'un noeud jusqu'à vos parents, puis descendre d'un noeud jusqu'à vos frères/soeurs, ce qui représente 2 degrés de séparation. Quatre degrés vous séparent de vos cousins/cousines —deux jusqu'à vos grands-parents, puis deux jusqu'à eux en passant par votre tante/oncle.

Les variables externes sont disponibles dans le fichier *kinship_var.sav*. En outre, une configuration initiale à partir d'une analyse antérieure est disponible dans le fichier *kinship_ini.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*.

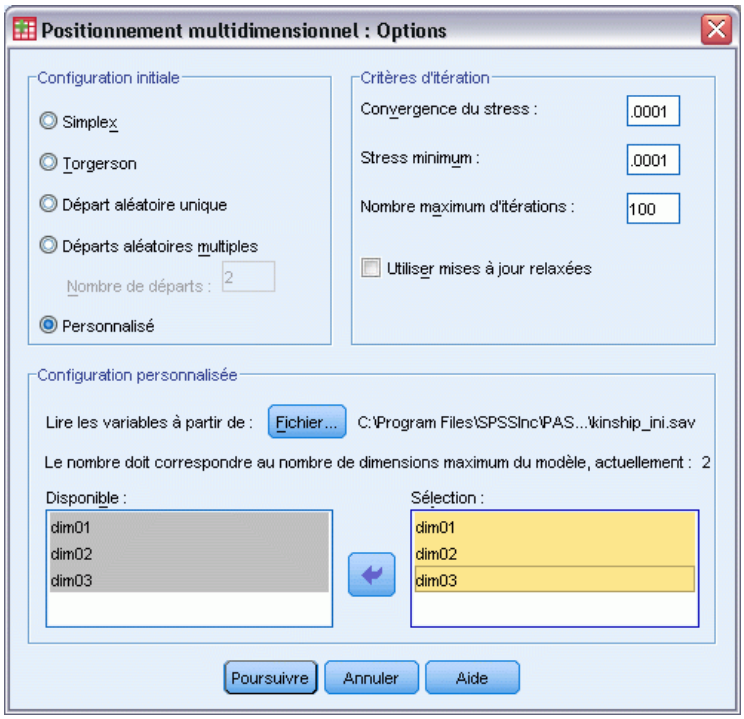
Exécution de l'analyse

Figure 14-9
Boîte de dialogue Modèle



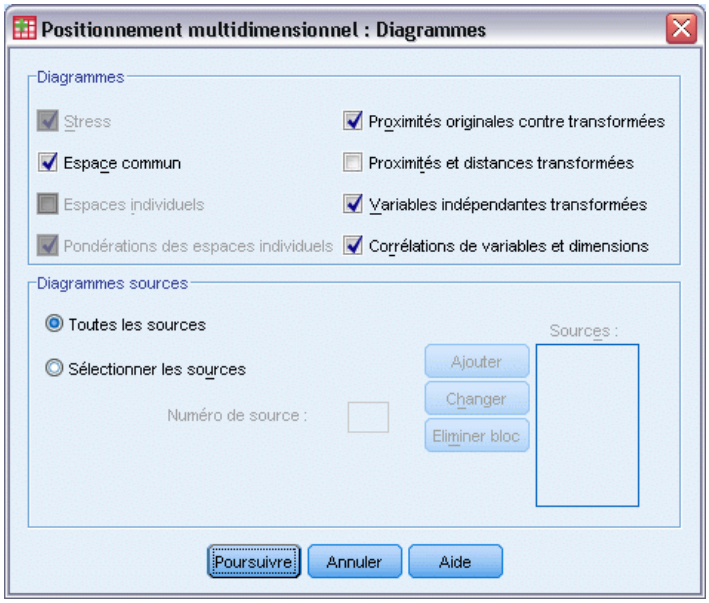
- Pour obtenir une solution tridimensionnelle, affichez à nouveau la boîte de dialogue Positionnement multidimensionnel, puis cliquez sur Modèle.
- Tapez 3 comme nombres minimum et maximum de dimensions.
- Cliquez sur Poursuivre.
- Cliquez sur Options dans la boîte de dialogue Positionnement multidimensionnel.

Figure 14-10
Options



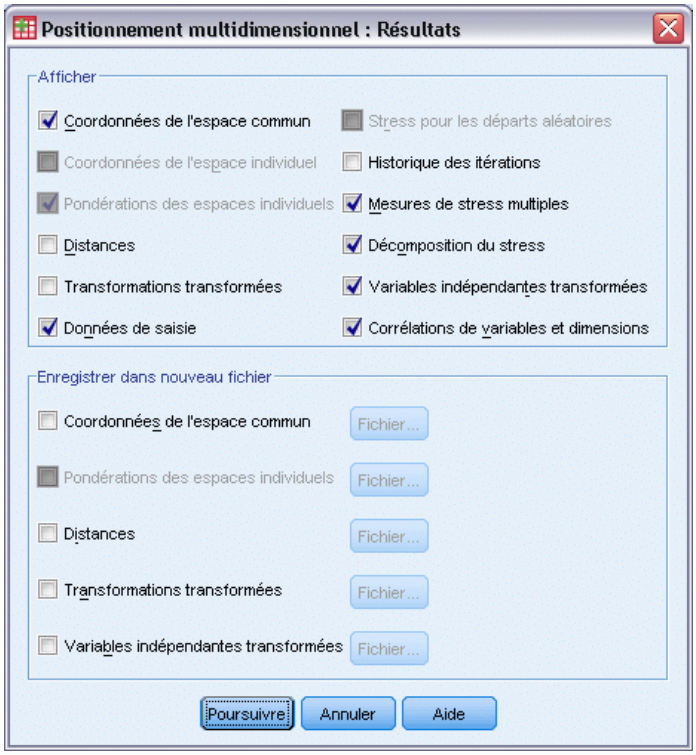
- Sélectionnez Personnalisée pour la configuration initiale.
- Sélectionnez *kinship_ini.sav* comme fichier contenant les variables à lire.
- Sélectionnez les options *dim01*, *dim02* et *dim03* comme variables.
- Cliquez sur Poursuivre.
- Cliquez sur Diagrammes dans la boîte de dialogue Positionnement multidimensionnel.

Figure 14-11
Boîte de dialogue Diagrammes



- Sélectionnez les options Proximités originales et transformées et Variables explicatives transformées.
- Cliquez sur Poursuivre.
- Cliquez sur Résultat dans la boîte de dialogue Positionnement multidimensionnel.

Figure 14-12
Résultat



- Sélectionnez les options Données d'entrée, Décomposition du stress et Corrélation des variables et dimensions.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Positionnement multidimensionnel.

Mesures de stress

Les mesures de stress et d'ajustement donnent une indication du degré d'éloignement entre les distances de la solution et les distances initiales.

Figure 14-13
Mesures de stress et d'ajustement

Mesures de stress et d'ajustement

Stress brut normalisé	,06234
Stress I	,24968[a]
Stress II	,87849[a]
Stress S	,14716[b]
Dispersion expliquée par	,93766
Coefficient de congruence de Tucker	,96833

PROXSCAL minimise le stress brut normalisé.

a. Facteur de codage optimal = 1,066.

b. Facteur de codage optimal = ,984.

Chacune des quatre statistiques de stress mesure le non-ajustement des données, tandis que la dispersion représentée et le coefficient de Tucker de congruence mesurent l'ajustement. Les mesures de stress faibles (jusqu'à un minimum de 0) et les mesures d'ajustement élevées (jusqu'à un maximum de 1) indiquent de bonnes solutions.

Figure 14-14
Décomposition du stress ligne normalisé

		Source						
		SRC_1	SRC_2	SRC_3	SRC_4	SRC_5	SRC_6	Moyenne
Objet	Tante	,0991	,0754	,0629	,0468	,0391	,0489	,0620
	Frère	,1351	,0974	,0496	,0813	,0613	,0597	,0807
	Cousin	,0325	,0336	,0480	,0290	,0327	,0463	,0370
	Fille	,0700	,0370	,0516	,0229	,0326	,0207	,0391
	Père	,0751	,0482	,0521	,0225	,0272	,0298	,0425
	Petite-fille	,1410	,0736	,0801	,0707	,0790	,0366	,0802
	Grand père	,1549	,1057	,0858	,0821	,0851	,0576	,0952
	Grand mère	,1550	,0979	,0858	,0844	,0816	,0627	,0946
	Petit-fils	,1374	,0772	,0793	,0719	,0791	,0382	,0805
	Mère	,0813	,0482	,0526	,0229	,0260	,0227	,0423
	Neveu	,0843	,0619	,0580	,0375	,0317	,0273	,0501
	Nièce	,0850	,0577	,0503	,0353	,0337	,0260	,0480
	Soeur	,1361	,0946	,0496	,0816	,0629	,0588	,0806
	Fils	,0689	,0373	,0456	,0242	,0337	,0253	,0392
	Oncle	,0977	,0761	,0678	,0489	,0383	,0498	,0631
Moyenne		,1035	,0681	,0613	,0508	,0496	,0407	,0623

La décomposition du stress facilite l'identification des sources et des objets contribuant le plus au stress global de la solution. Dans le cas présent, la majeure partie du stress parmi les sources est attribuable aux sources 1 et 2 tandis que, parmi les objets, elle est imputable aux éléments *Frère*, *Petite-fille*, *Grand-père*, *Grand-mère*, *Petit-fils* et *Soeur*.

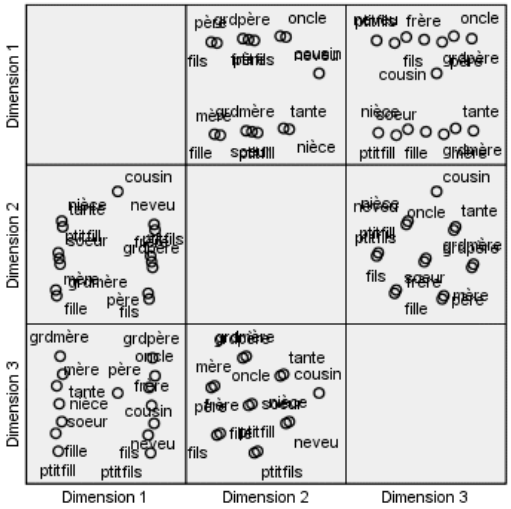
Les deux sources représentant la majeure partie du stress sont les deux groupes ayant trié les termes une seule fois. Ces informations suggèrent que les étudiants ont considéré plusieurs critères lors du tri des termes et que les étudiants qui étaient autorisés à opérer deux tris se sont focalisés sur une partie de ces critères pour le premier tri, puis ont pris en compte les autres critères à l'occasion du second tri.

Les objets qui représentent la majeure partie du stress sont ceux ayant un *degré* égal à 2. Ces personnes sont des relations n'appartenant pas à la famille "nucléaire" (*Mère*, *Père*, *Fille*, *Fils*), mais qui sont néanmoins plus proches que les autres relations. Cette position intermédiaire pourrait facilement créer un écart lors du tri de ces termes.

Coordonnées finales de l'espace commun

Le diagramme de l'espace commun fournit une représentation visuelle des relations entre les objets.

Figure 14-15
Coordonnées de l'espace commun



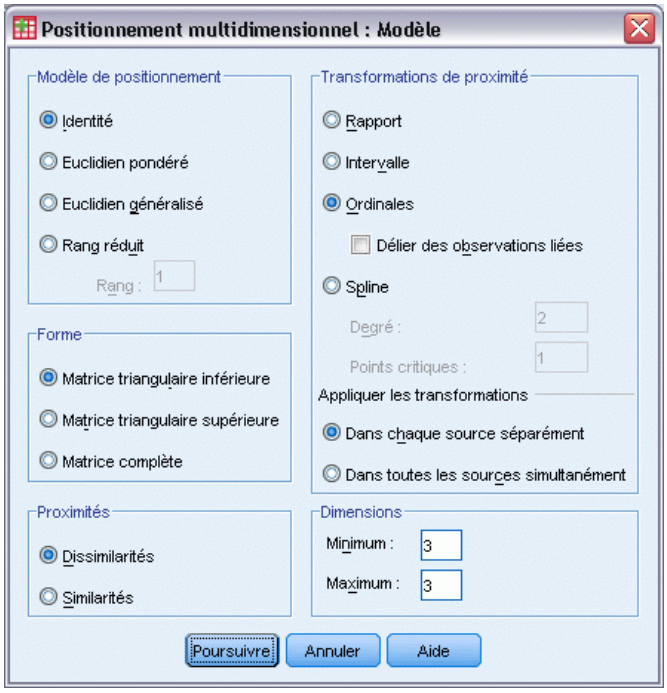
Observez les coordonnées finales des objets dans les dimensions 1 et 3 ; il s'agit du diagramme situé dans l'angle inférieur gauche de la matrice de diagrammes de dispersion. Ce diagramme montre que la dimension 1 (sur l'axe *x*) est corrélée avec la variable *sexe* et que la dimension 3 (sur l'axe *y*) est corrélée avec la variable *génér*. De gauche à droite, vous pouvez constater que la dimension 1 sépare les termes femme et homme, entre lesquels figure le terme à la fois masculin et féminin *Cousin/Cousine*. De bas en haut du diagramme, les valeurs croissantes le long de l'axe correspondent aux termes plus anciens.

Observez maintenant les coordonnées finales des objets dans les dimensions 2 et 3 ; il s'agit du diagramme situé au milieu à droite de la matrice de diagrammes de dispersion. Ce diagramme indique que la deuxième dimension (le long de l'axe *y*) correspond à la variable *degré*, les valeurs les plus élevées le long de l'axe correspondant à des termes relevant davantage de la famille « nucléaire ».

Solution tridimensionnelle avec transformations personnalisées

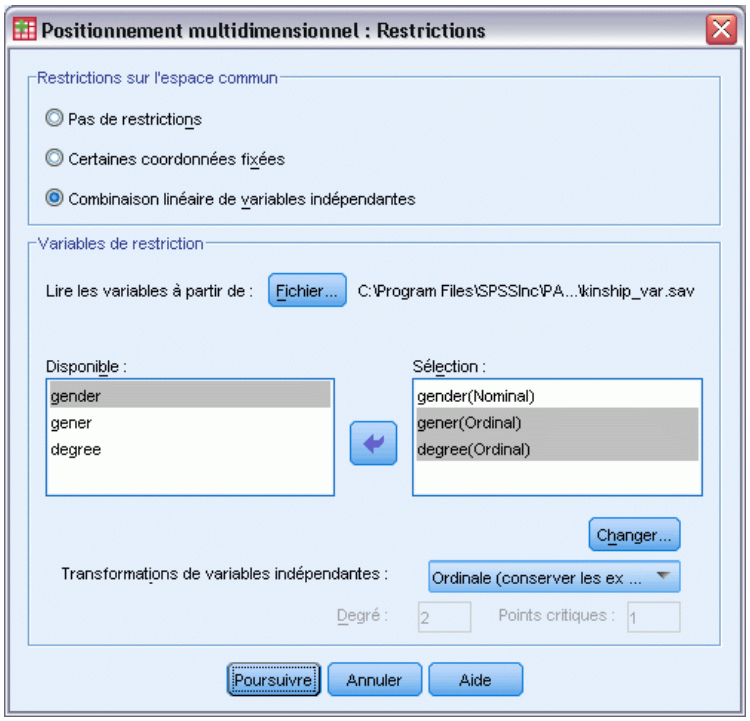
La solution précédente a été calculée à l'aide de la transformation de ratio par défaut pour les proximités et des transformations d'intervalles pour les variables indépendantes *génér* et *degré*. Les résultats sont assez bons, mais vous pouvez les améliorer à l'aide d'autres transformations. Par exemple, les proximités *sexe* et *degré* sont toutes naturellement ordonnées, mais une transformation ordinale permet de mieux les modéliser qu'une transformation linéaire.

Figure 14-16
Boîte de dialogue Modèle



- Pour réexécuter l'analyse, en codant les proximités *génér* et *degré* au niveau ordinal (conservation des ex aequo), affichez à nouveau la boîte de dialogue Positionnement multidimensionnel, puis cliquez sur Modèle.
- Sélectionnez l'option Ordinal comme transformation de proximités.
- Cliquez sur Poursuivre.
- Cliquez sur Restrictions dans la boîte de dialogue Positionnement multidimensionnel.

Figure 14-17
Boîte de dialogue Restrictions



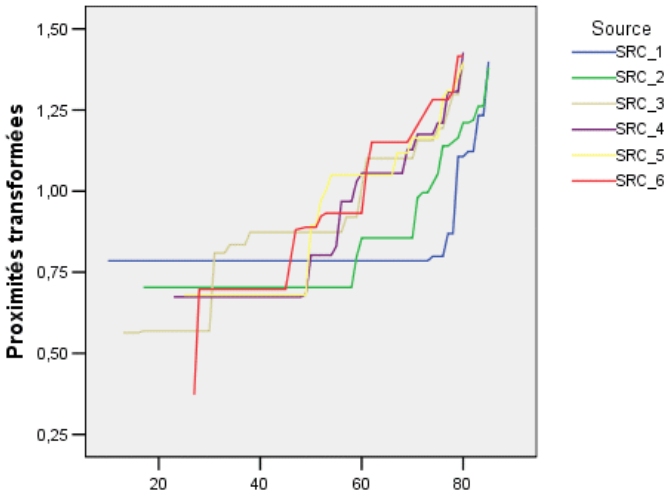
- Sélectionnez les options *sexe* et *degré*.
- Sélectionnez l'option Ordinal (conserver les ex-aequo) dans la liste déroulante Transformations des variables indépendantes.
- Cliquez sur Changer.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Positionnement multidimensionnel.

Diagrammes de transformation

Les diagrammes de transformation sont un premier indice efficace pour déterminer si les transformations initiales étaient appropriées. Si les diagrammes sont à peu près linéaires, l'hypothèse linéaire est appropriée. Sinon, vérifiez si les mesures de stress indiquent une amélioration de l'ajustement, et si le diagramme de l'espace commun facilite l'interprétation.

Chacune des variables indépendantes obtenant des transformations à peu près linéaires, il peut s'avérer approprié de les interpréter en tant que données numériques. Toutefois, les proximités n'obtenant pas de transformation linéaire, il est possible que la transformation ordinale convienne davantage pour celles-ci.

Figure 14-18
Transformations transformées



Mesures de stress

Le stress de la solution actuelle prend en charge l'argument de codage des proximités au niveau ordinal.

Figure 14-19
Mesures de stress et d'ajustement

Stress brut normalisé	,03137
Stress I	,17712[a]
Stress II	,61987[a]
Stress S	,07953[b]
Dispersion expliquée par	,96863
Coefficient de congruence de Tucker	,98419

PROXSCAL minimise le stress brut normalisé.

a. Facteur de codage optimal = 1,032.

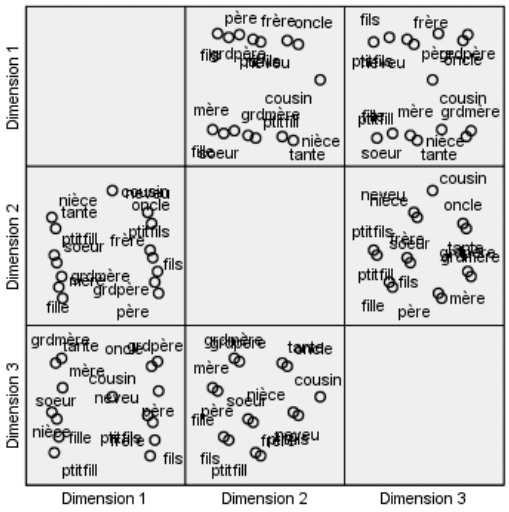
b. Facteur de codage optimal = ,980.

Le stress ligne normalisé de la solution antérieure a pour valeur 0,06234. Le codage des variables à l'aide de transformations personnalisées divise par 2 la valeur du stress, qui passe à 0,03137.

Coordonnées finales de l'espace commun

Les diagrammes de l'espace commun offrent essentiellement la même interprétation des dimensions que la solution précédente.

Figure 14-20
Coordonnées de l'espace commun



Analyse

Il est préférable de traiter les proximités en tant que variables ordinales, dans la mesure où les mesures de stress affichent une amélioration sensible. Ensuite, vous pouvez, si vous le souhaitez, "délier" les variables ordinales—c'est-à-dire, permettre à des valeurs équivalentes des variables initiales d'obtenir différentes valeurs transformées. Par exemple, dans la première source, les proximités entre *Tante* et *Fils*, ainsi qu'entre *Tante* et *Petit-fils*, ont pour valeur 85. L'approche "liée" des variables ordinales oblige les valeurs transformées de ces proximités à être équivalentes, mais vous n'avez aucune raison particulière de supposer qu'elles doivent l'être. Dans ce cas, vous pouvez autoriser la suppression des liens des proximités de manière à éviter toute restriction superflue.

Lectures recommandées

Pour plus d'informations sur le positionnement multidimensionnel, reportez-vous aux documents suivants :

Commandeur, J. J. F., et W. J. Heiser. 1993. *Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices*. Leiden: Department of Data Theory, University of Leiden.

De Leeuw, J., et W. J. Heiser. 1980. Multidimensional scaling with restrictions on the configuration. Dans : *Multivariate Analysis, Vol. V,* P. R. Krishnaiah, éd. Amsterdam: North-Holland.

Heiser, W. J. 1981. *Unfolding analysis of proximity data*. Leiden: Department of Data Theory, University of Leiden.

Heiser, W. J., et F. M. T. A. Busing. 2004. Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations. Dans : *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, éd. Thousand Oaks, Californie: Sage Publications, Inc..

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, .

Kruskal, J. B. 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, .

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrika*, 27, .

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function II. *Psychometrika*, 27, .

Chapitre
15

Dépliage multidimensionnel

La procédure de dépliage multidimensionnel tente de trouver une échelle quantitative commune vous permettant d'examiner les relations entre deux ensembles d'objets de manière visuelle.

Exemple \: Préférences alimentaires du petit-déjeuner

Dans une étude classique (Green et Rao, 1972), on a demandé à 21 étudiants en MBA (Master of Business Administration) de l'école de Wharton et à leurs conjoints de classer 15 aliments du petit-déjeuner selon leurs préférences, de 1= "aliment préféré" à 15= "aliment le moins apprécié". Ces informations sont regroupées dans le fichier *breakfast_overall.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*.

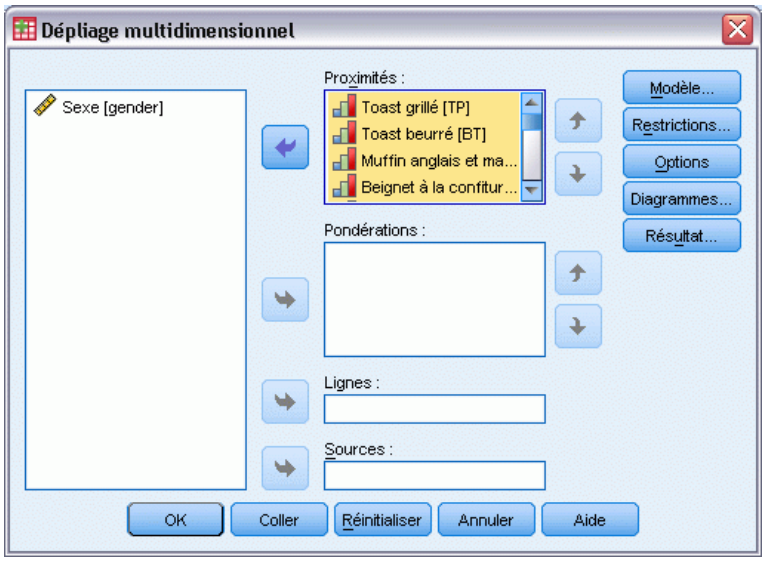
Le résultat de l'étude illustre un exemple de problème de dégénérescence typique, inhérent à la plupart des algorithmes de dépliage multidimensionnel, résolu en pénalisant le coefficient de variation des proximités transformées (Busing, Groenen, et Heiser, 2005). Vous allez voir ce qu'est une solution dégénérée et comment résoudre le problème à l'aide du dépliage multidimensionnel, qui permet de déterminer la logique suivie par les individus dans leur classement. La syntaxe servant à reproduire ces analyses se trouve dans *prefscal_breakfast-overall.sps*.

Création d'une solution dégénérée

- Pour lancer une analyse Dépliage multidimensionnel, choisissez les options suivantes dans les menus :

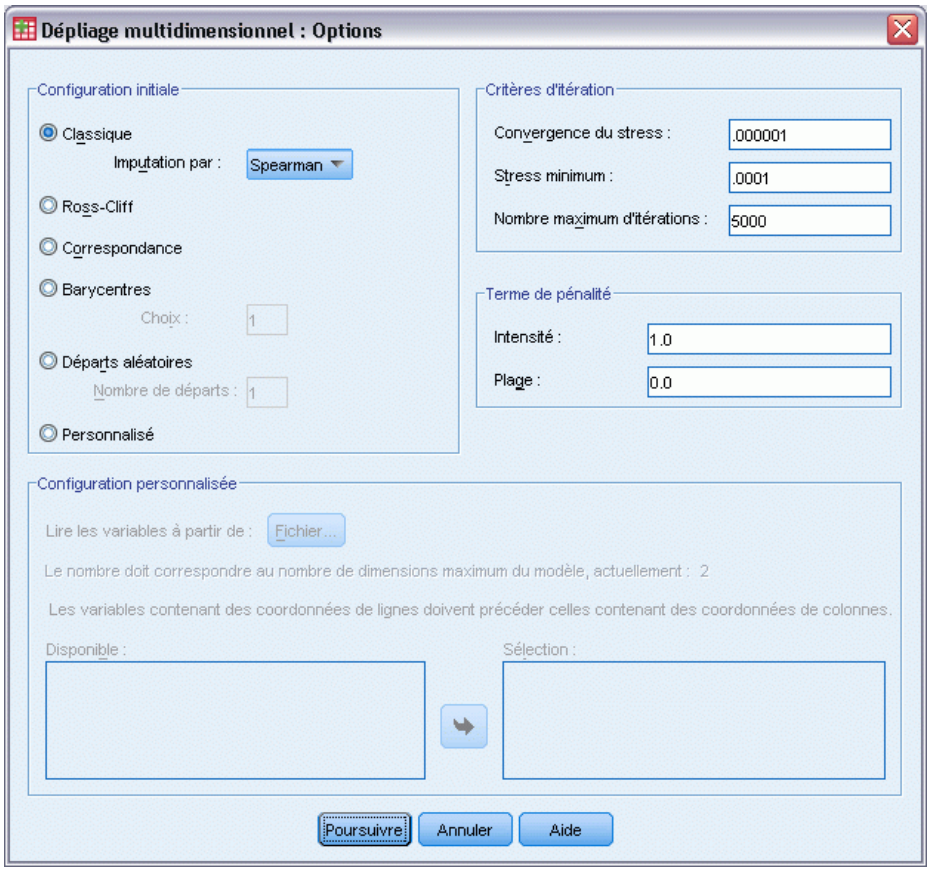
  Analyse > Echelle > Dépliage multidimensionnel (PREFSCAL)...

Figure 15-1
Boîte de dialogue principale Dépliage multidimensionnel



- Sélectionnez les options allant de *Pain grillé* à *Tartine beurrée* comme variables de proximité.
- Cliquez sur Options.

Figure 15-2
Options



- Sélectionnez Spearman comme méthode d'imputation du départ classique.
- Dans le groupe Terme de pénalité, tapez 1,0 comme valeur du paramètre Force et 0,0 comme valeur du paramètre Intervalle. Ceci désactive le terme de pénalité.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Dépliage multidimensionnel.

Voici la syntaxe de commande générée par ces sélections :

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INITIAL=CLASSICAL (SPEARMAN)
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(1.0) OMEGA(0.0)
  /PRINT=MEASURES COMMON
  /PLOT=COMMON .
```

- Cette syntaxe indique une analyse des variables *tb (pain grillé)* à *cmb (tartine beurrée)*.

- La sous-commande INITIAL spécifie que les valeurs de départ sont imputées à l'aide des distances de Spearman.
- Les valeurs spécifiées dans la sous-commande PENALTY annulent le terme de pénalité et, par conséquent, la procédure minimise la mesure du stress-I de Kruskal. La solution obtenue est donc dégénérée.
- La sous-commande PLOT demande des diagrammes de l'espace commun.
- Tous les autres paramètres sont réinitialisés à leur valeur par défaut.

Mesures

Figure 15-3
Mesures de la solution dégénérée

Itérations		154
Valeur de fonction finale		,0000990
Eléments de valeur de fonction	Elément de stress	,0000990
	Elément de pénalité	1,0000000
Médiocrité d'ajustement	Stress normalisé	,0000000
	Stress-I de Kruskal	,0000990
	Stress-II de Kruskal	,6129749
	S-Stress-I de Young	,0001980
	S-Stress-II de Young	,7703817
Qualité de l'ajustement :	Dispersion représentée	1,0000000
	Variance expliquée par	,6230788
	Ordres de préférence récupérés	,7074830
	Rho de Spearman	,7450748
	Tau-b de Kendall	,6218729
Coefficients de variation	Proximités de variation	,5590170
	Proximités de variation transformées	,0000924
	Distances de variation	,1808765
Indices de dégénérescence	Somme des carrés des indices de fusion de DeSarbo	117,3115413
	Index de non-dégénérescence simple de Shepard	,0000000

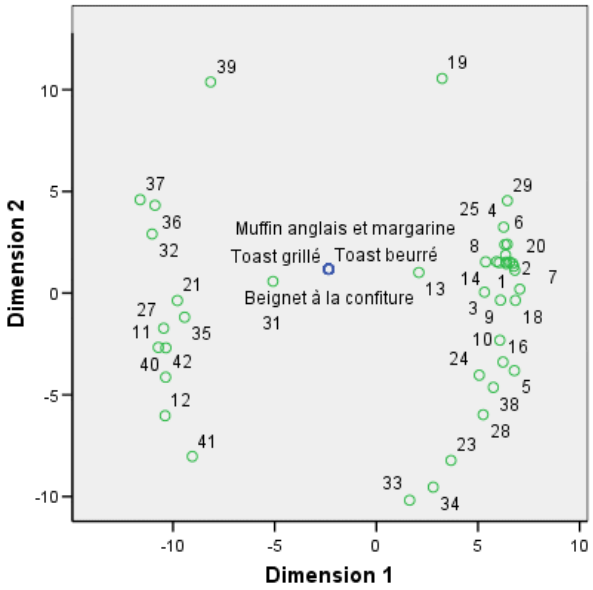
L'algorithme converge vers une solution après 154 itérations et applique un stress pénalisé (marqué comme la valeur de la fonction finale) de 0,0000990. Etant donné que le terme de pénalité a été désactivé, la mesure du stress pénalisé est égale au stress-I de Kruskal (la partie stress de la valeur de la fonction est équivalente à la mesure du défaut de l'ajustement de Kruskal). Des valeurs de stress basses indiquent généralement que la solution est bien adaptée aux données, mais il existe plusieurs signes d'avertissement d'une solution dégénérée.

- Le coefficient de variation des proximités transformées est très faible comparé au coefficient de variation des proximités d'origine. Ceci suggère que les proximités transformées de chaque ligne sont quasi-constantes et que la solution ne montrera ainsi aucune discrimination entre les objets.

- La somme des carrés des indices d'"intermixité" de DeSarbo sont une mesure de l'intermixité des points des différents sous-ensembles. L'absence d'intermixité est un signe d'avertissement d'une dégénérescence probable de la solution. Plus la valeur rapportée est proche de 0, plus la solution est intermixée. Plus elle est élevée, moins la solution est intermixée.
- L'index estimatif de non-dégénérescence de Shepard, rapporté sous forme d'un pourcentage des différentes distances, est égal à 0. Il s'agit là d'une indication numérique claire d'une différence insuffisante entre les distances et donc de la dégénérescence probable de la solution.

Espace commun

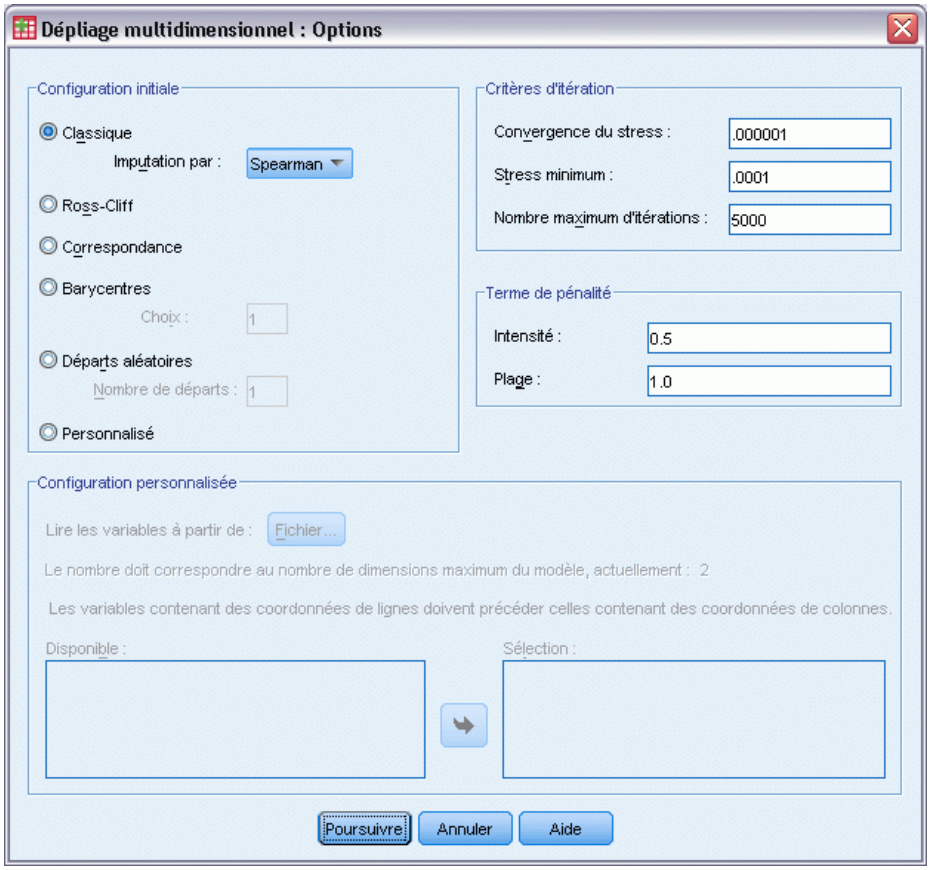
Figure 15-4
Diagramme joint de l'espace commun pour une solution dégénérée



Le diagramme joint de l'espace commun des objets de lignes et de colonnes apporte une confirmation visuelle de la dégénérescence de la solution. Les objets de lignes (individus) se situent à la circonférence d'un cercle centré sur les objets de colonnes (aliments du petit-déjeuner), dont les coordonnées se sont réduites à un point unique.

Exécution d'une analyse non dégénérée

Figure 15-5
Options



- Pour produire une solution non dégénérée, cliquez sur l'outil Rappeler boîte de dialogue et sélectionnez Dépliage multidimensionnel.
- Cliquez sur Options dans la boîte de dialogue Dépliage multidimensionnel.
- Dans le groupe Terme de pénalité, tapez 0,5 comme valeur du paramètre Force et 1,0 comme valeur du paramètre Intervalle. Ceci désactive le terme de pénalité.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Dépliage multidimensionnel.

Voici la syntaxe de commande générée par ces sélections :

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INITIAL=CLASSICAL (SPEARMAN)
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(1.0) OMEGA(0.0)
```

```
/PRINT=MEASURES COMMON
/PLOT=COMMON .
```

- La seule différence réside dans la sous-commande `PENALTY`. `LAMBDA` et `OMEGA` ont été définies respectivement sur 0,5 et 1,0, leurs valeurs pas défaut.

Mesures

Figure 15-6
Mesures de la solution non dégénérée

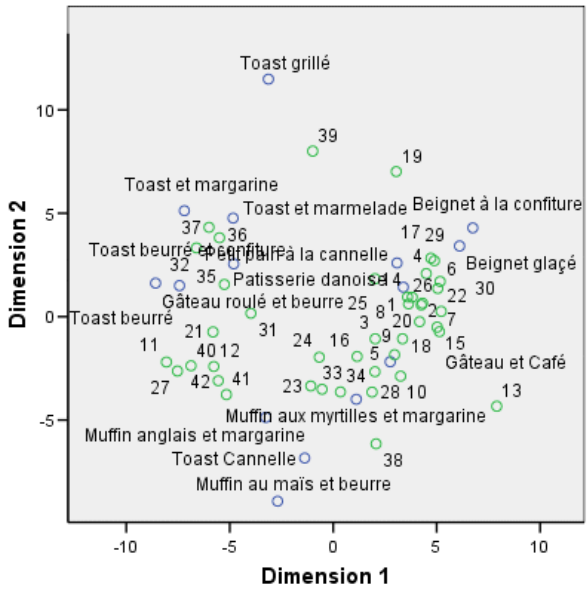
Itérations		157
Valeur de fonction finale		,6848930
Eléments de valeur de fonction	Elément de stress	,2428268
	Elément de pénalité	1,9317409
Médiocrité d'ajustement	Stress normalisé	,0583589
	Stress-I de Kruskal	,2415758
	Stress-II de Kruskal	,5875599
	S-Stress-I de Young	,3446361
	S-Stress-II de Young	,5030127
Qualité de l'ajustement :	Dispersion représentée	,9416411
	Variance expliquée par	,7651552
	Ordres de préférence récupérés	,7818594
	Rho de Spearman	,8179181
	Tau-b de Kendall	,6916725
Coefficients de variation	Proximités de variation	,5590170
	Proximités de variation transformées	,6006156
	Distances de variation	,4833617
Indices de dégénérescence	Somme des carrés des indices de fusion de DeSarbo	,1590979
	Index de non-dégénérescence simple de Shepard	,7895692

Les problèmes relevés dans les mesures de la solution dégénérée sont à présent corrigés.

- Le stress normalisé n'est plus égal à 0.
- Le coefficient de variation des proximités transformées présente maintenant une valeur similaire au coefficient de variation des proximités d'origine.
- Les indices d'"intermixité" de DeSarbo sont beaucoup plus proches de 0, indiquant une grande amélioration de l'intermixité de la solution.
- L'index estimatif de non-dégénérescence de Shepard, rapporté sous forme d'un pourcentage des différentes distances, est environ égal à 80 %. Les différences entre les distances sont suffisantes et la solution est probablement non dégénérée.

Espace commun

Figure 15-7
Diagramme joint de l'espace commun pour une solution non dégénérée



Le diagramme joint de l'espace commun permet une interprétation des dimensions. La dimension horizontale semble indiquer une discrimination entre les pains mous et durs ou encore les toasts, les aliments les plus mous se trouvant dans la partie droite de l'axe. La dimension verticale ne permet pas une interprétation claire, peut-être uniquement une discrimination basée sur la commodité, les aliments les plus formels se trouvant dans la partie inférieure de l'axe.

Ceci conduit à la formation de plusieurs groupes d'aliments. Par exemple, les pains aux raisins, les brioches et les beignets forment un groupe d'aliments mous et quelque peu informels. Les croissants et pains au chocolat forment un groupe d'aliments plus durs et plus formels. Les toasts et tartines forment un groupe d'aliments durs et quelque peu informels. Le pain grillé est un aliment dur, extrêmement informel.

Les individus représentés par les objets de lignes se divisent en plusieurs groupes bien délimités, selon leurs préférences pour les aliments mous ou durs, avec de nombreuses variations intra-groupes le long de la dimension verticale.

Exemple \: Dépliage tridimensionnel des préférences relatives aux aliments du petit-déjeuner

Dans une étude classique (Green et al., 1972), on a demandé à 21 étudiants en MBA (Master of Business Administration) de l'école de Wharton et à leurs conjoints de classer 15 aliments du petit-déjeuner selon leurs préférences, de 1= "aliment préféré" à 15= "aliment le moins apprécié". Leurs préférences ont été enregistrées dans six scénarios différents, allant de « Préférence générale » à « En-cas avec boisson uniquement ». Ces informations sont regroupées dans le

fichier *breakfast.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*.

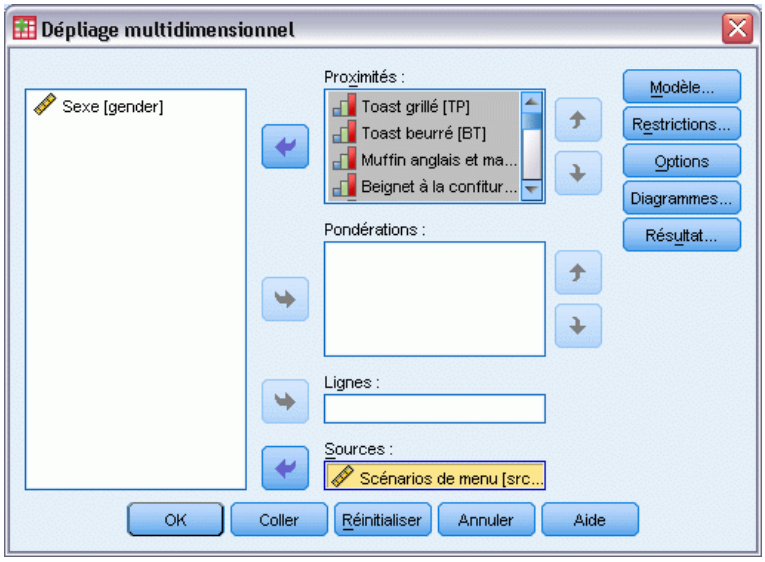
Les six scénarios peuvent être traités en tant que sources distinctes. Utilisez la procédure PREFSCAL pour effectuer un dépliage tridimensionnel des lignes, des colonnes et des sources. La syntaxe servant à reproduire ces analyses se trouve dans *prefscal_breakfast.sps*.

Exécution de l'analyse

- Pour lancer une analyse Dépliage multidimensionnel, choisissez les options suivantes dans les menus :

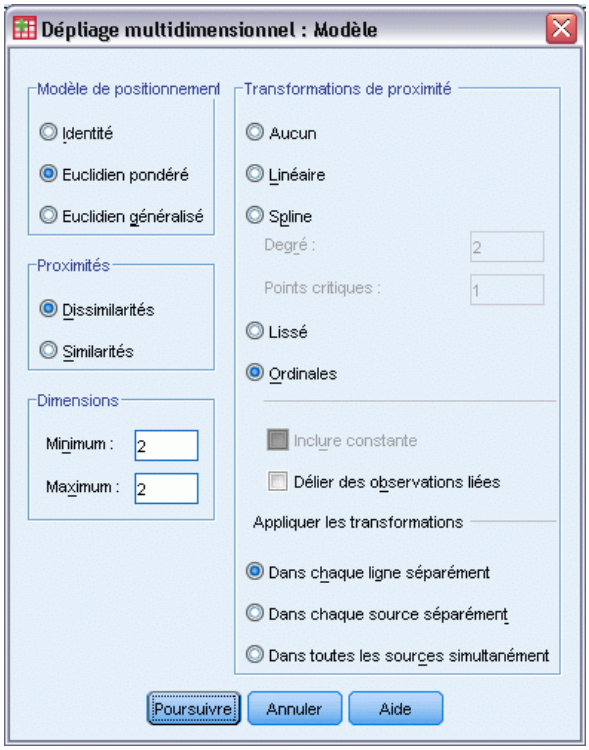
  Analyse > Echelle > Dépliage multidimensionnel (PREFSCAL)...

Figure 15-8
Boîte de dialogue principale Dépliage multidimensionnel



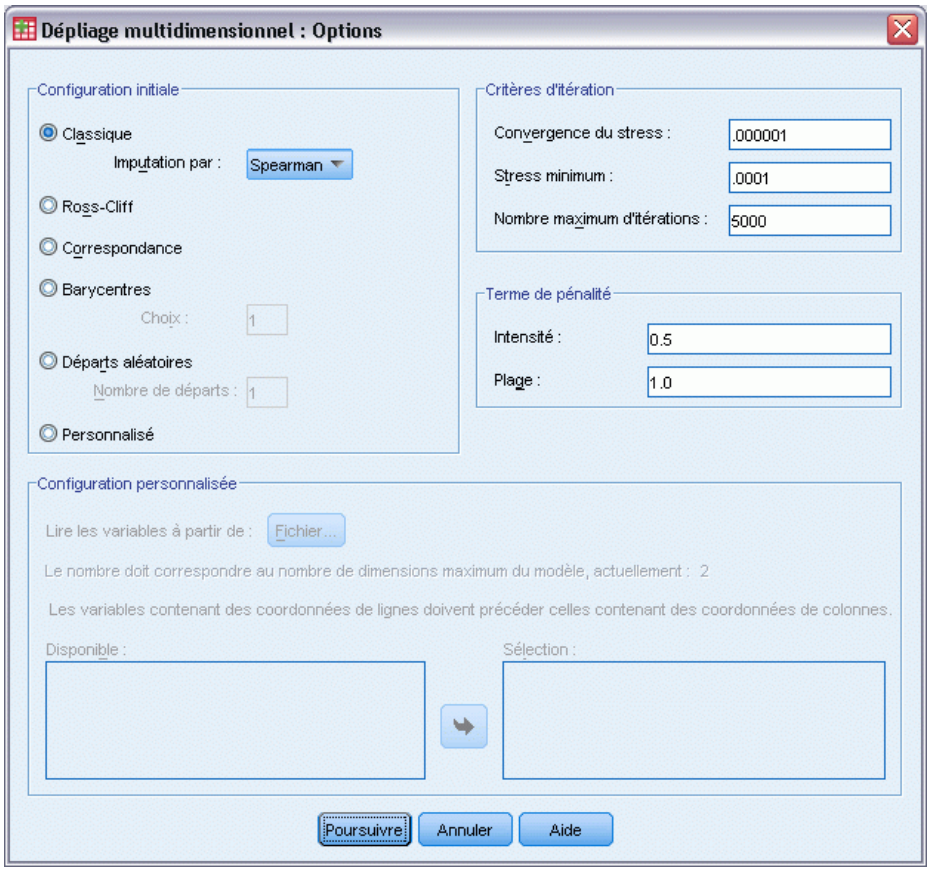
- Sélectionnez les options allant de *Pain grillé* à *Tartine beurrée* comme variables de proximité.
- Sélectionnez *Scénarios* comme variable source.
- Cliquez sur Modèle.

Figure 15-9
Boîte de dialogue Modèle



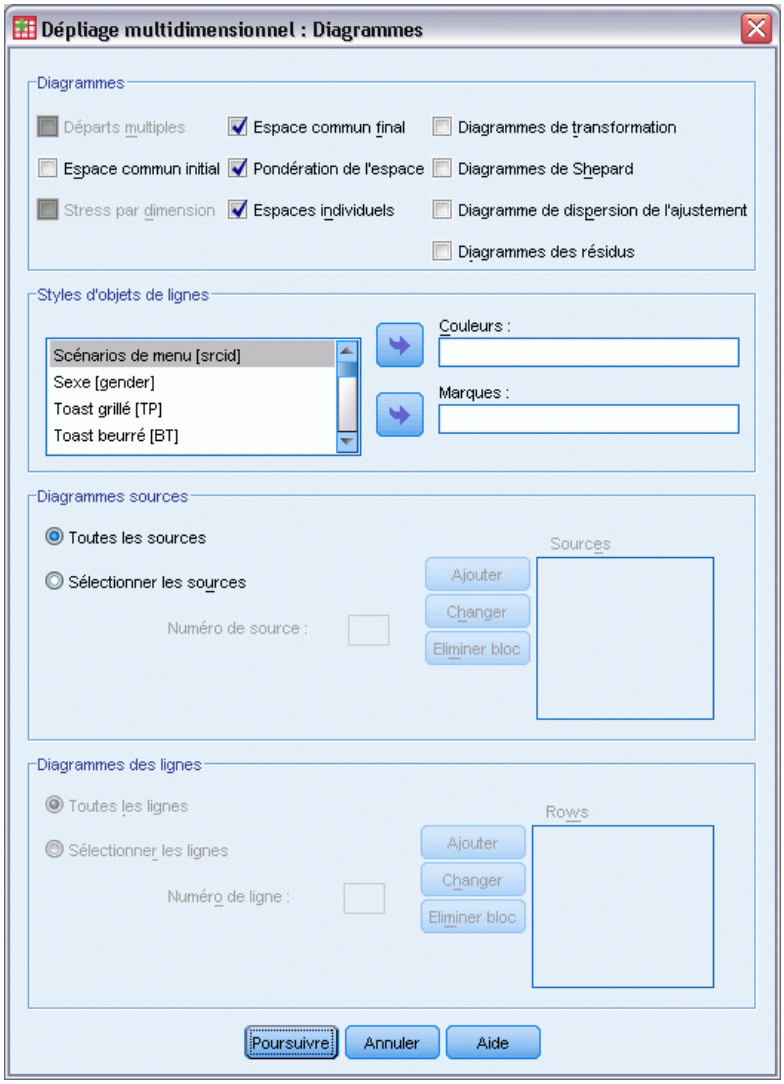
- Sélectionnez Euclidien pondéré comme modèle de positionnement.
- Cliquez sur Poursuivre.
- Cliquez sur Options dans la boîte de dialogue Dépliage multidimensionnel.

Figure 15-10
Options



- Sélectionnez Spearman comme méthode d'imputation du départ classique.
- Cliquez sur Poursuivre.
- Cliquez sur Diagrammes dans la boîte de dialogue Dépliage multidimensionnel.

Figure 15-11
Boîte de dialogue Diagrammes



- Sélectionnez l'option Espaces individuels dans le groupe Diagrammes.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Dépliage multidimensionnel.

Voici la syntaxe de commande générée par ces sélections :

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INPUT=SOURCES(srcid )
  /INITIAL=CLASSICAL (SPEARMAN)
  /CONDITION=ROW
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /MODEL=WEIGHTED
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
```

```
MAXITER(5000)
/PENALTY=LAMBDA(0.5) OMEGA(1.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON WEIGHTS INDIVIDUAL ( ALL ) .
```

- Cette syntaxe indique une analyse des variables *tb (pain grillé)* à *cmb (tartine beurrée)*. La variable *srcid* est utilisée pour identifier les sources.
- La sous-commande `INITIAL` spécifie que les valeurs de départ sont calculées à l'aide des distances de Spearman.
- La sous-commande `MODEL` spécifie un modèle Euclidien pondéré, qui permet à chaque espace individuel de pondérer les dimensions de l'espace commun d'une manière différente.
- La sous-commande `PLOT` demande des diagrammes de l'espace commun, des espaces individuels et des pondérations des espaces individuels.
- Tous les autres paramètres sont réinitialisés à leur valeur par défaut.

Mesures

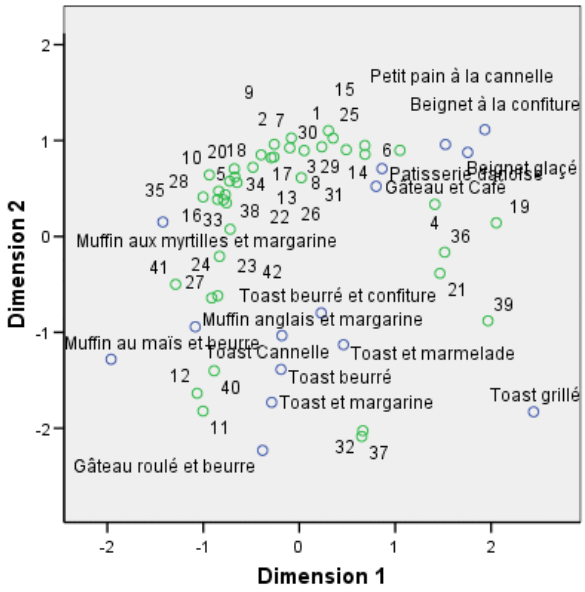
Figure 15-12
Mesures

Itérations		481
Valeur de fonction finale		,8199642
Eléments de valeur de fonction	Elément de stress	,3680994
	Elément de pénalité	1,8265211
Médiocrité d'ajustement	Stress normalisé	,1335343
	Stress-I de Kruskal	,3654234
	Stress-II de Kruskal	,9780824
	S-Stress-I de Young	,4938016
	S-Stress-II de Young	,6912352
Qualité de l'ajustement :	Dispersion représentée	,8664657
	Variance expliquée par	,5024853
	Ordres de préférence récupérés	,7025321
	Rho de Spearman	,6271702
	Tau-b de Kendall	,4991188
Coefficients de variation	Proximités de variation	,5590170
	Proximités de variation transformées	,6378878
	Distances de variation	,4484515
Indices de dégénérescence	Somme des carrés des indices de fusion de DeSarbo	,2199287
	Index de non-dégénérescence simple de Shepard	,7643613

L'algorithme converge après 481 itérations, avec une mesure du stress pénalisé finale de 0,8199642. Les coefficients de variation et l'index de Shepard sont suffisamment élevés et les indices de DeSarbo suffisamment bas pour suggérer qu'il n'existe aucun problème de dégénérescence.

Espace commun

Figure 15-13
Diagramme join de l'espace commun



Le diagramme joint de l'espace commun montre une configuration finale très similaire à l'analyse bidimensionnelle des préférences générales, avec une solution transposée au-dessus de la ligne des 45 °. Ainsi, la dimension verticale semble indiquer une discrimination entre les pains mous et durs ou encore les toasts, les aliments les plus mous se trouvant dans la partie supérieure de l'axe. La dimension horizontale ne permet pas une interprétation claire, peut-être uniquement une discrimination basée sur la commodité, les aliments les plus formels se trouvant dans la partie gauche de l'axe.

Les individus représentés par les objets de lignes se divisent toujours en plusieurs groupes bien délimités, selon leurs préférences pour les aliments mous ou durs avec de nombreuses variations intra-groupes le long de la dimension horizontale.

Espaces individuels

Figure 15-14
Pondérations des dimensions

		Dimension		
		1	2	Spécificité[a]
Source	Préférence générale	3,235	4,297	,186
	Petit déjeuné, avec jus, oeufs au bacon et boisson	4,883	2,193	,457
	Petit déjeuné, avec jus, céréale et boisson	4,131	3,438	,109
	Petit déjeuné avec jus, crêpes, saucisses et boisson	4,291	3,267	,164
	Petit déjeuné avec seulement une boisson	3,124	4,413	,223
	Encas avec seulement une boisson	2,750	4,541	,313
Importance :[b]		,504	,496	

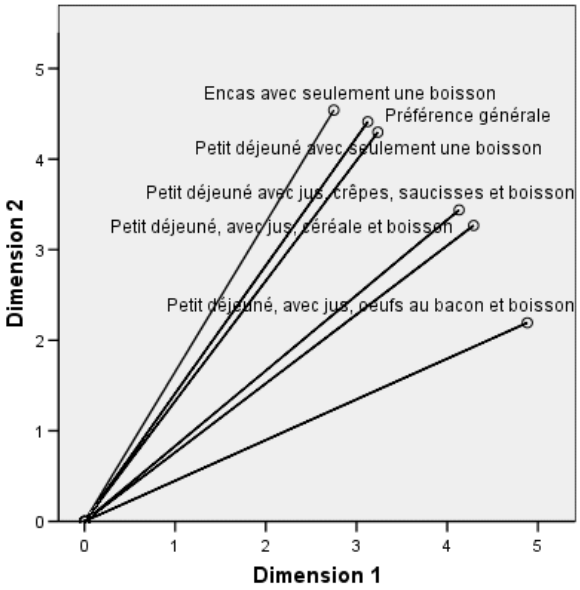
a. La spécificité indique la spécificité d'une source. La plage de spécificité est comprise entre zéro et un. Zéro indique une source moyenne ayant des pondérations de dimension identiques et un, une source très spécifique dotée d'une pondération élevée de dimension exceptionnelle et d'autres pondérations proches de zéro.

b. Importance relative de chaque dimension, fournie en tant que rapport entre la somme des carrés d'une dimension et la somme totale des carrés.

Un espace individuel est calculé pour chaque source. Les pondérations des dimensions indiquent l'impact des différents espaces individuels sur les dimensions de l'espace commun. Une pondération plus élevée indique une plus grande distance à l'intérieur de l'espace individuel et donc une plus grande discrimination entre les objets de la dimension en question pour cet espace individuel.

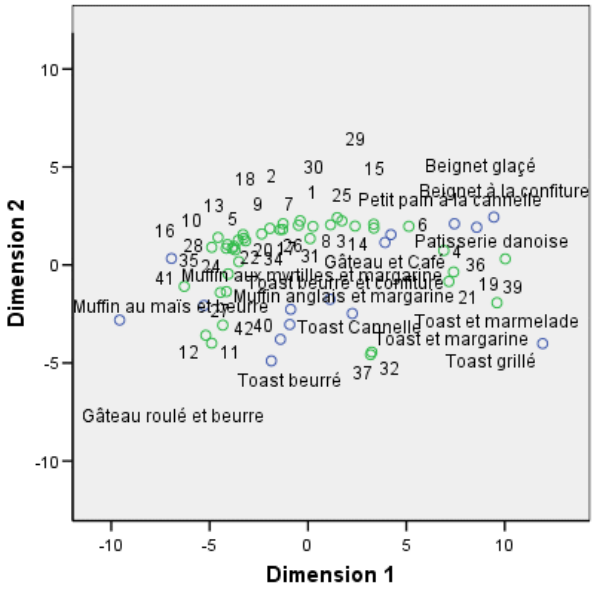
- La **spécificité** est une mesure de la différence entre l'espace individuel et l'espace commun. Un espace individuel identique à l'espace commun présenterait des pondérations de dimensions identiques et une spécificité de 0, alors qu'un espace individuel spécifique à une dimension particulière présenterait une seule pondération de dimension élevée et une spécificité de 1. Dans le cas présent, les sources les plus divergentes sont *Petit-déjeuner avec jus de fruit, jambon, oeufs et boisson* et *En-cas avec boisson chaude uniquement*.
- L'**importance** est la mesure de la contribution relative de chaque dimension à la solution. Dans le cas présent, les dimensions présentent une importance égale.

Figure 15-15
Pondérations des dimensions



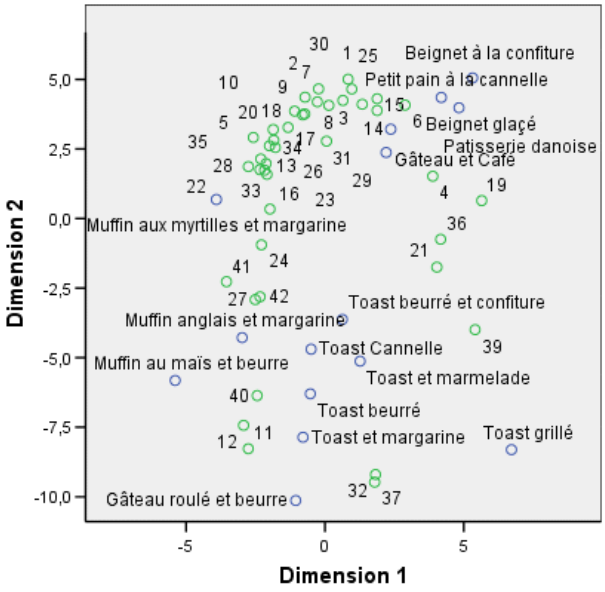
Le diagramme des pondérations des dimensions offre une vue d'ensemble du tableau des pondérations. Les groupes *Petit-déjeuner avec jus de fruit, jambon, oeufs et boisson* et *En-cas avec boisson uniquement* sont les plus proches des axes des dimensions, mais aucun des deux n'est spécifiquement rattaché à une dimension particulière.

Figure 15-16
Diagramme joint de l'espace individuel Petit-déjeuner avec jus de fruit, jambon, oeufs et boisson



Le diagramme joint de l'espace individuel *Petit-déjeuner avec jus de fruit, jambon, oeurs et boisson* illustre l'effet de ce scénario sur les préférences. La source repose majoritairement sur la premère dimension, donc la différenciation entre les aliments est principalement due à la première dimension.

Figure 15-17
Diagramme joint de l'espace individuel En-cas avec boisson uniquement



Le diagramme joint de l'espace individuel *En-cas avec boisson uniquement* illustre l'effet de ce scénario sur les préférences. La source repose majoritairement sur la deuxième dimension, donc la différenciation entre les aliments est principalement due à la deuxième dimension. Cependant, une mineure partie de la différenciation se fait également le long de la première dimension en raison de la spécificité relativement basse de la source.

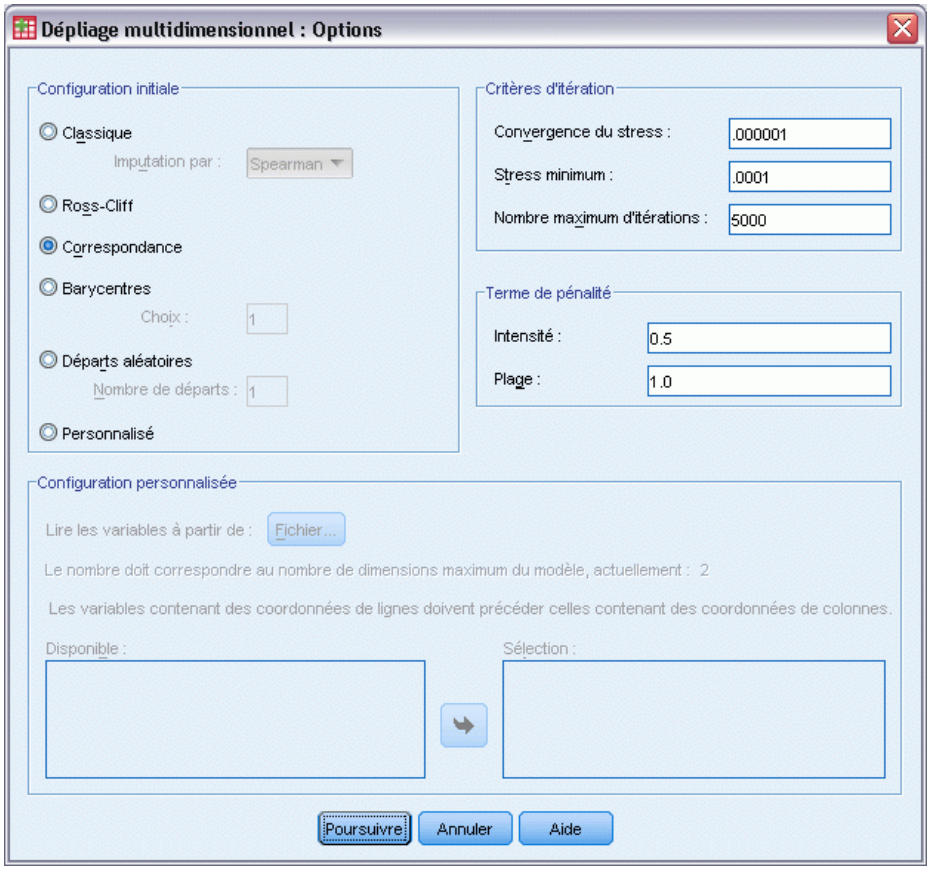
Utilisation d'une configuration initiale différente

La configuration finale peut dépendre des points de départs donnés à l'algorithme. Idéalement, la structure générale de la solution doit rester la même, sans quoi il peut s'avérer difficile de déterminer laquelle est correcte. Cependant, des variations structurelles de détail peuvent être envisagées dans différentes configurations initiales, comme par exemple l'utilisation d'un départ par correspondance dans l'analyse tridimensionnelle des données du petit-déjeuner.

▶ Pour produire une solution avec départ par correspondance, cliquez sur l'outil Rappeler boîte de dialogue et sélectionnez Dépliage multidimensionnel.

▶ Cliquez sur Options dans la boîte de dialogue Dépliage multidimensionnel.

Figure 15-18
Options



▶ Sélectionnez Correspondance dans le groupe Configuration initiale.

▶ Cliquez sur Poursuivre.

▶ Cliquez sur OK dans la boîte de dialogue Dépliage multidimensionnel.

Voici la syntaxe de commande générée par ces sélections :

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INPUT=SOURCES(srcid )
  /INITIAL=CORRESPONDENCE
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(1.0) OMEGA(0.0)
  /PRINT=MEASURES COMMON
```

```
/PLOT=COMMON WEIGHTS INDIVIDUAL ( ALL ) .
```

- La seule différence réside dans la sous-commande `INITIAL`. La configuration de départ a été définie sur `CORRESPONDENCE`, qui utilise les résultats d'une analyse des correspondances sur les données inversées (similitudes au lieu des différences), avec une normalisation symétrique des coordonnées des lignes et des colonnes.

Mesures

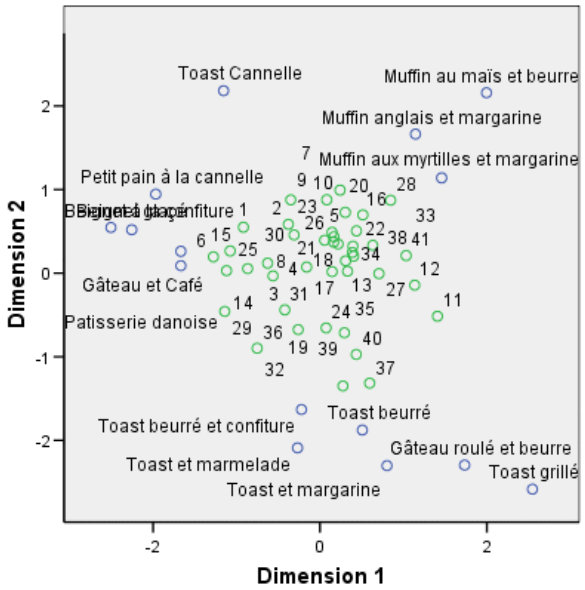
Figure 15-19
Mesures de la configuration initiale des correspondances

Itérations		385
Valeur de fonction finale		,8140741
Eléments de valeur de fonction	Elément de stress	,3493640
	Elément de pénalité	1,8969229
Médiocrité d'ajustement	Stress normalisé	,1212145
	Stress-I de Kruskal	,3481587
	Stress-II de Kruskal	1,0770522
	S-Stress-I de Young	,4812632
	S-Stress-II de Young	,6871733
Qualité de l'ajustement :	Dispersion représentée	,8787855
	Variance expliquée par	,5183498
	Ordres de préférence récupérés	,7174981
	Rho de Spearman	,6446272
	Tau-b de Kendall	,5165230
Coefficients de variation	Proximités de variation	,5590170
	Proximités de variation transformées	,6122308
	Distances de variation	,4043695
Indices de dégénérescence	Somme des carrés des indices de fusion de DeSarbo	1,7571887
	Index de non-dégénérescence simple de Shepard	,7532124

L'algorithme converge après 385 itérations, avec une mesure du stress pénalisé finale de 0,8140741. Les valeurs de cette statistique, du défaut de l'ajustement, de la qualité de l'ajustement, des coefficients de variation et de l'index de Shepard sont toutes très similaires à celles de la solution utilisant le départ de Spearman classique. Les indices de DeSarbo sont quelque peu différents, avec une valeur de 1,7571887 au lieu de 0,2199287, ce qui suggère que la solution utilisant le départ par correspondance n'est pas aussi mixte. Pour voir dans quelle mesure ceci affecte la solution, reportez-vous au diagramme joint de l'espace commun.

Espace commun

Figure 15-20
Diagramme joint de l'espace commun pour la configuration initiale des correspondances



Le diagramme joint de l'espace commun montre une configuration finale similaire à l'analyse faite avec la configuration initiale de Spearman classique. Cependant, les objets de colonnes (aliments du petit-déjeuner) se situent autour des objets de lignes (individus) au lieu que l'ensemble soit intermixé.

Espaces individuels

Figure 15-21
Pondérations des dimensions pour la configuration initiale des correspondances

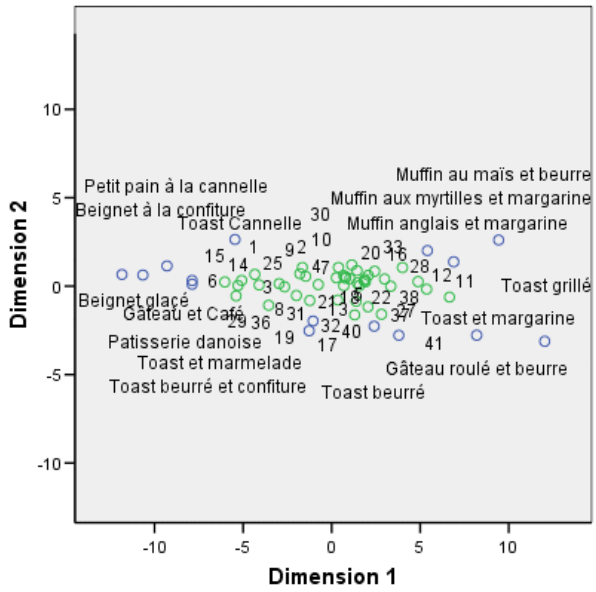
		Dimension		
		1	2	Spécificité[a]
Source	Préférence générale	2,836	3,877	,279
	Petit déjeuné, avec jus, oeufs au bacon et boisson	4,727	1,207	,636
	Petit déjeuné, avec jus, céréale et boisson	4,183	2,377	,263
	Petit déjeuné avec jus, crêpes, saucisses et boisson	4,412	1,993	,389
	Petit déjeuné avec seulement une boisson	2,605	4,050	,351
	Encas avec seulement une boisson	1,864	4,415	,552
Importance :[b]		,556	,444	

a. La spécificité indique la spécificité d'une source. La plage de spécificité est comprise entre zéro et un. Zéro indique une source moyenne ayant des pondérations de dimension identiques et un, une source très spécifique dotée d'une pondération élevée de dimension exceptionnelle et d'autres pondérations proches de zéro.

b. Importance relative de chaque dimension, fournie en tant que rapport entre la somme des carrés d'une dimension et la somme totale des carrés.

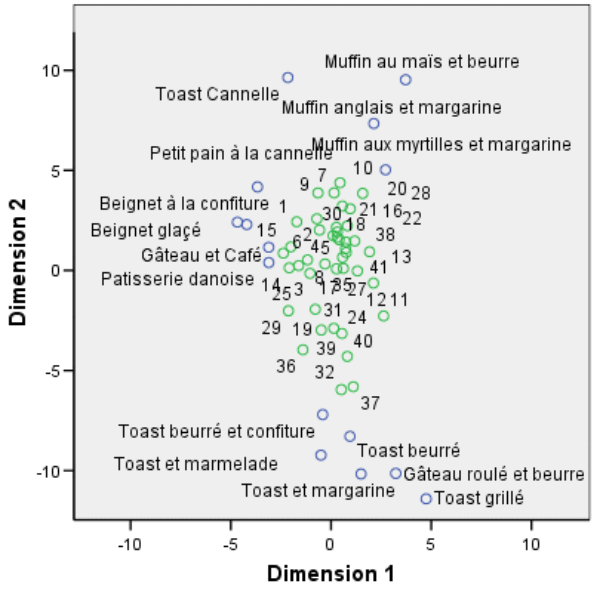
Dans la configuration initiale des correspondances, chaque espace individuel présente une spécificité plus élevée, c'est à dire que chaque situation dans laquelle les participants ont classé les aliments de petit-déjeuner est plus fortement associée à une dimension spécifique. Les sources les plus divergentes sont toujours *Petit-déjeuner avec jus de fruit, jambon, oeufs et boisson* et *En-cas avec boisson uniquement*.

Figure 15-22
Diagramme joint de l'espace individuel Petit-déjeuner avec jus de fruit, jambon, oeufs et boisson pour la configuration initiale des correspondances



La spécificité plus élevée apparaît clairement dans le diagramme joint de l'espace individuel *Petit-déjeuner avec jus de fruit, jambon, oeufs et boisson*. La source affecte encore plus fortement la première dimension que dans le cas du départ de Spearman classique, si bien que les objets de colonnes et de lignes présentent une variation un peu moins importante sur l'axe vertical et un peu plus importante sur l'axe horizontal.

Figure 15-23
Diagramme joint de l'espace individuel En-cas avec boisson uniquement pour la configuration initiale des correspondances



Le diagramme joint de l'espace individuel *En-cas avec boissson uniquement* montre que les objets de lignes et de colonnes sont plus proches d'une ligne verticale que dans le cas du départ de Spearman classique.

Exemple \: Examen de la justesse de la relation comportement-situation

Dans un exemple classique (Price et Bouffard, 1974), il a été demandé à 52 étudiants de noter les combinaisons établies à partir de 15 situations et de 15 comportements sur une échelle de 0 à 9, où 0 = "extrêmement approprié" et 9 = "extrêmement inapproprié". En effectuant la moyenne des résultats de l'ensemble des individus, on constate une certaine différence entre les valeurs.

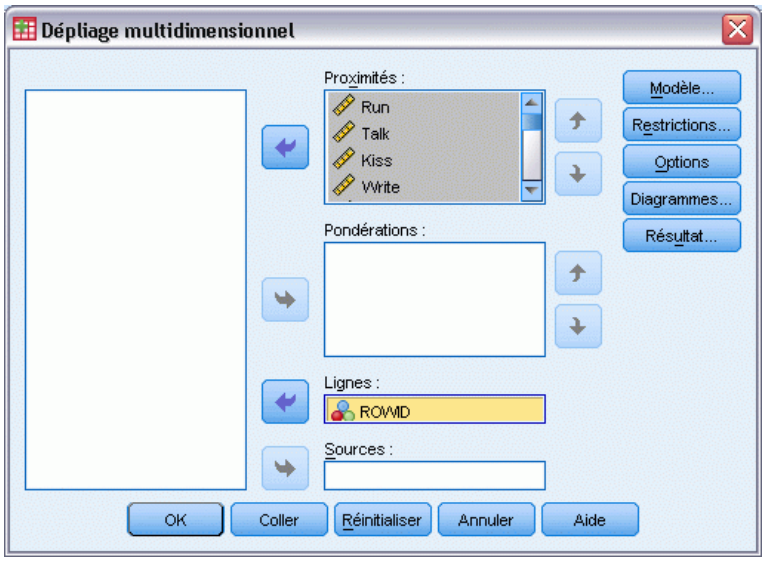
Ces informations sont regroupées dans le fichier *behavior.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*. Utilisez le dépliage multidimensionnel pour établir des groupes de situations similaires et les comportements avec lesquels elles sont le plus souvent associées. La syntaxe servant à reproduire ces analyses se trouve dans *prefscal_behavior.sps*.

Exécution de l'analyse

- Pour lancer une analyse Dépliage multidimensionnel, choisissez les options suivantes dans les menus :

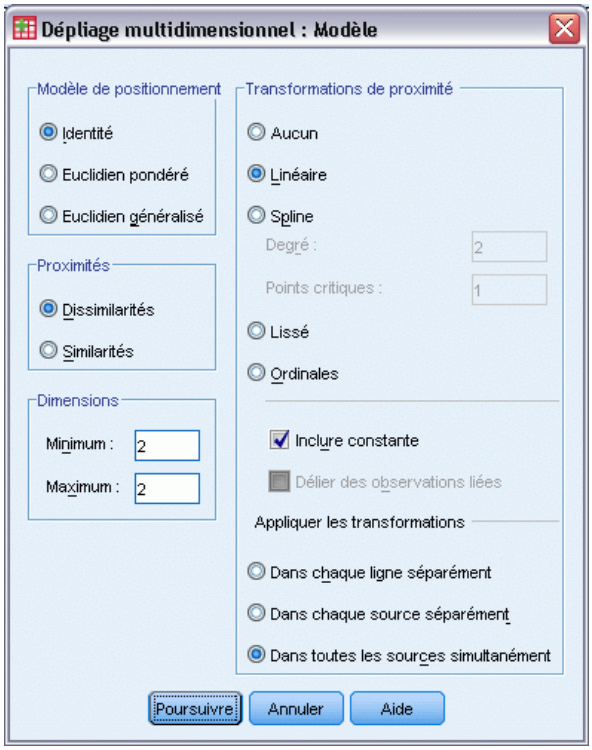
  Analyse > Echelle > Dépliage multidimensionnel (PREFSCAL)...

Figure 15-24
Boîte de dialogue principale Dépliage multidimensionnel



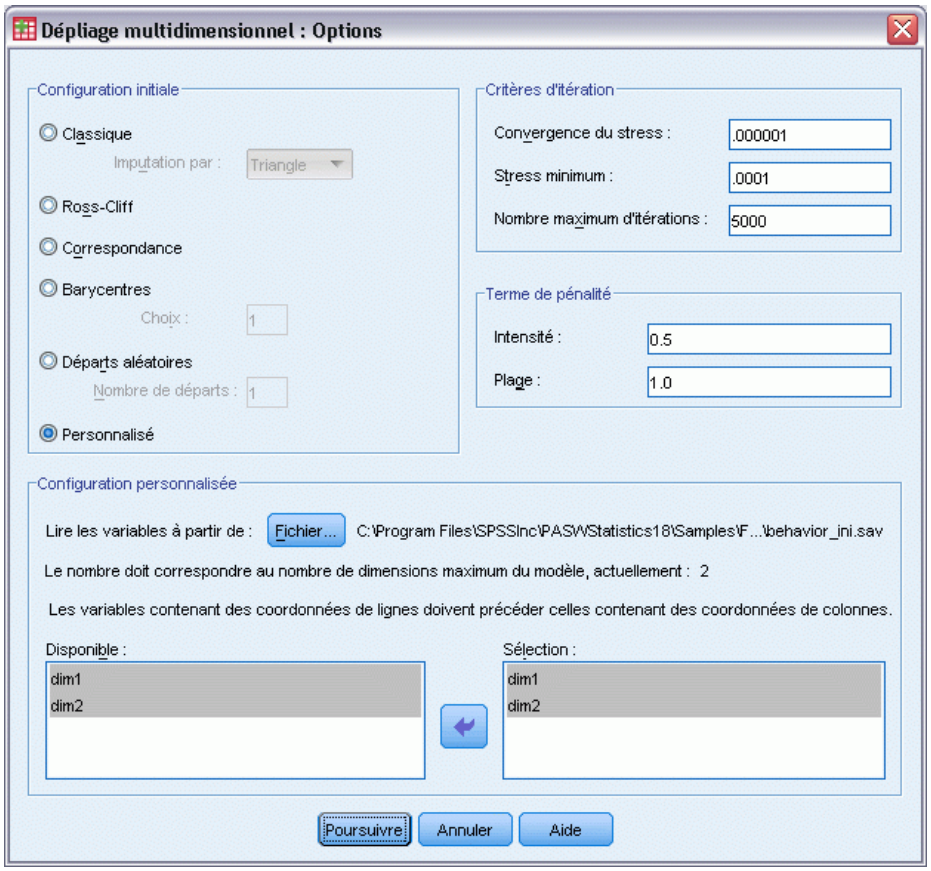
- Sélectionnez les options allant de *Courir* à *Crier* comme variables de proximité.
- Sélectionnez *ROWID* comme variable de ligne.
- Cliquez sur Modèle.

Figure 15-25
Boîte de dialogue Modèle



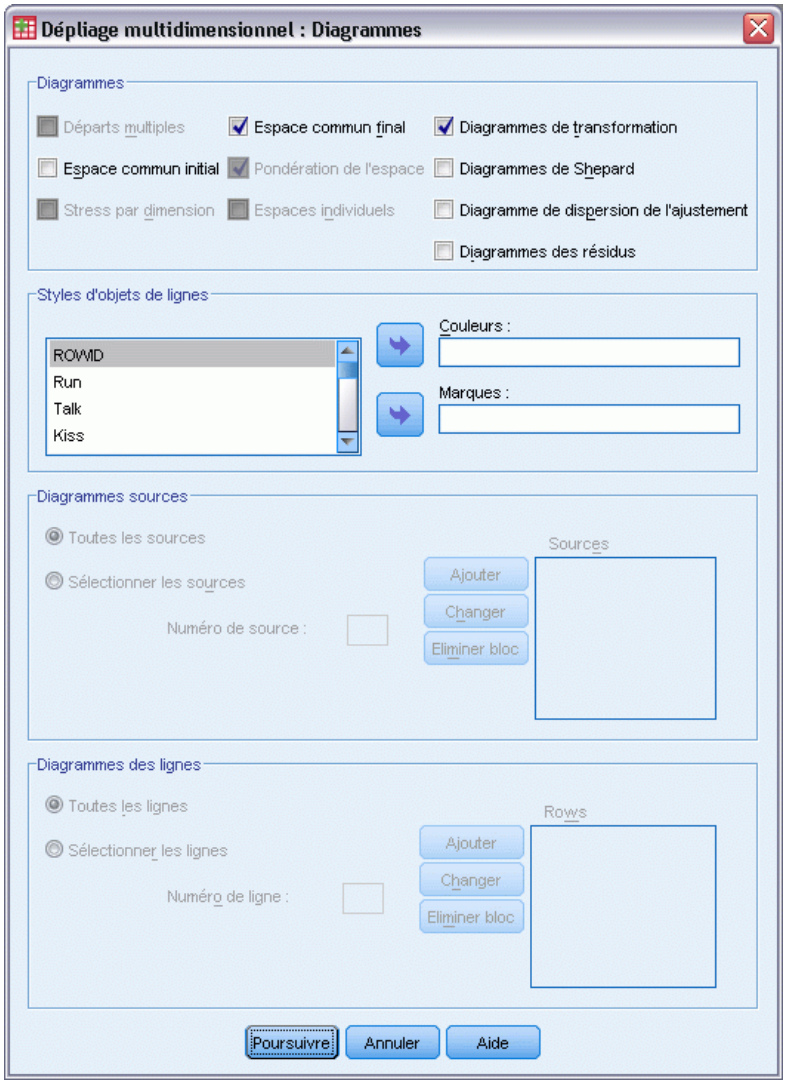
- Sélectionnez Linéaire comme transformation de proximité et choisissez Inclure une constante.
- Choisissez d'appliquer les transformations Dans toutes les sources simultanément.
- Cliquez sur Poursuivre.
- Cliquez sur Options dans la boîte de dialogue Dépliage multidimensionnel.

Figure 15-26
Options



- Sélectionnez Personnalisée dans le groupe Configuration initiale.
- Accédez au fichier *behavior_ini.sav* et choisissez-le comme fichier contenant la configuration initiale personnalisée. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Categories 20*.
- Sélectionnez *dim1* et *dim2* comme variables spécifiant la configuration initiale.
- Cliquez sur Poursuivre.
- Cliquez sur Diagrammes dans la boîte de dialogue Dépliage multidimensionnel.

Figure 15-27
Boîte de dialogue Diagrammes



- Sélectionnez Diagrammes de transformation dans le groupe Diagrammes.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Dépliage multidimensionnel.

Voici la syntaxe de commande générée par ces sélections :

```
PREFSCAL
  VARIABLES=Run Talk Kiss Write Eat Sleep Mumble Read Fight Belch Argue Jump
  Cry Laugh Shout
  /INPUT=ROWS(ROWID )
  /INITIAL=( 'samplesDirectory/behavior_ini.sav' )
  dim1 dim2
  /CONDITION=UNCONDITIONAL
  /TRANSFORMATION=LINEAR (INTERCEPT)
  /PROXIMITIES=DISSIMILARITIES
```

```
/MODEL=IDENTITY
/CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
MAXITER(5000)
/PENALTY=LAMBDA(1.0) OMEGA(0.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON TRANSFORMATIONS .
```

- Cette syntaxe spécifie une analyse des variables *courir* à *crier*. La variable *idligne* est utilisée pour identifier les lignes.
- La sous-commande `INITIAL` spécifie que les valeurs de départ doivent être tirées du fichier *behavior_ini.sav*. Les coordonnées des lignes et des colonnes sont empilées, avec les coordonnées des colonnes à la suite des coordonnées des lignes.
- La sous-commande `CONDITION` spécifie que toutes les proximités peuvent être comparées entre elles. Ceci est vérifié pour cette analyse : en comparant les proximités obtenues pour les comportements Courir dans un parc et Courir dans une église vous observez que l'un des deux comportements est considéré moins approprié que l'autre.
- La sous-commande `TRANSFORMATION` indique une transformation linéaire des proximités avec constante. Ceci est approprié si une différence de 1 point dans les proximités est observée dans tout l'intervalle. En d'autres termes, si les étudiants ont attribué leurs notes de manière à ce que la différence entre 0 et 1 est la même que la différence entre 5 et 6, alors une transformation linéaire est appropriée.
- La sous-commande `PLOT` demande des diagrammes de l'espace commun et des diagrammes de transformation.
- Tous les autres paramètres sont réinitialisés à leur valeur par défaut.

Mesures

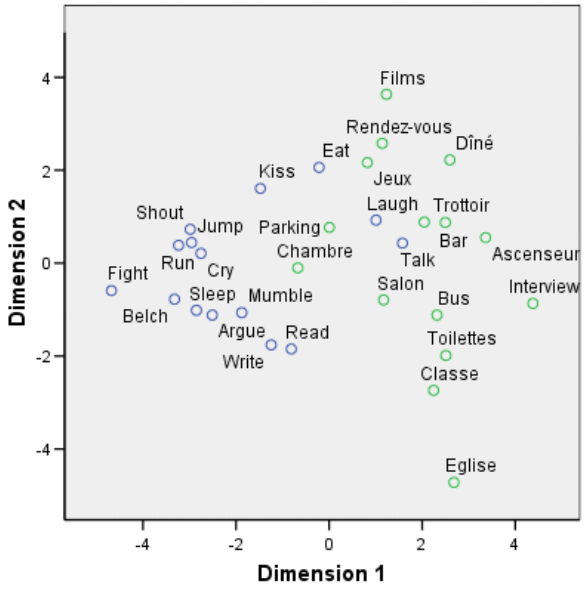
Figure 15-28
Mesures

Itérations		169
Valeur de fonction finale		,6427725
Eléments de valeur de fonction	Elément de stress	,1900001
	Elément de pénalité	2,1745069
Médiocrité d'ajustement	Stress normalisé	,0361000
	Stress-I de Kruskal	,1900001
	Stress-II de Kruskal	,5224668
	S-Stress-I de Young	,2760971
	S-Stress-II de Young	,4525933
Qualité de l'ajustement :	Dispersion représentée	,9639000
	Variance expliquée par	,8082862
	Ordres de préférence récupérés	,8608333
	Rho de Spearman	,8981120
	Tau-b de Kendall	,7202452
Coefficients de variation	Proximités de variation	,5138436
	Proximités de variation transformées	,4751934
	Distances de variation	,3912592
Indices de dégénérescence	Somme des carrés des indices de fusion de DeSarbo	,4957969
	Index de non-dégénérescence simple de Shepard	,7173810

L'algorithme converge après 169 itérations, avec une mesure du stress pénalisé finale de 0,6427725. Les coefficients de variation et l'index de Shepard sont suffisamment élevés et les indices de DeSarbo suffisamment bas pour suggérer qu'il n'existe aucun problème de dégénérescence.

Espace commun

Figure 15-29
Diagramme join de l'espace commun

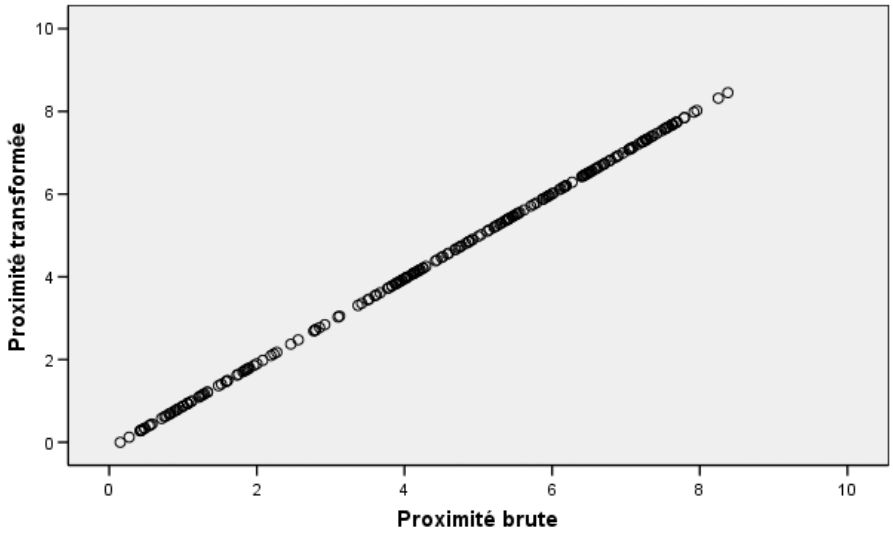


La dimension horizontale apparaît plus fortement associée aux objets de colonnes (comportements) et établit une discrimination entre les comportements inappropriés (se battre, roter) et les comportements plus appropriés. La dimension verticale apparaît plus fortement associée aux objets de lignes (situations) et définit différentes restrictions parmi les relations comportement-situation établies.

- Dans la partie inférieure de la dimension verticale se situent les situations (à l'église, en classe) restreintes aux types de comportements plus calmes et introspectifs (lire, écrire). Ainsi, ces comportements sont contenus dans la partie inférieure de l'axe vertical.
- Dans la partie supérieure de la dimension verticale se trouvent les situations (films, jeux, rendez-vous) restreintes aux types de comportements sociables/extrovertis (manger, embrasser, rire). Ainsi, ces comportements sont contenus dans la partie supérieure de l'axe vertical.
- Au centre de la dimension verticale, les situations sont réparties distinctivement le long de la dimension horizontale selon le caractère restrictif de la situation. Les situations les plus éloignées des comportements (en entretien) sont les plus restrictives, alors que celles les proches des comportements (dans la chambre, au parc) sont généralement les moins restrictives.

Transformations de proximité

Figure 15-30
Diagramme de transformation



Inconditionnel Linéaire transformation avec ordonnée

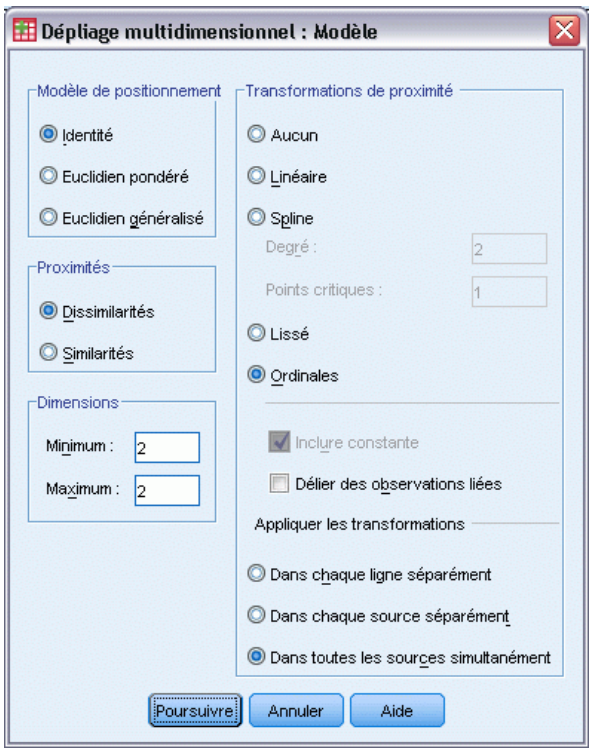
Les proximités étaient traitées comme linéaires dans cette analyse, de manière à ce que le diagramme représentant les valeurs transformées en fonction des proximités d'origine forme une ligne droite. L'ajustement de cette solution est bon, mais un meilleur ajustement peut être obtenu par une transformation différente des proximités.

Modification de la transformation des proximités (ordinale)

- Pour produire une solution avec transformation ordinale des proximités, cliquez sur l'outil Rappeler boîte de dialogue et sélectionnez Dépliage multidimensionnel.

▶ Cliquez sur Modèle dans la boîte de dialogue Dépliage multidimensionnel.

Figure 15-31
Boîte de dialogue Modèle



▶ Sélectionnez l'option Ordinal comme transformation de proximités.

▶ Cliquez sur Poursuivre.

▶ Cliquez sur OK dans la boîte de dialogue Dépliage multidimensionnel.

Voici la syntaxe de commande générée par ces sélections :

```
PREFSCAL
  VARIABLES=Run Talk Kiss Write Eat Sleep Mumble Read Fight Belch Argue Jump
  Cry Laugh Shout
  /INPUT=ROWS(ROWID )
  /INITIAL=( 'samplesDirectory/behavior_ini.sav' )
  dim1 dim2
  /CONDITION=UNCONDITIONAL
  /TRANSFORMATION=LINEAR (INTERCEPT)
  /PROXIMITIES=DISSIMILARITIES
  /MODEL=IDENTITY
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(1.0) OMEGA(0.0)
  /PRINT=MEASURES COMMON
  /PLOT=COMMON TRANSFORMATIONS .
```

- La seule différence réside dans la sous-commande TRANSFORMATION. La transformation est définie sur ORDINAL, ce qui préserve l'ordre des proximités mais ne nécessite pas que les valeurs transformées soient proportionnelles aux valeurs d'origine.

Mesures

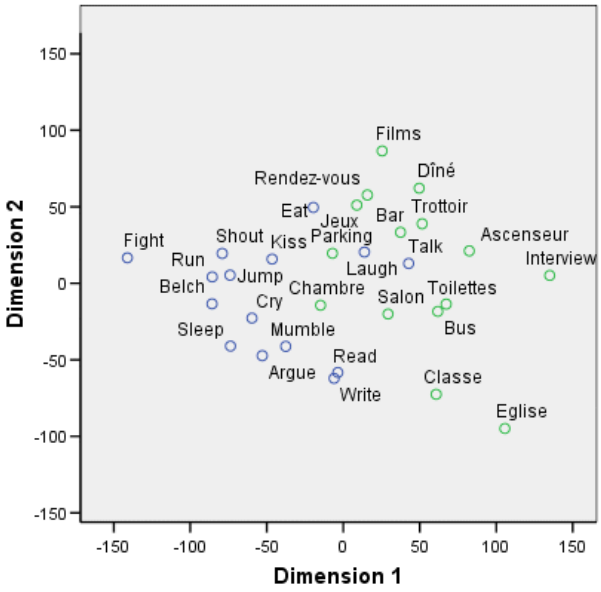
Figure 15-32
Mesures de la solution avec transformation ordinale

Itérations		268
Valeur de fonction finale		,6044671
Eléments de valeur de fonction	Elément de stress	,1747239
	Elément de pénalité	2,0911875
Médiocrité d'ajustement	Stress normalisé	,0305285
	Stress-I de Kruskal	,1747239
	Stress-II de Kruskal	,4444641
	S-Stress-I de Young	,2707147
	S-Stress-II de Young	,3978003
Qualité de l'ajustement :	Dispersion représentée	,9694715
	Variance expliquée par	,8454488
	Ordres de préférence récupérés	,8574206
	Rho de Spearman	,9032676
	Tau-b de Kendall	,7532788
Coefficients de variation	Proximités de variation	,5138436
	Proximités de variation transformées	,4930018
	Distances de variation	,4284849
Indices de dégénérescence	Somme des carrés des indices de fusion de DeSarbo	,3610680
	Index de non-dégénérescence simple de Shepard	,7469048

L'algorithme converge après 268 itérations, avec une mesure du stress pénalisé finale de 0.6044671. Cette statistique et les autres mesures sont légèrement meilleures pour cette solution que pour celle obtenue par transformation linéaire des proximités.

Espace commun

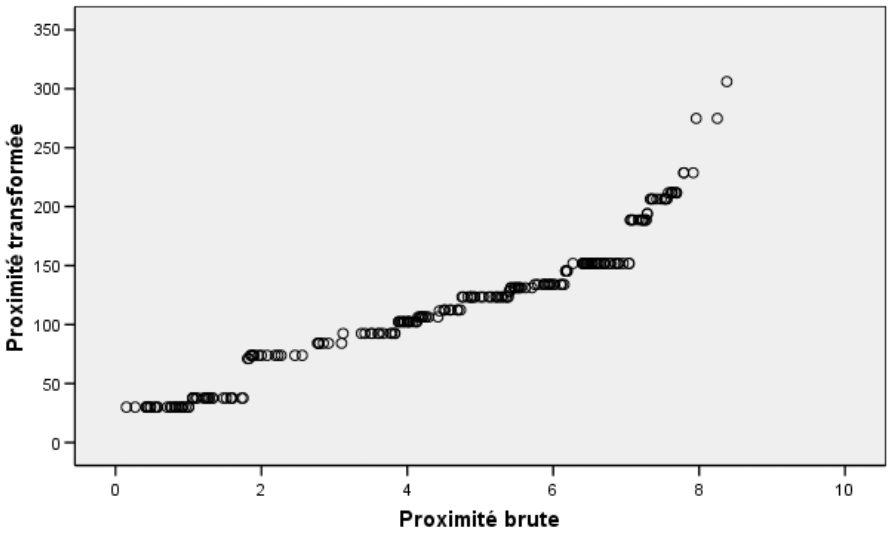
Figure 15-33
Diagramme joint de l'espace commun pour la solution avec transformation ordinale



L'interprétation de l'espace commun est la même pour les deux solutions. Cette solution (transformation ordinale) présente une variation relativement plus faible sur la dimension verticale que sur la dimension horizontale contrairement à la solution avec transformation linéaire.

Transformations de proximité

Figure 15-34
Diagramme de transformation de la solution avec transformation ordinale



A l'exception des valeurs aux proximités les plus élevées, qui se distinguent du reste des valeurs, la transformation ordinale des proximités est relativement linéaire. Ces proximités élevées non-linéaires constituent la principale différence entre les solutions ordinale et linéaire ; cependant, nous ne disposons pas de suffisamment d'informations pour déteminer si cette tendance non-linéaire dans les valeurs les plus élevées s'avère être une tendance vérifiée ou une anomalie.

Lectures recommandées

Reportez-vous aux écrits suivants pour plus d'informations :

Busing, F. M. T. A., P. J. F. Groenen, et W. J. Heiser. 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, 70, .

Green, P. E., et V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Price, R. H., et D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Annexe

A

Fichiers d'exemple

Les fichiers d'exemple installés avec le produit figurent dans le sous-répertoire *Echantillons* du répertoire d'installation. Il existe un dossier distinct au sein du sous-répertoire Echantillons pour chacune des langues suivantes : Anglais, Français, Allemand, Italien, Japonais, Coréen, Polonais, Russe, Chinois simplifié, Espangol et Chinois traditionnel.

Seuls quelques fichiers d'exemples sont disponibles dans toutes les langues. Si un fichier d'exemple n'est pas disponible dans une langue, le dossier de langue contient la version anglaise du fichier d'exemple.

Descriptions

Voici de brèves descriptions des fichiers d'exemple utilisés dans divers exemples à travers la documentation.

- **accidents.sav.**Ce fichier de données d'hypothèse concerne une société d'assurance qui étudie les facteurs de risque liés à l'âge et au sexe dans les accidents de la route survenant dans une région donnée. Chaque observation correspond à une classification croisée de la catégorie d'âge et du sexe.
- **adl.sav.** Ce fichier de données d'hypothèse concerne les mesures entreprises pour identifier les avantages d'un type de thérapie proposé aux patients qui ont subi une attaque cardiaque. Les médecins ont assigné de manière aléatoire les patients du sexe féminin ayant subi une attaque cardiaque à un groupe parmi deux groupes possibles. Le premier groupe a fait l'objet de la thérapie standard tandis que le second a bénéficié en plus d'une thérapie émotionnelle. Trois mois après les traitements, les capacités de chaque patient à effectuer les tâches ordinaires de la vie quotidienne ont été notées en tant que variables ordinales.
- **advert.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un détaillant pour examiner la relation existant entre l'argent dépensé dans la publicité et les ventes résultantes. Pour ce faire, il collecte les chiffres des ventes passées et les coûts associés à la publicité.
- **aflatoxin.sav.** Ce fichier de données d'hypothèse concerne le test de l'aflatoxine dans des récoltes de maïs. La concentration de ce poison varie largement d'une récolte à l'autre et au sein de chaque récolte. Un processeur de grain a reçu 16 échantillons issus de 8 récoltes de maïs et a mesuré les niveaux d'alfatoxine en parties par milliard (PPB).
- **anorectic.sav.** En cherchant à développer une symptomatologie standardisée du comportement anorexique/boulimique, des chercheurs(Van der Ham, Meulman, Van Strien, et Van Engeland, 1997) ont examiné 55 adolescents souffrant de troubles alimentaires. Chaque patient a été observé quatre fois sur une période de quatre années, soit un total de 220 observations. A chaque observation, les patients ont été notés pour chacun des 16 symptômes. En raison de l'absence de scores de symptôme pour le patient 71/visite 2, le patient 76/visite 2 et le patient 47/visite 3, le nombre d'observations valides est de 217.

- **bankloan.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une banque pour réduire le taux de défaut de paiement. Il contient des informations financières et démographiques sur 850 clients existants et éventuels. Les premières 700 observations concernent des clients auxquels des prêts ont été octroyés. Les 150 dernières observations correspondant aux clients éventuels que la banque doit classer comme bons ou mauvais risques de crédit.
- **bankloan_binning.sav.** Ce fichier de données d'hypothèse concerne des informations financières et démographiques sur 5 000 clients existants.
- **behavior.sav.** Dans un exemple classique (Price et Bouffard, 1974), on a demandé à 52 étudiants de noter les combinaisons établies à partir de 15 situations et de 15 comportements sur une échelle de 0 à 9, où 0 = « extrêmement approprié » et 9 = « extrêmement inapproprié ». En effectuant la moyenne des résultats de l'ensemble des individus, on constate une certaine différence entre les valeurs.
- **behavior_ini.sav.** Ce fichier de données contient la configuration initiale d'une solution bidimensionnelle pour *behavior.sav*.
- **brakes.sav.** Ce fichier de données d'hypothèse concerne le contrôle qualité effectué dans une usine qui fabrique des freins à disque pour des voitures haut de gamme. Le fichier de données contient les mesures de diamètre de 16 disques de 8 machines de production. Le diamètre cible des freins est de 322 millimètres.
- **breakfast.sav.** Au cours d'une étude classique (Green et Rao, 1972), on a demandé à 21 étudiants en MBA (Master of Business Administration) de l'école de Wharton et à leurs conjoints de classer 15 aliments du petit-déjeuner selon leurs préférences, de 1= « aliment préféré » à 15= « aliment le moins apprécié ». Leurs préférences ont été enregistrées dans six scénarios différents, allant de « Préférence générale » à « En-cas avec boisson uniquement ».
- **breakfast-overall.sav.** Ce fichier de données contient les préférences de petit-déjeuner du premier scénario uniquement, « Préférence générale ».
- **broadband_1.sav.** Ce fichier de données d'hypothèse concerne le nombre d'abonnés, par région, à un service haut débit. Le fichier de données contient le nombre d'abonnés mensuels de 85 régions sur une période de quatre ans.
- **broadband_2.sav.** Ce fichier de données est identique au fichier *broadband_1.sav* mais contient les données relatives à trois mois supplémentaires.
- **car_insurance_claims.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs (McCullagh et Nelder, 1989) qui concerne des actions en indemnisation pour des voitures. Le montant d'action en indemnisation moyen peut être modelé comme présentant une distribution gamma, à l'aide d'une fonction de lien inverse pour associer la moyenne de la variable dépendante à une combinaison linéaire de l'âge de l'assuré, du type de véhicule et de l'âge du véhicule. Le nombre d'actions entreprises peut être utilisé comme pondération de positionnement.
- **car_sales.sav.** Ce fichier de données contient des estimations de ventes hypothétiques, des barèmes de prix et des spécifications physiques concernant divers modèles et marques de véhicule. Les barèmes de prix et les spécifications physiques proviennent tour à tour de *edmunds.com* et des sites des constructeurs.
- **car_sales_uprepared.sav.** Il s'agit d'une version modifiée de *car_sales.sav* qui n'inclut aucune version transformée des champs.

- **carpet.sav.** Dans un exemple courant (Green et Wind, 1973), une société intéressée par la commercialisation d'un nouveau nettoyeur de tapis souhaite examiner l'influence de cinq critères sur la préférence du consommateur : la conception du conditionnement, la marque, le prix, une étiquette *Economique* et une garantie satisfait ou remboursé. Il existe trois niveaux de critère pour la conception du conditionnement, suivant l'emplacement de l'applicateur, trois marques (*K2R*, *Glory* et *Bissell*), trois niveaux de prix et deux niveaux (non ou oui) pour chacun des deux derniers critères. Dix consommateurs classent 22 profils définis par ces critères. La variable *Préférence* indique le classement des rangs moyens de chaque profil. Un rang faible correspond à une préférence élevée. Cette variable reflète une mesure globale de préférence pour chaque profil.
- **carpet_prefs.sav.** Ce fichier de données repose sur le même exemple que celui décrit pour *carpet.sav*, mais contient les classements réels issus de chacun des 10 clients. On a demandé aux consommateurs de classer les 22 profils de produits, du préféré au moins intéressant. Les variables *PREF1* à *PREF22* contiennent les identificateurs des profils associés, tels qu'ils sont définis dans *carpet_plan.sav*.
- **catalog.sav.** Ce fichier de données contient des chiffres de ventes mensuelles hypothétiques relatifs à trois produits vendus par une entreprise de vente par correspondance. Les données relatives à cinq variables explicatives possibles sont également incluses.
- **catalog_seasfac.sav.** Ce fichier de données est identique à *catalog.sav* mais contient en plus un ensemble de facteurs saisonniers calculés à partir de la procédure de désaisonnalisation, ainsi que les variables de date correspondantes.
- **cellular.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un opérateur téléphonique pour réduire les taux de désabonnement. Des scores de propension au désabonnement sont attribués aux comptes, de 0 à 100. Les comptes ayant une note égale ou supérieure à 50 sont susceptibles de changer de fournisseur.
- **ceramics.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fabricant pour déterminer si un nouvel alliage haute qualité résiste mieux à la chaleur qu'un alliage standard. Chaque observation représente un test séparé de l'un des deux alliages ; le degré de chaleur auquel l'alliage ne résiste pas est enregistré.
- **cereal.sav.** Ce fichier de données d'hypothèse concerne un sondage de 880 personnes interrogées sur leurs préférences de petit-déjeuner et sur leur âge, leur sexe, leur situation familiale et leur mode de vie (actif ou non actif, selon qu'elles pratiquent une activité physique au moins deux fois par semaine). Chaque observation correspond à un répondant distinct.
- **clothing_defects.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de textile. Dans chaque lot produit à l'usine, les inspecteurs prélèvent un échantillon de vêtements et comptent le nombre de vêtements qui ne sont pas acceptables.
- **coffee.sav.** Ce fichier de données concerne l'image perçue de six marques de café frappé (Kennedy, Riquier, et Sharp, 1996). Pour chacun des 23 attributs d'image de café frappé, les personnes sollicitées ont sélectionné toutes les marques décrites par l'attribut. Les six marques sont appelées AA, BB, CC, DD, EE et FF à des fins de confidentialité.
- **contacts.sav.** Ce fichier de données d'hypothèse concerne les listes de contacts d'un groupe de représentants en informatique d'entreprise. Chaque contact est classé selon le service de l'entreprise où il travaille et le classement de son entreprise. Sont également enregistrés le

montant de la dernière vente effectuée, le temps passé depuis la dernière vente et la taille de l'entreprise du contact.

- **creditpromo.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un grand magasin pour évaluer l'efficacité d'une promotion récente de carte de crédit. A cette fin, 500 détenteurs de carte ont été sélectionnés au hasard. La moitié a reçu une publicité faisant la promotion d'un taux d'intérêt réduit sur les achats effectués dans les trois mois à venir. L'autre moitié a reçu une publicité saisonnière standard.
- **customer_dbase.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour utiliser les informations figurant dans sa banque de données et proposer des offres spéciales aux clients susceptibles d'être intéressés. Un sous-groupe de la base de clients a été sélectionné au hasard et a reçu des offres spéciales. Les réponses des clients ont été enregistrées.
- **customer_information.sav.** Un fichier de données d'hypothèse qui contient les informations postales du client, telles que le nom et l'adresse.
- **customer_subset.sav.** Un sous-ensemble de 80 observations de *customer_dbase.sav*.
- **debate.sav.** Ce fichier de données d'hypothèse concerne des réponses appariées à une enquête donnée aux participants à un débat politique avant et après le débat. Chaque observation représente un répondant distinct.
- **debate_aggregate.sav.** Il s'agit d'un fichier de données d'hypothèse qui rassemble les réponses dans le fichier *debate.sav*. Chaque observation correspond à une classification croisée de préférence avant et après le débat.
- **demo.sav.** Ce fichier de données d'hypothèse concerne une base de données clients achetée en vue de diffuser des offres mensuelles. Les données indiquent si le client a répondu ou non à l'offre et contiennent diverses informations démographiques.
- **demo_cs_1.sav.** Ce fichier de données d'hypothèse concerne la première mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à une ville différente. La région, la province, le quartier et la ville sont enregistrés.
- **demo_cs_2.sav.** Ce fichier de données d'hypothèse concerne la seconde mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à un ménage différent issu des villes sélectionnées à la première étape. La région, la province, le quartier, la ville, la sous-division et l'identification sont enregistrés. Les informations d'échantillonnage des deux premières étapes de la conception sont également incluses.
- **demo_cs.sav.** Ce fichier de données d'hypothèse concerne des informations d'enquête collectées via une méthode complexe d'échantillonnage. Chaque observation correspond à un ménage différent et diverses informations géographiques et d'échantillonnage sont enregistrées.
- **dmdata.sav.** Ceci est un fichier de données d'hypothèse qui contient des informations démographiques et des informations concernant les achats pour une entreprise de marketing direct. *dmdata2.sav* contient les informations pour un sous-ensemble de contacts qui ont reçu un envoi d'essai, et *dmdata3.sav* contient des informations sur les contacts restants qui n'ont pas reçu l'envoi d'essai.

- **dietstudy.sav.** Ce fichier de données d'hypothèse contient les résultats d'une étude portant sur le régime de Stillman(Rickman, Mitchell, Dingman, et Dalen, 1974). Chaque observation correspond à un sujet distinct et enregistre son poids en livres avant et après le régime, ainsi que ses niveaux de triglycérides en mg/100 ml.
- **dvdplayer.sav.** Ce fichier de données d'hypothèse concerne le développement d'un nouveau lecteur DVD. A l'aide d'un prototype, l'équipe de marketing a collecté des données de groupes spécifiques. Chaque observation correspond à un utilisateur interrogé et enregistre des informations démographiques sur cet utilisateur, ainsi que ses réponses aux questions portant sur le prototype.
- **german_credit.sav.** Ce fichier de données provient de l'ensemble de données « German credit » figurant dans le référentiel Machine Learning Databases (Blake et Merz, 1998) de l'université de Californie, Irvine.
- **grocery_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *grocery_coupons.sav* dans lequel les achats hebdomadaires sont organisés par client distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, le montant dépensé enregistré est à présent la somme des montants dépensés au cours des quatre semaines de l'enquête.
- **grocery_coupons.sav.** Il s'agit d'un fichier de données d'hypothèse qui contient des données d'enquête collectées par une chaîne de magasins d'alimentation qui chercher à déterminer les habitudes de consommation de ses clients. Chaque client est suivi pendant quatre semaines et chaque observation correspond à une semaine distincte. Les informations enregistrées concernent les endroits où le client effectue ses achats, la manière dont il les effectue, ainsi que les sommes dépensées en provisions au cours de cette semaine.
- **guttman.sav.** Bell (Bell, 1961) a présenté un tableau pour illustrer les groupes sociaux possibles. Guttman (Guttman, 1968) a utilisé une partie de ce tableau, dans lequel cinq variables décrivant des éléments tels que l'interaction sociale, le sentiment d'appartenance à un groupe, la proximité physique des membres et la formalité de la relation, ont été croisées avec sept groupes sociaux théoriques, dont les foules (par exemple, le public d'un match de football), l'audience (par exemple, au cinéma ou dans une salle de classe), le public (par exemple, les journaux ou la télévision), les bandes (proche d'une foule, mais qui serait caractérisée par une interaction beaucoup plus intense), les groupes primaires (intimes), les groupes secondaires (volontaires) et la communauté moderne (groupement lâche issu d'une forte proximité physique et d'un besoin de services spécialisés).
- **health_funding.sav.** Ce fichier de données d'hypothèse concerne des données sur le financement des soins de santé (montant par groupe de 100 individus), les taux de maladie (taux par groupe de 10 000 individus) et les visites chez les prestataires de soins de santé (taux par groupe de 10 000 individus). Chaque observation représente une ville différente.
- **hivassay.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un laboratoire pharmaceutique pour développer une analyse rapide de détection d'infection HIV. L'analyse a pour résultat huit nuances de rouge, les nuances les plus marquées indiquant une plus forte probabilité d'infection. Un test en laboratoire a été effectué sur 2 000 échantillons de sang, la moitié de ces échantillons étant infectée par le virus HIV et l'autre moitié étant saine.
- **hourlywagedata.sav.** Ce fichier de données d'hypothèse concerne les salaires horaires d'infirmières occupant des postes administratifs et dans les services de soins, et affichant divers niveaux d'expérience.

- **insurance_claims.sav.** Il s'agit d'un fichier de données hypothétiques qui concerne une compagnie d'assurance souhaitant développer un modèle pour signaler des réclamations suspectes, potentiellement frauduleuses. Chaque observation correspond à une réclamation distincte.
- **insure.sav.** Ce fichier de données d'hypothèse concerne une compagnie d'assurance qui étudie les facteurs de risque indiquant si un client sera amené à déclarer un incident au cours d'un contrat d'assurance vie d'une durée de 10 ans. Chaque observation figurant dans le fichier de données représente deux contrats, l'un ayant enregistré une réclamation et l'autre non, appariés par âge et sexe.
- **judges.sav.** Ce fichier de données d'hypothèse concerne les scores attribués par des juges expérimentés (plus un juge enthousiaste) à 300 performances de gymnastique. Chaque ligne représente une performance distincte ; les juges ont examiné les mêmes performances.
- **kinship_dat.sav.** Rosenberg et Kim (Rosenberg et Kim, 1975) se sont lancés dans l'analyse de 15 termes de parenté (cousin/cousine, fille, fils, frère, grand-mère, grand-père, mère, neveu, nièce, oncle, père, petite-fille, petit-fils, sœur, tante). Ils ont demandé à quatre groupes d'étudiants (deux groupes de femmes et deux groupes d'hommes) de trier ces termes en fonction des similarités. Deux groupes (un groupe de femmes et un groupe d'hommes) ont été invités à effectuer deux tris, en basant le second sur un autre critère que le premier. Ainsi, un total de six "sources" a été obtenu. Chaque source correspond à une matrice de proximité 15×15, dont le nombre de cellules est égal au nombre de personnes dans une source moins le nombre de fois où les objets ont été partitionnés dans cette source.
- **kinship_ini.sav.** Ce fichier de données contient une configuration initiale d'une solution tridimensionnelle pour *kinship_dat.sav*.
- **kinship_var.sav.** Ce fichier de données contient les variables indépendantes *sexe*, *génér*(ation) et *degré* (de séparation) permettant d'interpréter les dimensions d'une solution pour *kinship_dat.sav*. Elles permettent en particulier de réduire l'espace de la solution à une combinaison linéaire de ces variables.
- **marketvalues.sav.** Ce fichier de données concerne les ventes de maisons dans un nouvel ensemble à Algonquin (Illinois) au cours des années 1999–2000. Ces ventes relèvent des archives publiques.
- **nhis2000_subset.sav.**Le NHIS (National Health Interview Survey) est une enquête de grande envergure concernant la population des Etats-Unis. Des entretiens ont lieu avec un échantillon de ménages représentatifs de la population américaine. Des informations démographiques et des observations sur l'état de santé et le comportement sanitaire sont recueillies auprès des membres de chaque ménage. Ce fichier de données contient un sous-groupe d'informations issues de l'enquête de 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Fichier de données et documentation d'usage public. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accès en 2003.
- **ozone.sav.** Les données incluent 330 observations portant sur six variables météorologiques pour prévoir la concentration d'ozone à partir des variables restantes. Des chercheurs précédents (Breiman et Friedman, 1985), (Hastie et Tibshirani, 1990), ont décelé parmi ces variables des non-linéarités qui pénalisent les approches standard de la régression.

- **pain_medication.sav.** Ce fichier de données d'hypothèse contient les résultats d'un essai clinique d'un remède anti-inflammatoire traitant les douleurs de l'arthrite chronique. On cherche notamment à déterminer le temps nécessaire au médicament pour agir et les résultats qu'il permet d'obtenir par rapport à un médicament existant.
- **patient_los.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux de patients admis à l'hôpital pour suspicion d'infarctus du myocarde suspecté (ou « attaque cardiaque »). Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **patlos_sample.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux d'un échantillon de patients sous traitement thrombolytique après un infarctus du myocarde. Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **poll_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un enquêteur pour déterminer le niveau de soutien du public pour un projet de loi avant législature. Les observations correspondent à des électeurs enregistrés. Chaque observation enregistre le comté, la ville et le quartier où habite l'électeur.
- **poll_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des électeurs répertoriés dans le fichier *poll_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *poll.csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. Toutefois, ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS – Probability-Proportional-to-Size), il existe également un fichier contenant les probabilités de sélection conjointes (*poll_jointprob.sav*). Les variables supplémentaires correspondant à la répartition démographique des électeurs et à leur opinion sur le projet de loi proposé ont été collectées et ajoutées au fichier de données une fois l'échantillon prélevé.
- **property_assess.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur au niveau du comté pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés vendues dans le comté au cours de l'année précédente. Chaque observation du fichier de données enregistre la ville où se trouve la propriété, l'évaluateur ayant visité la propriété pour la dernière fois, le temps écoulé depuis cette évaluation, l'évaluation effectuée à ce moment-là et la valeur de vente de la propriété.
- **property_assess_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur du gouvernement pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés de l'état. Chaque observation du fichier de données enregistre le comté, la ville et le quartier où se trouve la propriété, le temps écoulé depuis la dernière évaluation et l'évaluation alors effectuée.
- **property_assess_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des propriétés répertoriées dans le fichier *property_assess_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *property_assess.csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. La variable supplémentaire *Valeur courante* a été collectée et ajoutée au fichier de données une fois l'échantillon prélevé.

- **recidivism.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis, ainsi que le temps écoulé jusqu'à la seconde arrestation si elle s'est produite dans les deux années suivant la première.
- **recidivism_cs_sample.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste libéré suite à la première arrestation en juin 2003 et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis et les données relatives à la seconde arrestation, si elle a eu lieu avant fin juin 2006. Les récidivistes ont été choisis dans plusieurs départements échantillonnés conformément au plan d'échantillonnage spécifié dans *recidivism_cs.csplan*. Ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS - Probability proportional to size), il existe également un fichier contenant les probabilités de sélection conjointes (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.**Un fichier de données d'hypothèse qui contient les données de transaction d'achat, y compris la date d'achat, le/les élément(s) acheté(s) et le montant monétaire pour chaque transaction.
- **salesperformance.sav.** Ce fichier de données d'hypothèse concerne l'évaluation de deux nouveaux cours de formation en vente. Soixante employés, divisés en trois groupes, reçoivent chacun une formation standard. En outre, le groupe 2 suit une formation technique et le groupe 3 un didacticiel pratique. A l'issue du cours de formation, chaque employé est testé et sa note enregistrée. Chaque observation du fichier de données représente un stagiaire distinct et enregistre le groupe auquel il a été assigné et la note qu'il a obtenue au test.
- **satisf.sav.** Il s'agit d'un fichier de données d'hypothèse portant sur une enquête de satisfaction effectuée par une société de vente au détail au niveau de quatre magasins. Un total de 582 clients ont été interrogés et chaque observation représente la réponse d'un seul client.
- **screws.sav.**Ce fichier de données contient des informations sur les descriptives des vis, des boulons, des écrous et des clous.(Hartigan, 1975).
- **shampoo_ph.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de produits capillaires. A intervalles réguliers, six lots de sortie distincts sont mesurés et leur pH enregistré. La plage cible est 4,5–5,5.
- **ships.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs (McCullagh et al., 1989) et concernant les dommages causés à des cargos par les vagues. Les effectifs d'incidents peuvent être modélisés comme des incidents se produisant selon un taux de Poisson en fonction du type de navire, de la période de construction et de la période de service. Les mois de service totalisés pour chaque cellule du tableau formé par la classification croisée des facteurs fournissent les valeurs d'exposition au risque.
- **site.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour choisir de nouveaux sites pour le développement de ses activités. L'entreprise a fait appel à deux consultants pour évaluer séparément les sites. Ces consultants, en plus de fournir un rapport approfondi, ont classé chaque site comme constituant une éventualité « bonne », « moyenne » ou « faible ».

- **smokers.sav.** Ce fichier de données est extrait de l'étude National Household Survey of Drug Abuse de 1998 et constitue un échantillon de probabilité des ménages américains. (*http://dx.doi.org/10.3886/ICPSR02934*) Ainsi, la première étape dans l'analyse de ce fichier doit consister à pondérer les données pour refléter les tendances de population.
- **stocks.sav** Ce fichier de données hypothétiques contient le cours et le volume des actions pour un an.
- **stroke_clean.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois celle-ci purgée via des procédures de l'option Validation de données.
- **stroke_invalid.sav.** Ce fichier de données d'hypothèse concerne l'état initial d'une base de données médicales et comporte plusieurs erreurs de saisie de données.
- **stroke_survival.** Ce fichier de données d'hypothèse concerne les temps de survie de patients qui quittent un programme de rééducation à la suite d'un accident ischémique et rencontrent un certain nombre de problèmes. Après l'attaque, l'occurrence d'infarctus du myocarde, d'accidents ischémiques ou hémorragiques est signalée, et le moment de l'événement enregistré. L'échantillon est tronqué à gauche car il n'inclut que les patients ayant survécu durant le programme de rééducation mis en place suite à une attaque.
- **stroke_valid.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois les valeurs vérifiées via la procédure Validation de données. Elle contient encore des observations anormales potentielles.
- **survey_sample.sav.** Ce fichier de données concerne des informations d'enquête dont des données démographiques et des mesures comportementales. Il est basé sur un sous-ensemble de variables de la 1998 NORC General Social Survey, bien que certaines valeurs de données aient été modifiées et que des variables supplémentaires fictives aient été ajoutées à titre de démonstration.
- **telco.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société de télécommunications pour réduire les taux de désabonnement de sa base de clients. Chaque observation correspond à un client distinct et enregistre diverses informations démographiques et d'utilisation de service.
- **telco_extra.sav.** Ce fichier de données est semblable au fichier de données *telco.sav* mais les variables de permanence et de dépenses des consommateurs transformées log ont été supprimées et remplacées par des variables de dépenses des consommateurs transformées log standardisées.
- **telco_missing.sav.** Ce fichier de données est un sous-ensemble du fichier de données *telco.sav* mais certaines des valeurs de données démographiques ont été remplacées par des valeurs manquantes.
- **testmarket.sav.** Ce fichier de données d'hypothèse concerne une chaîne de fast foods et ses plans marketing visant à ajouter un nouveau plat à son menu. Trois campagnes étant possibles pour promouvoir le nouveau produit, le nouveau plat est introduit sur des sites sur plusieurs marchés sélectionnés au hasard. Une promotion différente est effectuée sur chaque site et les ventes hebdomadaires du nouveau plat sont enregistrées pour les quatre premières semaines. Chaque observation correspond à un site-semaine distinct.
- **testmarket_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *testmarket.sav* dans lequel les ventes hebdomadaires sont organisées par site distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, les ventes

enregistrées sont à présent la somme des ventes réalisées au cours des quatre semaines de l'enquête.

- **tree_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_credit.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire.
- **tree_missing_data.sav** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire avec un grand nombre de valeurs manquantes.
- **tree_score_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_textdata.sav.** Ce fichier de données simples ne comporte que deux variables et vise essentiellement à indiquer l'état par défaut des variables avant affectation du niveau de mesure et des étiquettes de valeurs.
- **tv-survey.sav.** Ce fichier de données d'hypothèse concerne une enquête menée par un studio de télévision qui envisage de prolonger la diffusion d'un programme ou de l'arrêter. On a demandé à 906 personnes si elles regarderaient le programme dans diverses situations. Chaque ligne représente un répondant distinct et chaque colonne une situation distincte.
- **ulcer_recurrence.sav.** Ce fichier contient des informations partielles d'une enquête visant à comparer l'efficacité de deux thérapies de prévention de la récurrence des ulcères. Il fournit un bon exemple de données censurées par intervalle et a été présenté et analysé ailleurs (Collett, 2003).
- **ulcer_recurrence_recoded.sav.** Ce fichier réorganise les informations figurant dans le fichier *ulcer_recurrence.sav* pour que vous puissiez modéliser la probabilité d'événement pour chaque intervalle de l'enquête plutôt que la probabilité d'événement de fin d'enquête. Il a été présenté et analysé ailleurs (Collett et al., 2003).
- **verd1985.sav.** Ce fichier de données concerne une enquête (Verdegaal, 1985). Les réponses de 15 sujets à 8 variables ont été enregistrées. Les variables présentant un intérêt sont divisées en trois ensembles. Le groupe 1 comprend l'*âge* et la *situation familiale*, le groupe 2 les *animaux domestiques* et la *presse*, et le groupe 3 la *musique* et l'*habitat*. A la variable *animal domestique* est appliqué un codage nominal multiple et à *âge*, un codage ordinal ; toutes les autres variables ont un codage nominal simple.
- **virus.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fournisseur de services Internet pour déterminer les effets d'un virus sur ses réseaux. Il a suivi le pourcentage (approximatif) de trafic de messages électroniques infectés par un virus sur ses réseaux sur la durée, de la découverte à la circonscription de la menace.
- **wheeze_steubenville.sav.** Il s'agit d'un sous-ensemble d'une enquête longitudinale des effets de la pollution de l'air sur la santé des enfants (Ware, Dockery, Spiro III, Speizer, et Ferris Jr., 1984). Les données contiennent des mesures binaires répétées de l'état asthmatique d'enfants de la ville de Steubenville (Ohio), âgés de 7, 8, 9 et 10 ans, et indiquent si la mère fumait au cours de la première année de l'enquête.
- **workprog.sav.** Ce fichier de données d'hypothèse concerne un programme de l'administration visant à proposer de meilleurs postes aux personnes défavorisées. Un échantillon de participants potentiels au programme a ensuite été prélevé. Certains de ces participants ont

été sélectionnés au hasard pour participer au programme. Chaque observation représente un participant au programme distinct.

- **worldsales.sav** Ce fichier de données hypothétiques contient les revenus des ventes par continent et par produit.

Annexe

B

Remarques

Ces informations ont été développées pour les produits et services offerts dans le monde.

Il est possible qu'IBM n'offre pas dans les autres pays les produits, services et fonctionnalités décrits dans ce document. Contactez votre représentant local IBM pour obtenir des informations sur les produits et services actuellement disponibles dans votre région. Toute référence à un produit, programme ou service IBM n'implique pas que les seuls les produits, programmes ou services IBM peuvent être utilisés. Tout produit, programme ou service de fonctionnalité équivalente qui ne viole pas la propriété intellectuelle IBM peut être utilisé à la place. Cependant l'utilisateur doit évaluer et vérifier l'utilisation d'un produit, programme ou service non IBM.

IBM peut posséder des brevets ou des applications de brevet en attente qui couvrent les sujets décrits dans ce 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, Armonk, NY 10504-1785, États-Unis

Pour obtenir des informations de licence concernant la configuration de caractères codés sur deux octets (DBCS), veuillez contacter dans votre pays le département chargé de la propriété intellectuelle chez IBM ou envoyez vos commentaires par écrit à :

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japon.

Le paragraphe suivant ne s'applique pas au Royaume-Uni ni à aucun pays dans lequel ces dispositions sont contraires au droit local : INTERNATIONAL BUSINESS MACHINES FOURNIT CETTE PUBLICATION « EN L'ETAT » SANS GARANTIE D'AUCUNE SORTE, IMPLICITE OU EXPLICITE, Y COMPRIS, MAIS SANS ETRE LIMITE AUX GARANTIES IMPLICITES DE NON VIOLATION, DE QUALITE MARCHANDE OU D'ADAPTATION POUR UN USAGE PARTICULIER. Certains états n'autorisent pas l'exclusion de garanties explicites ou implicites lors de certaines transactions, par conséquent, il est possible que cet énoncé ne vous concerne pas.

Ces informations peuvent contenir des erreurs techniques ou des erreurs typographiques. Ces informations sont modifiées de temps en temps ; ces modifications seront intégrées aux nouvelles versions de la publication. IBM peut apporter des améliorations et/ou modifications des produits et/ou des programmes décrits dans cette publications à tout moment sans avertissement préalable.

Toute référence dans ces informations à des sites Web autres qu'IBM est fournie dans un but pratique uniquement et ne sert en aucun cas de recommandation pour ces sites Web. Le matériel contenu sur ces sites Web ne fait pas partie du matériel de ce produit IBM et l'utilisation de ces sites Web se fait à vos propres risques.

IBM peut utiliser ou distribuer les informations que vous lui fournissez, de la façon dont il le souhaite, sans encourir aucune obligation envers vous.

Les personnes disposant d'une licence pour ce programme et qui souhaitent obtenir des informations sur celui-ci pour activer : (i) l'échange d'informations entre des programmes créés de manière indépendante et d'autres programmes (notamment celui-ci) et (ii) l'utilisation mutuelle des informations qui ont été échangées, doivent contacter :

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, États-Unis.

Ces informations peuvent être disponibles, soumises à des conditions générales, et dans certains cas payantes.

Le programme sous licence décrit dans ce document et toute la documentation sous licence disponible pour ce programme sont fournis par IBM en conformité avec les conditions de l'accord du client IBM, avec l'accord de licence du programme international IBM et avec tout accord équivalent entre nous.

les informations concernant les produits autres qu'IBM ont été obtenues auprès des fabricants de ces produits, leurs annonces publiques ou d'autres sources publiques disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances, leur compatibilité ou toute autre fonctionnalité associée à des produits autres qu'IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits.

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. Tous ces noms sont fictifs et toute ressemblance avec des noms et des adresses utilisés par une entreprise réelle ne serait que pure coïncidence.

Si vous consultez la version papier de ces informations, il est possible que certaines photographies et illustrations en couleurs n'apparaissent pas.

Marques commerciales

IBM, le logo IBM, ibm.com et SPSS sont des marques commerciales d'IBM Corporation, déposées dans de nombreuses juridictions du monde entier. Une liste à jour des marques IBM est disponible sur Internet à l'adresse *http://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.

Java et toutes les marques et logos Java sont des marques commerciales de Sun Microsystems, Inc. aux Etats-Unis et/ou 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.

Ce produit utilise WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com/*.

Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés.

Les captures d'écran des produits Adobe sont reproduites avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran des produits Microsoft sont reproduites avec l'autorisation de Microsoft Corporation.

Bibliographie

Barlow, R. E., D. J. Bartholomew, D. J. Bremner, et H. D. Brunk. 1972. *Statistical inference under order restrictions*. New York: John Wiley and Sons.

Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

Benzécri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. Dans : *Methodologies of Pattern Recognition,* S. Watanabe, éd. New York: Academic Press.

Benzécri, J. P. 1992. *Correspondence analysis handbook*. New York: Marcel Dekker.

Bishop, Y. M., S. E. Feinberg, et P. W. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. Cambridge, Massachusetts: MIT Press.

Blake, C. L., et C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., et J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, .

Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics*, 18, .

Busing, F. M. T. A., P. J. F. Groenen, et W. J. Heiser. 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, 70, .

Carroll, J. D. 1968. Generalization of canonical correlation analysis to three or more sets of variables. Dans : *Proceedings of the 76th Annual Convention of the American Psychological Association, 3,* Washington, D.C.: American Psychological Association.

Collett, D. 2003. *Modelling survival data in medical research*, 2 éd. Boca Raton: Chapman & Hall/CRC.

Commandeur, J. J. F., et W. J. Heiser. 1993. *Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices*. Leiden: Department of Data Theory, University of Leiden.

De Haas, M., J. A. Algera, H. F. J. M. Van Tuijl, et J. J. Meulman. 2000. Macro and micro goal setting: In search of coherence. *Applied Psychology*, 49, .

De Leeuw, J. 1982. Nonlinear principal components analysis. Dans : *COMPSTAT Proceedings in Computational Statistics,* Vienne: Physica Verlag.

De Leeuw, J. 1984. *Canonical analysis of categorical data*, 2nd éd. Leiden: DSWO Press.

De Leeuw, J. 1984. The Gifi system of nonlinear multivariate analysis. Dans : *Data Analysis and Informatics III,* E. Diday, et al., éd..

De Leeuw, J., et W. J. Heiser. 1980. Multidimensional scaling with restrictions on the configuration. Dans : *Multivariate Analysis, Vol. V,* P. R. Krishnaiah, éd. Amsterdam: North-Holland.

De Leeuw, J., et J. Van Rijckevorsel. 1980. HOMALS and PRINCALS—Some generalizations of principal components analysis. Dans : *Data Analysis and Informatics,* E. Diday, et al., éd. Amsterdam: North-Holland.

De Leeuw, J., F. W. Young, et Y. Takane. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .

De Leeuw, J. 1990. Multivariate analysis with optimal scaling. Dans : *Progress in Multivariate Analysis,* S. Das Gupta, et J. Sethuraman, éds. Calcutta: Indian Statistical Institute.

Eckart, C., et G. Young. 1936. The approximation of one matrix by another one of lower rank. *Psychometrika*, 1, .

Fisher, R. A. 1938. *Statistical methods for research workers*. Edimbourg: Oliver and Boyd.

Fisher, R. A. 1940. The precision of discriminant functions. *Annals of Eugenics*, 10, .

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, 58, .

Gifi, A. 1985. *PRINCALS. Research Report UG-85-02*. Leiden: Department of Data Theory, University of Leiden.

Gifi, A. 1990. *Nonlinear multivariate analysis*. Chichester: John Wiley and Sons.

Gilula, Z., et S. J. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association*, 83, .

Gower, J. C., et J. J. Meulman. 1993. The treatment of categorical information in physical anthropology. *International Journal of Anthropology*, 8, .

Green, P. E., et V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., et Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. Dans : *The Prediction of Personal Adjustment,* P. Horst, éd. New York: Social Science Research Council.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, .

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

Hastie, T., et R. Tibshirani. 1990. *Generalized additive models*. Londres: Chapman and Hall.

Hastie, T., R. Tibshirani, et A. Buja. 1994. Flexible discriminant analysis. *Journal of the American Statistical Association*, 89, .

Hayashi, C. 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statitical Mathematics*, 2, .

Heiser, W. J. 1981. *Unfolding analysis of proximity data*. Leiden: Department of Data Theory, University of Leiden.

Heiser, W. J., et F. M. T. A. Busing. 2004. Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations. Dans : *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, éd. Thousand Oaks, Californie: Sage Publications, Inc..

Heiser, W. J., et J. J. Meulman. 1994. Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. Dans : *Correspondence Analysis in the Social Sciences: Recent Developments and Applications,* M. Greenacre, et J. Blasius, éds. New York: Academic Press.

Heiser, W. J., et J. J. Meulman. 1995. Nonlinear methods for the analysis of homogeneity and heterogeneity. Dans : *Recent Advances in Descriptive Multivariate Analysis,* W. J. Krzanowski, éd. Oxford: Oxford University Press.

Horst, P. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17, .

Horst, P. 1961. Relations among m sets of measures. *Psychometrika*, 26, .

Israëls, A. 1987. *Eigenvalue techniques for qualitative data*. Leiden: DSWO Press.

Kennedy, R., C. Riquier, et B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, .

Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika*, 58, .

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, .

Kruskal, J. B. 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, .

Kruskal, J. B. 1965. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society Series B*, 27, .

Kruskal, J. B. 1978. Factor analysis and principal components analysis: Bilinear methods. Dans : *International Encyclopedia of Statistics,* W. H. Kruskal, et J. M. Tanur, éds. New York: The Free Press.

Kruskal, J. B., et R. N. Shepard. 1974. A nonmetric variety of linear factor analysis. *Psychometrika*, 39, .

Krzanowski, W. J., et F. H. C. Marriott. 1994. *Multivariate analysis: Part I, distributions, ordination and inference*. Londres: Edward Arnold.

Lebart, L., A. Morineau, et K. M. Warwick. 1984. *Multivariate descriptive statistical analysis*. New York: John Wiley and Sons.

Lingoes, J. C. 1968. The multivariate analysis of qualitative data. *Multivariate Behavioral Research*, 3, .

Max, J. 1960. Quantizing for minimum distortion. *Proceedings IEEE (Information Theory)*, 6, .

McCullagh, P., et J. A. Nelder. 1989. *Generalized Linear Models*, 2nd éd. Londres: Chapman & Hall.

Meulman, J. J. 1982. *Homogeneity analysis of incomplete data*. Leiden: DSWO Press.

Meulman, J. J. 1986. *A distance approach to nonlinear multivariate analysis*. Leiden: DSWO Press.

Meulman, J. J. 1992. The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika*, 57, .

Meulman, J. J. 1993. Principal coordinates analysis with optimal transformations of the variables: Minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology*, 46, .

Meulman, J. J. 1996. Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification*, 13, .

Meulman, J. J. 2003. Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, 4, .

Meulman, J. J., et W. J. Heiser. 1997. Graphical display of interaction in multiway contingency tables by use of homogeneity analysis. Dans : *Visual Display of Categorical Data,* M. Greenacre, et J. Blasius, éds. New York: Academic Press.

Meulman, J. J., et P. Verboon. 1993. Points of view analysis revisited: Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika*, 58, .

Meulman, J. J., A. J. Van der Kooij, et A. Babinec. 2000. New features of categorical principal components analysis for complicated data sets, including data mining. Dans : *Classification, Automation and New Media,* W. Gaul, et G. Ritter, éds. Berlin: Springer-Verlag.

Meulman, J. J., A. J. Van der Kooij, et W. J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. Dans : *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, éd. Thousand Oaks, Californie: Sage Publications, Inc..

Nishisato, S. 1980. *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S. 1984. Forced classification: A simple application of a quantification method. *Psychometrika*, 49, .

Nishisato, S. 1994. *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Pratt, J. W. 1987. Dividing the indivisible: Using simple symmetry to partition variance explained. Dans : *Proceedings of the Second International Conference in Statistics,* T. Pukkila, et S. Puntanen, éds. Tampere, Finlande: Université de Tampere.

Price, R. H., et D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Ramsay, J. O. 1989. Monotone regression splines in action. *Statistical Science*, 4, .

Rao, C. R. 1973. *Linear statistical inference and its applications*, 2nd éd. New York: John Wiley and Sons.

Rao, C. R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. Dans : *Multivariate Analysis, Vol. 5,* P. R. Krishnaiah, éd. Amsterdam: North-Holland.

Rickman, R., N. Mitchell, J. Dingman, et J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .

Rosenberg, S., et M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, .

Roskam, E. E. 1968. *Metric analysis of ordinal data in psychology*. Voorschoten: VAM.

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrika*, 27, .

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function II. *Psychometrika*, 27, .

Shepard, R. N. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, .

Tenenhaus, M., et F. W. Young. 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, .

Theunissen, N. C. M., J. J. Meulman, A. L. Den Ouden, H. M. Koopman, G. H. Verrips, S. P. Verloove-Vanhorick, et J. M. Wit. 2003. Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology*, 4, .

Tucker, L. R. 1960. Intra-individual and inter-individual multidimensionality. Dans : *Psychological Scaling: Theory & Applications,* H. Gulliksen, et S. Messick, éds. New York: John Wiley and Sons.

Van der Burg, E. 1988. *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO Press.

Van der Burg, E., et J. De Leeuw. 1983. Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, .

Van der Burg, E., J. De Leeuw, et R. Verdegaal. 1988. Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 53, .

Van der Ham, T., J. J. Meulman, D. C. Van Strien, et H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, .

Van der Kooij, A. J., et J. J. Meulman. 1997. MURALS: Multiple regression and optimal scaling using alternating least squares. Dans : *Softstat '97,* F. Faulbaum, et W. Bandilla, éds. Stuttgart: Gustav Fisher.

Van Rijckevorsel, J. 1987. *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. Leiden: DSWO Press.

Verboon, P., et R. A. Van der Lans. 1994. Robust canonical discriminant analysis. *Psychometrika*, 59, .

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (en néerlandais)*. Leiden: Department of Data Theory, University of Leiden.

Vlek, C., et P. J. Stallen. 1981. Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance*, 28, .

Wagenaar, W. A. 1988. *Paradoxes of gambling behaviour*. Londres: Lawrence Erlbaum Associates, Inc.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, et B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, .

Winsberg, S., et J. O. Ramsay. 1980. Monotonic transformations to additivity using splines. *Biometrika*, 67, .

Winsberg, S., et J. O. Ramsay. 1983. Monotone spline transformations for dimension reduction. *Psychometrika*, 48, .

Wolter, K. M. 1985. *Introduction to variance estimation*. Berlin: Springer-Verlag.

Young, F. W. 1981. Quantitative analysis of qualitative data. *Psychometrika*, 46, .

Young, F. W., J. De Leeuw, et Y. Takane. 1976. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .

Young, F. W., Y. Takane, et J. De Leeuw. 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, .

Zeijl, E., Y. te Poel, M. du Bois-Reymond, J. Ravesloot, et J. J. Meulman. 2000. The role of parents and peers in the leisure activities of young adolescents. *Journal of Leisure Research*, 32, .

Index