IBM SPSS Modeler 18.1.1 入力ノード、プロセス・ノード、出力ノード

注記

本書および本書で紹介する製品をご使用になる前に、 423 ページの『特記事項』に記載されている情報をお読みください。

本書は、IBM SPSS Modeler バージョン 18 リリース 1 モディフィケーション 1 および新しい版で明記されない限り、以降のすべてのリリースおよびモディフィケーションに適用されます。

お客様の環境によっては、資料中の円記号がバックスラッシュと表示されたり、バックスラッシュが円記号と表示されたりする場合があります。

原典： IBM SPSS Modeler 18.1.1 Source, Process, and Output Nodes

発行： 日本アイ・ビー・エム株式会社

担当： トランスレーション・サービス・センター

目次

第 4 章 フィールド設定ノード 127

前書き

IBM® SPSS® Modeler は、IBM Corp. が開発した企業強化用のデータ・マイニング・ワークベンチです。SPSS Modeler を使用すると、企業はデータを詳しく調べることにより顧客および一般市民とのリレーションシップを強化することができます。企業は、SPSS Modeler を使用して得られた情報に基づいて利益をもたらす顧客を獲得し、抱き合わせ販売の機会を見つけ、新規顧客を引き付け、不正を発見し、リスクを減少させ、政府機関へのサービスの提供を改善することができます。

SPSS Modeler の視覚的インターフェースを使用すると、特定ビジネスの専門知識を適用し、より強力な予測モデルを実現し、解決までの時間を短縮します。SPSS Modeler では、予測、分類、セグメント化、および関連性検出アルゴリズムなど、さまざまなモデル作成手法を提供しています。モデルを作成した後は、IBM SPSS Modeler Solution Publisher により、企業全体の意思決定者やデータベースにモデルを配布することが可能になります。

IBM Business Analytics について

IBM Business Analytics ソフトウェアは、意思決定者がビジネス・パフォーマンスの改善のために信頼できる完全で整合性があり、正確な情報を提供します。ビジネス・インテリジェンス、予測分析、財務実績および戦略管理、分析アプリケーション の包括的なポートフォリオを利用することによって、現在の実績を明確、迅速に理解し、将来の結果を予測することができます。豊富な産業用ソリューション、証明された実践法、それに専門家によるサービスを組み合せることにより、あらゆる規模の会社組織が、最高の生産性を推進し、信頼できる意志決定を自動化し、そして、よりよい結果を実現させることができます。

このポートフォリオの一部として、IBM SPSS Predictive Analytics ソフトウェアを使用する組織は、将来のイベントを予測し、その洞察に基づいて積極的に行動し、より優れた業績を実現することができます。全世界の企業、政府、学術分野のお客様が IBM SPSS の技術を活用し、不正行為を減少させ、リスクを軽減させながら、顧客の獲得、保持、成長において、競争優位を高めることができます。IBM SPSS ソフトウェアを日常業務に組み込めば、組織は予測分析を活用する企業となり、ビジネス目標に対応するよう意思決定の方向付けと自動化を行い、高い競争力を獲得できるようになります。詳細な情報、または営業担当者へのお問い合わせ方法については、http://www.ibm.com/spss を参照してください。

技術サポート

お客様はテクニカル・サポートをご利用いただけます。IBM Corp. 製品の使用方法、または対応するハードウェア環境へのインストールについてサポートが必要な場合は、テクニカル・サポートにご連絡ください。テクニカル・サポートのご利用には、http://www.ibm.com/supportのIBM Corp. Web サイトをご覧ください。支援を要請される場合は、事前にユーザー、会社組織、そして、サポート契約を明確にしておいていただくよう、お願いします。

第 1 章 IBM SPSS Modeler について

IBM SPSS Modeler は、ビジネスの専門知識を活用して予測モデルを迅速に作成したり、また作成したモデルをビジネス・オペレーションに展開して意志決定を改善できるようにする、一連のデータ・マイニング・ツールです。IBM SPSS Modeler は業界標準の CRISP-DM モデルをベースに設計されたものであり、データ・マイニング・プロセス全体をサポートして、データに基づいてより良いビジネスの成果を達成できるようにします。

IBM SPSS Modeler には、マシン学習、人工知能、および統計に基づいたさまざまなモデル作成方法が用意されています。「モデル作成」パレットを利用して、データから新しい情報を引き出したり、予測モデルを作成することができます。各手法によって、利点や適した問題の種類が異なります。

SPSS Modeler は、 スタンドアロン製品として購入または SPSS Modeler Server と組み合わせてクライアントとして使用することができます。後のセクションで説明されているとおり、多くの追加オプションも使用することができます。 詳しくは、https://www.ibm.com/analytics/us/en/technology/spss/を参照してください。

IBM SPSS Modeler 製品

製品と関連するソフトウェアの IBM SPSS Modeler ファミリーの構成は次のとおりです。

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console (IBM SPSS Deployment Manager に付属)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server Adapters for IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler はこの製品のすべての機能を搭載したバージョンであり、ユーザーのパーソナル・コンピューターにインストールし、そのコンピューターで実行します。スタンドアロン製品としてローカル・モードで SPSS Modeler を実行するか、大規模なデータ・セットを使用する場合にパフォーマンスを向上させるために IBM SPSS Modeler Server と組み合わせて実行することができます。

SPSS Modeler を使用して、プログラミングの必要なく、正確な予測モデルを迅速かつ直感的に構築することができます。独自のビジュアル・インターフェースを使用すると、データ・マイニング・プロセスを簡単に視覚化することができます。製品に組み込まれている高度な分析の支援を受けて、データ内に隠れたパターンやトレンドを発見することができます。結果をモデル化し、結果に影響を与える要因を理解することにより、ビジネスチャンスを生かしてリスクを軽減できるようになります。

SPSS Modeler は SPSS Modeler Professional および SPSS Modeler Premium の 2 つのエディションで使用できます。 詳しくは、トピック 3 ページの『IBM SPSS Modeler のエディション』を参照してください。

IBM SPSS Modeler Server

SPSS Modeler は、クライアント/サーバー アーキテクチャーを使用して、リソース集中型の操作が必要な要求を、強力なサーバー ソフトウェアへ分散することにより、大規模なデータ セットに対するパフォーマンスを高速化します。

SPSS Modeler Server は、1 つまたは複数の IBM SPSS Modeler のインストールと組み合わせてサーバー・ホストで分散分析モードで継続的に実行する、別途ライセンスが必要な製品です。サーバーへの分散により、SPSS Modeler Server は、クライアント コンピューターにデータをダウンロードせずにサーバー上でメモリーを集中的に使用する操作を実行できるため、大きなデータ セットで優れたパフォーマンスを発揮します。さらに、IBM SPSS Modeler Server は、SQL の最適化とデータベース内のモデリング機能のサポートにより、パフォーマンスをさらに高め、より高度な自動化を実現しています。

IBM SPSS Modeler Administration Console

Modeler Administration Console は、SPSS Modeler Server 構成オプションの多くを管理するグラフィカル・ユーザー・インターフェースです。それらの構成オプションは、オプション・ファイルで設定することも可能です。コンソールは、IBM SPSS Deployment Manager に含まれています。コンソールを使用すると、SPSS Modeler Server インストール済み環境をモニターしたり、構成したりできます。SPSS Modeler Server の現在の顧客は、コンソールを無料で利用できます。アプリケーションは Windows コンピューターにのみインストールできますが、サポートされる任意のプラットフォームにインストールされたサーバーを管理できます。

IBM SPSS Modeler Batch

データマイニングは、通常、対話型のプロセスですが、グラフィカル・ユーザー・インターフェースを必要とせずに、コマンドラインから SPSS Modeler を実行することも可能です。例えば、ユーザーの介入なしで実行する長期実行または反復的なタスクがあります。SPSS Modeler Batch は、通常のユーザー・インターフェースにアクセスせずに SPSS Modeler の完全な分析機能のサポートを提供する製品の特別バージョンです。SPSS Modeler Batch を使用するには、SPSS Modeler Server が必要です。

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher は、SPSS Modeler ストリームのパッケージ版を作成することができるツールで、このストリームは外部ランタイム・エンジンによって実行することも、外部アプリケーションに埋め込むこともできます。このように、SPSS Modeler がインストールされていない環境で使用するための完全な SPSS Modeler ストリームを公開して展開することができます。SPSS Modeler Solution Publisher は、個別のライセンスが必要とされている IBM SPSS Collaboration and Deployment Services - Scoring サービスの一部として配布されています。このライセンスにより、SPSS Modeler Solution Publisher Runtime を利用できるようになり、公開されたストリームを実行することができます。

SPSS Modeler Solution Publisher について詳しくは、IBM SPSS Collaboration and Deployment Services の資料を参照してください。IBM SPSS Collaboration and Deployment Services Knowledge Center に『IBM SPSS Modeler Solution Publisher』と『IBM SPSS Analytics Toolkit』というセクションがあります。

IBM SPSS Collaboration and Deployment Services 用の IBM SPSS Modeler Server アダプター

さまざまな IBM SPSS Collaboration and Deployment Services 用のアダプターを使用すると、SPSS Modeler および SPSS Modeler Server を IBM SPSS Collaboration and Deployment Services リポジト

リーとインタラクティブに機能させることができます。このように、リポジトリーに展開された SPSS Modeler ストリームは、複数のユーザーで共有したり、シンクライアント アプリケーションである IBM SPSS Modeler Advantage からアクセスしたりできます。リポジトリーをホストするシステムに、アダプターをインストールします。

IBM SPSS Modeler のエディション

SPSS Modeler は次のエディションで使用できます。

SPSS Modeler Professional

SPSS Modeler Professional は、CRM システムで追跡する行動や対話、人口統計データ、購入行動や販売データなど、多くの構造化データを処理するために必要なすべてのツールを提供しています。

SPSS Modeler Premium

SPSS Modeler Premium は、特化したデータ、または構造化されていないテキスト・データを処理するために SPSS Modeler Professional を拡張する、別途ライセンスが必要な製品です。SPSS Modeler Premium には、以下の IBM SPSS Modeler Text Analytics が含まれます。

IBM SPSS Modeler Text Analytics は、高度な言語技術と Natural Language Processing (NLP) を使用して、構造化されていない多様なテキスト・データをすばやく処理し、重要なコンセプトを抽出および組織化し、そしてそのコンセプトをカテゴリー別に分類します。抽出されたコンセプトとカテゴリーを、人口統計のような既存の構造化データと組み合わせ、IBM SPSS Modeler の豊富なデータ・マイニング・ツールを適用する方法で、焦点を絞ったより良い決定を下すことができます。

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription は、従来の IBM SPSS Modeler クライアントとすべて同じ予測分析機能を提供します。Subscription エディションの場合、定期的に製品アップデートをダウンロードできます。

資料

資料は、SPSS Modeler の「ヘルプ」メニューから参照できます。この「ヘルプ」メニューから Knowledge Center を開きます。Knowledge Center は、製品の外部で公に利用できます。

各製品の完全な資料 (インストール手順を含む) は、PDF 形式でも提供されており、製品ダウンロードの一部として、個別の圧縮フォルダーに格納されています。PDF 文書は、Web (http://www.ibm.com/support/docview.wss?uid=swg27046871) からもダウンロードできます。

SPSS Modeler Professional ドキュメント

SPSS Modeler Professional のドキュメント スイート (インストール手順を除く) は次のとおりです。

- **IBM SPSS Modeler** ユーザーズ・ガイド**:** SPSS Modeler の使用への全体的な入門で、データ ストリームの作成方法、欠損値の処理方法、CLEM 式の作成方法、プロジェクトおよびレポートの処理方法、および IBM SPSS Collaboration and Deployment Services または IBM SPSS Modeler Advantage に展開するためのストリームのパッケージ方法が含まれています。
- 「**IBM SPSS Modeler** 入力ノード、プロセス・ノード、出力ノード」。各種形式のデータの読み取り、処理、および出力に使用するすべてのノードの説明です。これは、モデル作成ノード以外のすべてのノードについての説明です。

- 「**IBM SPSS Modeler** モデル作成ノード」。データ・マイニング・モデルの作成に使用するすべてのノードについての説明です。IBM SPSS Modeler には、マシン学習、人工知能、および統計に基づいたさまざまなモデル作成方法が用意されています。
- **IBM SPSS Modeler** アプリケーション・ガイド。このガイドの例では、特定のモデル作成手法および技法について、簡単に対象を絞って紹介します。本ガイドのオンライン バージョンは、「ヘルプ」メニューからも利用できます。詳しくは、トピック『アプリケーションの例』を参照してください。
- 「**IBM SPSS Modeler Python** スクリプトとオートメーション」。Python スクリプトによるシステムの自動化に関する情報です。ノードおよびストリームの操作に使用できるプロパティーを含めて説明します。
- **IBM SPSS Modeler** 展開ガイド**:** IBM SPSS Deployment Manager のもとで処理されるジョブ内のステップとして IBM SPSS Modeler のストリームを実行することに関する情報。
- **IBM SPSS Modeler CLEF** 開発者ガイド**:** CLEF では、IBM SPSS Modeler のノードとしてデータ処理ルーチンやモデル作成アルゴリズムなどのサード・パーティー製のプログラムを統合できます。
- 「**IBM SPSS Modeler** データベース内 マイニング・ガイド」。サード・パーティー製アルゴリズムを使用してご使用のデータベースの能力を利用してパフォーマンスを向上させ、分析機能の範囲を拡張する方法に関する情報を示します。
- **IBM SPSS Modeler Server** 管理およびパフォーマンス・ガイド**:** IBM SPSS Modeler Server の構成方法と管理方法に関する情報。
- 「**IBM SPSS Deployment Manager** ユーザー・ガイド」。IBM SPSS Modeler Server の監視や構成を行うための Deployment Manager アプリケーションに組み込まれている管理コンソール・ユーザー・インターフェースの使用法に関する情報。
- 「**IBM SPSS Modeler CRISP-DM** ガイド」。SPSS Modeler でのデータ・マイニングに対する CRISP-DM 方法の使用に関するステップバイステップのガイドです。
- **IBM SPSS Modeler Batch** ユーザーズ・ガイド**:** IBM SPSS Modeler をバッチ・モードで使用するための完全ガイドで、バッチ・モードでの実行およびコマンド・ライン引数の詳細について説明します。このガイドは、PDF 形式のみです。

SPSS Modeler Premium ドキュメント

SPSS Modeler Premium のドキュメント スイート (インストール手順を除く) は次のとおりです。

- 「**SPSS Modeler Text Analytics** ユーザーズ・ガイド」。SPSS Modeler でテキスト分析を使用する場合の情報。テキスト・マイニング・ノード、インタラクティブ・ワークベンチ、テンプレートなどについて説明します。

アプリケーションの例

SPSS Modeler のデータ・マイニング・ツールは、多様なビジネスおよび組織の問題解決を支援しますが、アプリケーションの例では、特定のモデル作成手法および技術に関する簡単で、目的に沿った説明を行います。ここで使用されるデータセットは、データ・マイニング作業者によって管理される巨大なデータ・ストアよりも非常に小さいですが、関係するコンセプトや方法は実際のアプリケーションの規模に応じて拡張できます。

例にアクセスするには、SPSS Modeler の「ヘルプ」メニューで「アプリケーションの例」をクリックします。

データ・ファイルとサンプル・ストリームは、製品のインストール・ディレクトリーの Demos フォルダーにインストールされています。詳しくは、 5 ページの『Demos フォルダー』を参照してください。

データベース・モデル作成の例 : 例は、『*IBM SPSS Modeler* データベース内マイニング・ガイド』を参照してください。

スクリプトの例 : 例は、『*IBM SPSS Modeler* スクリプトとオートメーション ガイド』を参照してください。

Demos フォルダー

アプリケーションの例で使用されるデータ・ファイルとサンプル・ストリームは、製品のインストール ディレクトリの Demos フォルダー (例: C:¥Program Files¥IBM¥SPSS¥Modeler¥<version>¥Demos) にインストールされています。このフォルダーには、Windowsの「スタート」メニューの IBM SPSS Modeler プログラム・グループから、または「ファイル」 > 「ストリームを開く」ダイアログ・ボックスの最近使ったディレクトリーのリストで「Demos」をクリックしてアクセスすることもできます。

ライセンスの追跡

SPSS Modeler を使用すると、ライセンスの使用状況が一定の間隔で追跡され、ログに記録されます。ログに記録されるライセンスメトリックは *AUTHORIZED_USER* と *CONCURRENT_USER* であり、ログに記録されるメトリックのタイプは、SPSS Modeler に使用するライセンスのタイプによって決まります。

作成されたログファイルは IBM License Metric Tool によって処理可能であり、そのファイルからライセンス使用状況レポートを生成できます。

ライセンスログファイルは、SPSS Modeler クライアントログファイルが記録されるディレクトリと同じディレクトリに作成されます (デフォルトでは %ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log)。

第 2 章 ソース・ノード

概要

ソース・ノードを使用すると、フラット・ファイル、IBM SPSS Statistics (.sav)、SAS、Microsoft Excel、および ODBC 準拠のリレーショナル・データベースなどのさまざまな形式で格納されたデータを、インポートできるようになります。ユーザー入力ノードを使用して、合成データを生成することもできます。

「入力」パレットには、次のノードがあります。

Analytic Server ソースにより、Hadoop 分散ファイル・システム (HDFS) でストリームを実行することができます。Analytic Server データ ソースの情報は、テキスト ファイルやデータベースなど、さまざまな場所から得られます。詳しくは、トピック 13 ページの『Analytic Server ソース・ノード』を参照してください。

データベース ノードは、Microsoft SQL Server、Db2、Oracle など ODBC (開放型データベース接続) を使用する他のさまざまなパッケージからデータをインポートするのに使用できます。詳しくは、トピック 19 ページの『データベース・ソース・ノード』を参照してください。

可変長ノードで、可変長フィールド・テキスト・ファイル、つまりフィールド数は一定でも各フィールド内の文字数が異なるレコードを含むファイルから、データを読み込みます。このノードは、固定長のヘッダー・テキストやある種の注釈があるファイルにも使用できます。詳しくは、トピック 28 ページの『可変長ファイル・ノード』を参照してください。

固定長ノードで、固定長フィールド・テキスト・ファイルからデータをインポートします。ここで、ファイルのフィールドは区切られていませんが、同じ位置から始まって長さは固定されています。コンピューター生成のデータや、旧来のシステムのデータなどは、しばしば固定長フィールド形式で保存されています。詳しくは、トピック 32 ページの『固定長ファイル・ノード』を参照してください。

Statistics ファイル ノードは、IBM SPSS Statistics で使用される *.sav* または *.zsav* ファイル形式のデータおよび IBM SPSS Modeler に保存されたキャッシュ ファイル (同じ形式を使用する) を読み込みます。

Data Collection ノードで、Data Collection Data Model に基づく市場調査ソフトウェアによって使用されるさまざまな形式の調査データをインポートします。このノードを使用するには、Data Collection Developer Library がインストールされている必要があります。詳しくは、トピック 35 ページの『Data Collectionノード』を参照してください。

IBM Cognos ソース・ノードは、Cognos Analytics データベースからデータをインポートします。

IBM Cognos TM1 入力ノードは、Cognos TM1 データベースからデータをインポートします。

SAS ファイル・ノードで、SAS データを IBM SPSS Modeler へインポートします。詳しくは、トピック 47 ページの『SAS ソース・ノード』を参照してください。

Excel ノードは、Microsoft Excel から .xlsx ファイル形式でデータをインポートします。ODBC データ・ソースは不要です。詳しくは、トピック 48 ページの『Excel ソース・ノード』を参照してください。

XML ソース・ノードを使用して、XML 形式のデータをストリームにインポートできます。ディレクトリーの 1 つのファイルまたはすべてのファイルをインポートできます。オプションで、XML 構造を読み込むスキーマ ファイルを指定できます。

ユーザー入力ノードを利用すれば、最初から、あるいは既存のデータを変更して、合成データを簡単に作成できます。これは、モデル作成用の検定データ・セットを作成する場合などに役立ちます。詳しくは、トピック 51 ページの『ユーザー入力ノード』を参照してください。

シミュレーション生成ノードにより、シミュレーション対象のデータを容易に生成することができます。このとき、ユーザー指定の統計分布を使用して最初から生成するか、既存の履歴データに対してシミュレーション適合ノードを実行して得られた分布を使用して自動的に生成することができます。これは、モデルの入力に不確定性がある状況で予測モデルの結果を評価するときに便利です。

データ・ビュー・ノードを使用すると、IBM SPSS Collaboration and Deployment Services 分析データ ビューで定義されたデータ・ソースにアクセスすることができます。分析データ ビューは、データにアクセスするための標準のインターフェースを定義し、複数の物理データ・ソースをそのインターフェースに関連付けます。詳しくは、トピック 70 ページの『データ・ビュー・ノード』を参照してください。

データ マイニング セッションにマップまたは地理空間データを導入するには、地理空間入力ノードを使用します。詳しくは、トピック 72 ページの『地理空間ソース・ノード』を参照してください。

ストリームを開始するには、ソース・ノードをストリーム領域に追加します。次に、配置したノードをダブルクリックして、ダイアログ・ボックスを表示します。このダイアログ・ボックス内のさまざまなタブで、データの読み込み、フィールドと値の表示ができ、フィルター、データ型、フィールドの役割、欠損値の検査などを含む多様なオプションを設定することができます。

フィールドのストレージと形式の設定

固定長、可変長、XML 入力およびユーザー入力の各入力ノードで「データ」タブにあるオプションを使用すると、データが IBM SPSS Modeler 内に読み込まれたり作成されたりする場合のフィールドのストレージ・タイプを指定できます。また、固定長、可変長およびユーザー入力ノードの場合、フィールド形式およびその他のメタデータも指定できます。

ほかのソースから読み込まれたデータの場合、ストレージは自動的に決定されますが、置換ノードまたはフィールド作成ノード内で `to_integer` のような変換関数を使用することで変更できます。

フィールド 現在のデータセットのフィールドを表示して選択するには、「フィールド」列を使用します。

上書き 「上書き」列のチェック ボックスを選択すると、「ストレージ」列と「入力形式」列のオプションが有効になります。

データ・ストレージ

ストレージは、フィールド中へのデータの格納方法を表しています。例えば、1 と 0 の値をとるフィールドは整数データを格納します。これはデータの使用法を記述する測定の尺度とは異なり、ストレージに影響を与えません。例えば、値 1 と 0 をとる整数フィールドの測定の尺度をフラグ型に設定することができます。通常は、1 = *True*、0 = *False* を示します。ストレージはソースで確定する必要がありますが、測定の尺度はストリームのどこでもデータ型ノードを使用して変更できます。詳しくは、トピック 145 ページの『尺度』を参照してください。

指定できるストレージ・タイプを次に示します。

- 文字列 非数値データ (別名、英数字データ) を含むフィールドに使用されます。文字列には、*fred*、*Class 2*、または *1234* など、任意の文字のシーケンスを含めることができます。注意を要するのは、文字列内の数字は計算には使えないことです。
- 整数 値が整数で示されるフィールドです。
- 実数 値は数字で示され、小数点を含むことがあります (整数に限定されません)。表示形式は「ストリーム プロパティー」ダイアログ ボックスで指定し、データ型ノード(「形式」タブ) の各フィールドでオーバーライドできます。
- 日付 年、月、日など、標準形式で指定された日付の値です (`2007-09-26` など)。「ストリーム・プロパティー」ダイアログ・ボックスで特定の形式を指定します。
- 時間 期間として測定される時間です。例えば、「ストリーム プロパティー」ダイアログ ボックスで指定した現在の時間形式に応じて、1 時間 26 分 38 秒続くサービスコールを「`01:26:38`」と表現することができます。

- タイムスタンプ 例えば 2007-09-26 09:04:00 のように、日付と時刻の両方の構成要素を含む値です。この場合も、「ストリーム・プロパティー」ダイアログボックスの現在の日付と時間の形式に従います。日付と時刻が別々の値ではなく 1 つの値として解釈されるようにするには、タイム・スタンプ値を二重引用符で囲む必要があります (この規則は、ユーザー入力ノードで値を入力する場合などに適用されます)。
- リスト SPSS Modeler バージョン 17 で、地理空間および集合の新しい尺度とともに導入されました。リストのストレージ フィールドには、単一のレコードに対する複数の値が入ります。その他すべてのストレージ タイプのリスト版が存在します。

表 1. リストのストレージ タイプを示すアイコン

アイコン	ストレージ・タイプ
[A]	文字列のリスト
[ ]	整数のリスト
[ ]	実数のリスト
[ ]	時間のリスト
[ ]	日付のリスト
[ ]	タイムスタンプのリスト
[[]]	0 よりも大きな深さを持つリスト

さらに、集合の尺度とともに使用する場合は、以下の尺度のリスト版もあります。

表 2. リストの尺度を示すアイコン

アイコン	尺度
[ ]	連続のリスト
[ ]	カテゴリのリスト
[ ]	フラグのリスト

表 2. リストの尺度を示すアイコン (続き)

アイコン	尺度
[]	名義のリスト
[]	順序のリスト

リストは、Analytic Server、地理空間、または可変長ファイルのいずれかのソース・ノードで SPSS Modeler にインポートするか、フィールド作成ノードまたは置換のフィールド操作ノードを使用してストリーム内で作成することができます。

リストと、集合および地理空間の尺度との相互作用について詳しくは、 12 ページの『リストス トレージおよび関連する尺度』を参照してください。

ストレージの変換: 置換ノードで `to_string` や `to_integer` などのさまざまな変換関数を使用して、フィールドのストレージを変換することができます。詳しくは、トピック 170 ページの『置換ノードを使ったストレージの変換』を参照してください。 変換関数および日付や時刻の値のような、入力に特別な型が必要なその他の関数は、「ストリームのプロパティー」ダイアログ・ボックスに指定されている現在の形式に依存します。例えば、値が *Jan 2003*、*Feb 2003* などの文字列フィールドを日付ストレージへ変換する場合、ストリームのデフォルトの日付形式として「**MON YYYY**」を選択します。変換関数は、フィールド作成ノードのフィールド作成計算時の一時変換でも、利用できます。また、フィールド作成ノードを使用して、カテゴリー値を含む文字列フィールドの読み取りなど、他の操作も実行できます。詳しくは、トピック 169 ページの『フィールド作成ノードを使用して値を再コード化する』を参照してください。

混在データの読み込み: 注意を要するのは、数値ストレージ (整数、実数、時間、タイムスタンプ、または日付のいずれか) を含むフィールドで読み込む場合、数値以外の値は、ヌル値またはシステム欠損値に設定されることです。これは、一部のアプリケーションと異なり、IBM SPSS Modeler では、1 つのフィールド内でストレージ・タイプが混在することは許されないためです。これを回避するには、ソース・ノードでストレージ・タイプを変更するか、必要な場合は外部アプリケーションで、データが混ざり合ったフィールドを文字列として読み込む必要があります。

フィールド入力形式 (固定長ファイル、可変長ファイル、ユーザー入力ノードのみ)

文字列と整数以外のすべてのストレージのデータ型の場合、ドロップダウン・リストを使用して選択したフィールドに対して、形式のオプションを指定できます。例えば、さまざまなロケールからのデータを結合する場合、1 つのフィールドの小数点区切り文字としてピリオドの指定が必要な場合がありますが、カンマが必要な場合もあります。

ソース・ノードで指定した入力オプションにより「ストリームのプロパティー」ダイアログ・ボックスで指定した形式オプションは無効になりますが、その後のストリームには反映されません。これらは、データの内容にしたがって入力を正確に解析するように作成されています。指定された形式は、データが IBM SPSS Modeler に読み込まれるときにそのデータを解析するのに使用され、IBM SPSS Modeler に読み込まれたあとでそのデータをフォーマットする方法を指定するものではありません。ストリームの他の場所でフィールドごとに形式オプションを指定するには、データ型ノードの「形式」タブを使用します。 詳しくは、トピック 157 ページの『フィールド形式の「設定」タブ』を参照してください。

オプションはストレージ・タイプによって異なります。例えば、実数の場合、小数点区切り文字として「ピリオド **(.)**」 または 「カンマ **(,)**」 を選択できます。タイムスタンプ・フィールドでは、ドロップダウン・リストで 「指定」 を選択すると、個別のダイアログ・ボックスが表示されます。 詳しくは、トピック 158 ページの『フィールド形式のオプションの設定』を参照してください。

すべてのストレージ・タイプに対して、「ストリームのデフォルト」 を指定し、インポートにストリームのデフォルト設定を使用できます。ストリームの設定は、「ストリームのプロパティー」ダイアログ・ボックスで指定します。

付加オプション

他にも「データ型」タブを使用して指定できるさまざまなオプションがあります。

- まだ現在のノードを通じて接続されていないデータ (学習データなど) のストレージ設定を表示するには、「未使用のフィールド設定を表示」 を選択します。古いフィールドを消去するには、「消去」 をクリックします。
- このダイアログ・ボックスで作業中は、任意の時点で「リフレッシュ」をクリックすると、フィールドがデータ・ソースから再ロードされます。これは、ソース・ノードへのデータ接続を変更したり、ダイアログ・ボックス中のタブ間を行き来して作業を行うような場合に役立ちます。

リストス トレージおよび関連する尺度

SPSS Modeler バージョン 17 で導入され、地理空間および集合の新しい尺度を処理するために、リスト ストレージ フィールドには、単一のレコードに対する複数の値が格納されます。リストは、大カッコ ([]) で囲みます。リストは、例えば `[1,2,4,16]` や `["abc", "def"]` などです。

リストは、3 つのソース・ノード (Analytic Server、地理空間、可変長ファイル) のいずれかで SPSS Modeler にインポートすることも、フィールド作成操作ノードまたはフィールド置換操作ノードを使用してストリーム内に作成することも、ランク付けされた条件の結合方法の使用時に結合ノードで生成することもできます。

リストは深さを持つと見なされます。例えば、`[1,3]` の形式の 1 組の大括弧で囲まれた項目を持つ単純なリストは、深さ 0 で IBM SPSS Modeler に記録されます。深さ 0 の単純なリストのほかに、入れ子になったリストを使用できます。その場合、リストのそれぞれの値もまたリストになります。

入れ子になったリストの深さは、関連する尺度によって異なります。「不明」の尺度の場合、深さの制限は設定されず、「集合」の尺度の場合、深さは 0 になります。「地理空間」の尺度の場合は、入れ子になった項目数に応じて、深さを 0 から 2 の範囲内で指定する必要があります。

深さが 0 のリストの場合は、地理空間または集合のいずれかの尺度を設定できます。これらの尺度はいずれも親尺度であり、「値」ダイアログボックスでサブ尺度の情報を設定します。集合のサブ尺度は、そのリストに含まれる要素の尺度を決定します。「不明」と「地理空間」の尺度を除くすべての尺度は、集合のサブ尺度として使用することができます。地理空間の尺度にはポイント、行ストリング、多角形、複数点、複数行ストリング、および多角形群の 6 つのサブ尺度があります。詳しくは、 148 ページの『地理空間のサブ尺度』を参照してください。

注**:** 「集合」の尺度は、深さが 0 のリストでのみ使用することができます。「地理空間」の尺度は、最大深度が 2 のリストでのみ使用することができます。「不明」の尺度は、任意の深さのリストで使用することができます。

深さ 0 のリストと入れ子にしたリストの違いを、ポイントと行ストリングの地理空間サブ尺度の構造を使用して以下に例示します。

- ポイントの地理空間サブ尺度に、0 のフィールドの深さがある場合。

  `[1,3]` 2 つの座標

  `[1,3,-1]` 3 つの座標
- 行ストリングの地理空間サブ尺度に、1 のフィールドの深さがある場合。

  `[ [1,3], [5,0] ]` 2 つの座標

  `[ [1,3,-1], [5,0,8] ]` 3 つの座標

ポイントのフィールド (深さ 0) は通常のリストであり、それぞれの値が 2 つまたは 3 つの座標で構成されています。行ストリングのフィールド (深さ 1) はポイントのリストであり、それぞれのポイントがさらに一連のリスト値で構成されています。

リストの作成について詳しくは、 166 ページの『リスト フィールドまたは地理空間フィールドの作成』を参照してください。

サポートされない制御文字

SPSS Modeler の一部のプロセスでは、各種制御文字が含まれるデータを扱うことができません。データでこれらの文字が使用されていると、以下の例のようなエラー・メッセージが表示される場合があります。

```
Unsupported control characters found in values of field {0}
```

サポートされない文字は、`0x0` から `0x3F` までの文字と、`0x7F` の文字です。ただし、タブ (`0x9(¥t)`)、改行 (`0xA(¥n)`)、および復帰 (`0xD(¥r)`) の各文字は問題の原因にはなりません。

ストリーム内で、ソース・ノードの後に、サポートされない文字に関連するエラー・メッセージが表示された場合は、フィルター・ノードと CLEM 式 **`stripctrlchars`** を使用してそれらの文字を置き換えます。

Analytic Server ソース・ノード

Analytic Server ソースにより、Hadoop 分散ファイル・システム (HDFS) でストリームを実行することができます。Analytic Server データ・ソースの情報は、以下のようなさまざまな場所から得られたものです。

- HDFS 上のテキストファイル
- データベース
- HCatalog

通常、Analytic Server 入力を持つストリームは HDFS で実行します。ただし、HDFS での実行がサポートされないノードがストリームに含まれる場合は、可能な限り多くのストリームを Analytic Server に「プッシュバック」し、SPSS Modeler Server が残りのストリームを処理します。非常に大きいデータ・セットの場合は、ストリーム内にサンプル・ノードを配置するなどの方法でサブサンプルを抽出する必要があります。

管理者が定義したデフォルト接続の代わりに独自の Analytic Server 接続を使用する場合は、「デフォルトの **Analytic Server** を使用 **(Use default Analytic Server)**」の選択を解除して、接続を選択します。複数の Analytic Server 接続のセットアップ方法について詳しくは、Analytic Server に接続中を参照してください。

データ・ソース**:** 自分または SPSS Modeler Server 管理者が既に接続を確立したと想定して、使用するデータを含むデータ・ソースを選択します。データ・ソースは、そのソースに関連付けられたファイルおよびメタデータを含みます。「選択」をクリックすると、使用可能なデータ・ソースのリストが表示されます。詳しくは、トピック『データ ソースの選択』を参照してください。

新しいデータ・ソースを作成するか既存のデータ・ソースを編集する必要がある場合は、「データ・ソース・エディターの起動**...**」をクリックします。

なお、複数の Analytic Server 接続を使用すると、データの流れの制御に役立ちます。例えば、Analytic Server 入力ノードおよびエクスポート・ノードを使用している場合、あるストリームのそれぞれのブランチ内に異なる Analytic Server 接続を使用することで、各ブランチの実行時にブランチがそれ自体の Analytic Server を使用して IBM SPSS Modeler Server にデータが引き出されないようにすることができます。1 つのブランチに複数の Analytic Server 接続が含まれている場合は、Analytic Server から IBM SPSS Modeler Server にデータが引き出されることに注意してください。制限などの詳細については、Analytic Server のストリームのプロパティーを参照してください。

データ ソースの選択

データ・ソース・テーブルには、使用可能なデータ・ソースのリストが表示されます。使用するソースを選択し、「**OK**」 をクリックします。

「所有者を表示」をクリックすると、データ・ソースの所有者が表示されます。

「フィルター基準」を使用すると、データ・ソースのリストを「キーワード」でフィルターに掛けてデータ・ソース名やデータ・ソースの説明とフィルター基準を照合したり、「所有者」でフィルターに掛けたりすることができます。フィルター基準としては、文字列、数値、またはワイルドカード文字 (後述) を組み合わせて入力することができます。検索文字列では大文字/小文字が区別されます。「更新」をクリックすると、データ・ソース・テーブルが更新されます。

_ 下線を使用すると、検索文字列の中の任意の 1 文字を表すことができます。

% パーセント記号を使用すると、検索文字列の中の 0 個以上の文字の並びを表すことができます。

資格情報の修正

Analytic Server にアクセスするための資格情報が SPSS Modeler Server にアクセスするための資格情報と異なる場合は、Analytic Server でストリームを実行するときに Analytic Server の資格情報を入力する必要があります。資格情報が不明な場合は、サーバー管理者に問い合わせてください。

サポートされるノード

多くの SPSS Modeler ノードで HDFS での実行がサポートされていますが、ノードによっては実行方法に差異があるものや、現在サポートされていないものもあります。このトピックでは、現行レベルのサポートについて詳しく説明します。

一般

- 一部の文字は、引用符で囲まれた Modeler フィールド名では通常受け入れられますが、Analytic Server では受け入れられません。
- Analytic Server で実行する Modeler ストリームは、1 つ以上の Analytic Server ソース・ノードで開始し、単一のモデル作成ノードまたは Analytic Server エクスポート ノードで終了しなければなりません。

- 連続型対象のストレージを、整数ではなく実数として設定することを推奨します。スコアリング・モデルは、連続型対象の出力データ・ファイルに常に実数値を書き込みます。これに対し、スコアの出力データ・モデルは、対象のストレージに従います。そのため、連続型対象が整数型のストレージを持つ場合は、書き込まれる値とスコアのデータ・モデルが一致せず、この不一致が原因で、スコアリングしたデータを読み込むときにエラーが発生します。
- フィールドの尺度が「地理空間」である場合、関数の @OFFSET はサポートされません。

ソース

- Analytic Server ソース・ノード以外のノードで開始するストリームはローカル環境で実行されます。

レコード設定

すべてのレコード操作がサポートされます。ただし、TS のストリーミング ノードと Space-Time-Box ノードは除きます。サポートされるノードの機能に関するより詳細な注意点を以下に示します。

条件抽出

- フィールド作成ノードがサポートする機能と同じ一連の機能をサポートします。

サンプリング

- ブロック・レベル・サンプリングはサポートされていません。
- 複雑なサンプリング方法はサポートされていません。
- **「サンプルを破棄」**を指定したときの最初の n 件のサンプリングはサポートされていません。
- N>20000 を指定したときの最初の n 件のサンプリングはサポートされていません。
- **「最大サンプル数」**が設定されていないときは、n 件ごとのサンプリングはサポートされていません。
- N *「最大サンプル数」> 20000 のときは、n 件ごとのサンプリングはサポートされていません。
- 無作為 % ブロック レベルのサンプリングはサポートされていません。
- 現在、無作為 % では、シードの提供がサポートされています。

レコード集計

- 連続キーはサポートされていません。データをソートするように設定されている既存のストリームを再使用し、この設定を集計ノードで使用する場合は、ソート ノードを削除するようにそのストリームを変更してください。
- 順序統計 (中央値、第 1 四分位数、第 3 四分位数) は概算され、「最適化」タブを通じてサポートされます。

ソート

- 「最適化」タブはサポートされていません。

分散環境では、ソート・ノードによって設定されたレコード順序を保持する操作の数が限られます。

- ソート・ノードの後にエクスポート・ノードを使用すると、ソートされたデータ・ソースが生成されます。
- レコードのサンプリングが「初めの **n** 件」であるサンプル・ノードをソート・ノードの後に使用すると、先頭から *N* 件のレコードが返されます。

一般にソート・ノードは、ソートされたレコードが必要になる操作のできるだけ近くに配置してください。

レコード結合

- 順序による結合はサポートされていません。
- 「最適化」タブはサポートされていません。
- 結合操作は、比較的速度が遅い操作です。HDFS に使用可能なスペースがある場合、まずデータ・ソースを結合し、以降のストリームでその結合済みソースを使用する方が、それぞれのストリームでデータ・ソースを結合するよりかなり早くなる場合があります。

R 変換(R)

ノードの R シンタックスは、レコード単位の操作で構成されている必要があります。

フィールド設定

すべてのフィールド操作がサポートされます。ただし、匿名化ノード、行列入替ノード、時間区分ノード、時系列ノードは除きます。サポートされるノードの機能に関するより詳細な注意点を以下に示します。

データの自動準備

- ノードの学習はサポートされていません。学習済み自動データ準備ノードでの、新しいデータへの変換の適用はサポートされます。

フィールド作成

- すべてのフィールド作成関数がサポートされます。ただし、シーケンス関数は除きます。
- カウント型としてのフィールドの新規作成は、本質的にシーケンス操作であるため、サポートされていません。
- 分割フィールドを作成し、それらのフィールドを同じストリームで分割として使用することはできません。分割フィールドを作成するストリームと、そのフィールドを分割として使用するストリームの 2 つのストリームを作成する必要があります。

置換

- フィールド作成ノードがサポートする機能と同じ一連の機能をサポートします。

データ分割

以下の機能はサポートされていません。

- 最適カテゴリー化
- ランク
- 分位 -> 分位方法 (Tiling): 値の合計
- 分位 -> 同順位: 「現在のまま保持」および「無作為割当」
- 分位 -> カスタム N: 100 を超える値と、100 % N が 0 に等しくない任意の N 値。

RFM 分析

- 同順位の処理方法としての「現在のまま保持」オプションはサポートされません。RFM リーセンシー・スコア、フリクエンシー・スコア、およびマネタリー・スコアは、同じデータから Modeler によって計算されたスコアと常に一致するわけではありません。スコアの範囲は同じですが、スコアの割り当て (ビン番号) は 1 だけ異なる場合があります。

グラフ作成

グラフ作成ノードはすべてサポート対象です。

モデリング

サポートされるモデル作成ノードは、時系列、TCM、Isotonic-AS、拡張モデル、Tree-AS、C&R ツリー、Quest、CHAID、線型、Linear-AS、ニューラル ネットワーク、GLE、LSVM、TwoStep-AS、ランダム ツリー、STP、アソシエーション ルール、XGBoost-AS、ランダム フォレスト、K-Means-AS です。これらのノードの機能に関する注意を以下に示します。

線型　ビッグ・データに対するモデルを作成する場合は、通常、目的を「特に大きなデータセット」に変更するか、分割を指定してください。

- 既存の PSM モデルの学習継続はサポートされていません。
- 「標準」のモデル作成目的は、分割フィールドが定義されており、各分割のレコード数が大きすぎない場合にのみ推奨します。ここで、大きすぎるかどうかは、Hadoop クラスター内の個々のノードの性能に応じて判断します。対照的に、分割の定義が細かすぎてレコードが少なくなり、モデルを作成できない事態に陥らないように注意する必要もあります。
- 「ブースティング」の目的はサポートされていません。
- 「バギング」の目的はサポートされていません。
- レコードが少ない場合、「特に大きなデータセット」の目的は推奨しません。多くの場合、モデルが作成されないか、質の悪いモデルが作成されてしまいます。
- 自動データ準備はサポートされていません。欠損値が多いデータに対するモデルを作成するときに問題が発生する可能性があります。通常、それらの欠損値は自動データ準備の一環として代入されます。回避策は、選択した欠損値を代入するための詳細設定を指定してツリー・モデルまたはニューラル・ネットワークを使用することです。
- 分割モデルの場合、精度統計値は計算されません。

ニューラル・ネットワーク

ビッグ・データに対するモデルを作成する場合は、通常、目的を「特に大きなデータセット」に変更するか、分割を指定してください。

- 既存の標準モデルまたは PSM モデルの学習継続はサポートされていません。
- 「標準」のモデル作成目的は、分割フィールドが定義されており、各分割のレコード数が大きすぎない場合にのみ推奨します。ここで、大きすぎるかどうかは、Hadoop クラスター内の個々のノードの性能に応じて判断します。対照的に、分割の定義が細かすぎてレコードが少なくなり、モデルを作成できない事態に陥らないように注意する必要もあります。
- 「ブースティング」の目的はサポートされていません。
- 「バギング」の目的はサポートされていません。
- レコードが少ない場合、「特に大きなデータセット」の目的は推奨しません。多くの場合、モデルが作成されないか、質の悪いモデルが作成されてしまいます。
- データに多数の欠損値が存在する場合は、詳細設定を使用して欠損値を代入してください。
- 分割モデルの場合、精度統計値は計算されません。

C&R ツリー、**CHAID**、および **Quest**

ビッグ・データに対するモデルを作成する場合は、通常、目的を「特に大きなデータセット」に変更するか、分割を指定してください。

- 既存の PSM モデルの学習継続はサポートされていません。
- 「標準」のモデル作成目的は、分割フィールドが定義されており、各分割のレコード数が大きすぎない場合にのみ推奨します。ここで、大きすぎるかどうかは、Hadoop クラスター内の個々のノードの性能に応じて判断します。対照的に、分割の定義が細かすぎてレコードが少なくなり、モデルを作成できない事態に陥らないように注意する必要もあります。
- 「ブースティング」の目的はサポートされていません。
- 「バギング」の目的はサポートされていません。
- レコードが少ない場合、「特に大きなデータセット」の目的は推奨しません。多くの場合、モデルが作成されないか、質の悪いモデルが作成されてしまいます。
- 対話式セッションはサポートされていません。
- 分割モデルの場合、精度統計値は計算されません。
- 分割フィールドが存在する場合、Modeler でローカルに作成されるツリー モデルは、Analytic Server で作成されるツリー モデルとは多少異なり、その結果、異なるスコアが生成されます。どちらのケースのアルゴリズムも有効です。Analytic Server で使用されるアルゴリズムの方が新しくなっています。ツリー アルゴリズムには多数のヒューリスティック規則が使用される傾向があるため、この 2 つのコンポーネント間での差異は正常なものです。

モデル・スコアリング

モデル作成用にサポートされているすべてのモデルが、スコアリングでもサポートされます。また、次のノード用にローカルに作成されるモデル ナゲットが、スコアリングでサポートされます: C&RT、Quest、CHAID、線型、ニューラル ネットワーク (モデルが標準、ブースティング、バギング、非常に大きなデータセットに該当するかどうかは無関係)、回帰、C5.0、ロジスティック、一般化線型、GLMM、Cox、SVM、ベイズ ネットワーク、TwoStep、KNN、ディシジョン リスト、判別分析、自己学習、異常値検出、Apriori、Carma、K-Means、Kohonen、R、およびテキスト マイニング。

- 未調整傾向も調整済み傾向もスコアリングされません。回避策として、次の式でフィールド作成ノードを使用して、手動で未調整傾向を計算することにより、同じ効果を得ることができます。
  if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif

R ナゲットの R シンタックスは、レコード単位の操作で構成されている必要があります。

出力 クロス集計、精度分析、データ検査、変換、グローバルの設定、記述統計、平均、およびテーブルの各ノードがサポートされています。サポートされるノードの機能に関するより詳細な注意点を以下に示します。

データ検査

データ検査ノードは、連続型フィールドのモードは生成できません。

平均値

平均ノードは、標準誤差および 95% 信頼区間は生成できません。

テーブル

テーブル ノードは、上流の操作の結果を含む一時 Analytic Server データ ソースの書き込みによってサポートされます。それにより、テーブル ノードはそのデータ ソースの内容をページ処理します。

エクスポート

ストリームは Analytic Server ソース・ノードで開始し、Analytic Server エクスポート・ノード

以外のエクスポート・ノードで終了することができますが、データは HDFS から SPSS Modeler Server に移動した後、最終的にエクスポート先に移動します。

データベース・ソース・ノード

データベース・ソース・ノードは、Microsoft SQL Server、Db2、Oracle など ODBC (開放型データベース接続) を使用する他のさまざまなパッケージからデータをインポートするのに使用できます。

データベースを読み書きするには、ODBC データ・ソースがインストールされていて、該当するデータベースに対して必要に応じて読み取り権限や書き込み権限が設定されている必要があります。IBM SPSS Data Access Pack には、この目的で使用できる ODBC ドライバが含まれています。また、これらのドライバは、ダウンロード サイトから入手できます。ODBC データ・ソースの権限の作成または設定についてわからないことがある場合は、データベース管理者に問い合わせてください。

サポートされている ODBC ドライバー

IBM SPSS Modeler での使用がサポートおよびテストされているデータベースおよび ODBC ドライバーの最新情報については、当社サポート・サイト (http://www.ibm.com/support) にある製品互換性マトリクスを参照してください。

ドライバーをインストールする場所

注: ODBC ドライバーは、処理が実行される各コンピューターにインストールして設定する必要があります。

- ローカル (スタンドアロン) モードで IBM SPSS Modeler を実行する場合は、ドライバーをローカル・コンピューターにインストールする必要があります。
- IBM SPSS Modeler をリモートの IBM SPSS Modeler Server に対して分散モードで実行する場合、ODBC ドライバーは IBM SPSS Modeler Server がインストールされたコンピューターにインストールする必要があります。UNIX システムの IBM SPSS Modeler Server を使用している場合は、このセクションの『UNIX システムの ODBC ドライバーの設定』も参照してください。
- IBM SPSS Modeler と IBM SPSS Modeler Server の両方から同じデータ・ソースにアクセスする必要がある場合、ODBC ドライバーは両方のコンピューターにインストールする必要があります。
- 端末サービスを介して IBM SPSS Modeler を実行する場合、ODBC ドライバーは IBM SPSS Modeler がインストールされた端末サービス・サーバーにインストールする必要があります。

データベースのデータへのアクセス

データベースのデータにアクセスするには、以下のステップを実行します。

- 使用するデータベースに ODBC ドライバーをインストールして、データ・ソースを構成します。
- 「データベース・ノード」ダイアログ・ボックスで、テーブル・モードまたは SQL クエリー・モードを使用してデータベースに接続します。
- データベースからテーブルを選択します。
- 「データベース・ノード」ダイアログ・ボックスのタブを使用して、データ・フィールドの使用タイプを変更したり、フィールドをフィルタリングすることができます。

上記のステップについての詳細は関連資料のトピックで説明しています。

注: SPSS Modeler からデータベースのストアド プロシージャ (SP) を呼び出した場合に、期待していた SP の出力ではなく、`RowsAffected` という名前の単一の出力フィールドが返されることがあります。この

現象は、SP の出力データ モデルを判別できるだけの十分な情報が ODBC から返されなかった場合に発生します。SPSS Modeler では、出力を返す SP のサポートが限定されているため、SP を使用する代わりに、SP から SELECT を抽出して以下のいずれかの操作を使用することをお勧めします。

- SELECT に基づくビューを作成し、データベース・ソース・ノードでそのビューを選択する。
- データベース・ソース・ノードで直接 SELECT を使用する。

データベース・ノード・オプションの設定

「データベース入力ノード」ダイアログ・ボックスの「データ」タブにあるオプションを使用して、データベースにアクセスし、選択したテーブルからデータを読み込むことができます。

「モード」。ダイアログ・ボックスのコントロールを使用してテーブルに接続するには、「テーブル」 を選択します。

SQL を使用して、下で選択したデータベースに問い合わせるには、「**SQL** クエリー」 を選択します。 詳しくは、トピック 27 ページの『データベースの照会』を参照してください。

データ・ソース **:**テーブルおよび SQL クエリー・モードのどちらでも、データ・ソース・フィールドに名前を入力するか、またはドロップダウン・リストから 「新規データベース接続の追加」 を選択できます。

ダイアログ・ボックスを使用して、データベースに接続してテーブルを選択するには、次のオプションを使用します。

テーブル名 **:** アクセスするテーブル名が既知の場合に、「テーブル名」フィールドにその名前を入力します。そうでない場合は、「選択」 ボタンをクリックして、利用可能なテーブルを表示したダイアログ・ボックスを開きます。

表および列名を引用符で囲む**:** クエリーをデータベースに送信するときにテーブル名と列名を引用符で囲むかどうかを指定します (例えば、テーブル名と列名にスペースや句読点が含まれているような場合)。

- 「必要に応じて」 オプションを選択すると、非標準文字が含まれている場合にのみ、テーブル名とフィールド名が引用符で囲まれます。非標準文字とは、非 ASCII 文字、スペース文字、およびピリオド (.) 以外の非英数文字を指します。
- すべて のテーブル名とフィールド名を引用符で囲む場合は、「常時」を選択します。
- テーブル名とフィールド名を引用符で囲まない 場合は、「しない」を選択します。

前後のスペースを削除 **:** 文字列の前後のスペースを除去する場合に選択します。

注: SQL プッシュバックを使用する文字列と使用しない文字列と比較すると、接尾空白を含むさまざまな結果を生成する場合があります。

Oracle からの空の文字列の読み取り**:** Oracle データベースとの間で読み書きするときには、Oracle が IBM SPSS Modeler やその他のほとんどのデータベースとは異なり、空の文字列値をヌル値と同様に処理および格納することに注意してください。つまり、Oracle データベースから抽出されたデータは、同じデータがファイルやその他のデータベースから抽出された場合とは異なって動作し、また異なる結果が返ることがあります。

データベース接続の追加

データベースを開くには、最初に、接続するデータ・ソースを選択します。「データ」タブの「データ・ソース」ドロップダウン・リストで 「新規データベース接続の追加」 を選択します。

「データベース接続」ダイアログ・ボックスが表示されます。

注: 別の方法として、メイン メニューから「ツール」 > 「データベース...」を選択してこのダイアログを開くこともできます。

データ・ソース。使用できるデータ・ソースの一覧が表示されています。目的のデータベースが表示されていない場合は、リストを下方へスクロールしてください。データ・ソースを選択してパスワードを入力したら、「接続」 をクリックします。リストを更新するには、「リフレッシュ」 をクリックします。

ユーザー名およびパスワード: データ・ソースがパスワードで保護されている場合は、ユーザー名および関連付けられているパスワードを入力します。

資格情報。資格情報を IBM SPSS Collaboration and Deployment Services で構成した場合は、このオプションを選択すると、リポジトリー内の資格情報を参照することができます。資格情報のユーザー名とパスワードは、データベースにアクセスするために必要なユーザー名とパスワードに一致している必要があります。

接続。現在接続しているデータベースが表示されます。

- デフォルト。オプションで、 1 つの接続をデフォルトとして選択できます。これにより、データベース入力ノードまたはエクスポートノードはこの接続をデータ・ソースとして事前定義します。必要に応じて編集可能です。
- 保存。オプションで、後続のセッションで再度表示する接続を選択します。
- データ・ソース。現在接続しているデータベースの接続文字列。
- プリセット。(* 文字を使用して) プリセット値がデータベース接続に指定されているかどうかを示します。プリセット値を指定するには、データベース接続に対応する行のこの列をクリックして、リストから「指定」を選択します。 詳しくは、トピック 23 ページの『データベース接続のプリセット値の指定』を参照してください。

接続を削除するには、目的の接続をリストから選択し、「削除」 をクリックします。

選択が完了したら、「**OK**」 をクリックします。

データベースを読み書きするには、ODBC データ・ソースがインストールされていて、該当するデータベースに対して必要に応じて読み取り権限や書き込み権限が設定されている必要があります。IBM SPSS Data Access Pack には、この目的で使用できる ODBC ドライバが含まれています。また、これらのドライバは、ダウンロード サイトから入手できます。ODBC データ・ソースの権限の作成または設定についてわからないことがある場合は、データベース管理者に問い合わせてください。

サポートされている ODBC ドライバー

IBM SPSS Modeler での使用がサポートおよびテストされているデータベースおよび ODBC ドライバーの最新情報については、当社サポート・サイト (http://www.ibm.com/support) にある製品互換性マトリクスを参照してください。

ドライバーをインストールする場所

注: ODBC ドライバーは、処理が実行される各コンピューターにインストールして設定する必要があります。

- ローカル (スタンドアロン) モードで IBM SPSS Modeler を実行する場合は、ドライバーをローカル・コンピューターにインストールする必要があります。

- IBM SPSS Modeler をリモートの IBM SPSS Modeler Server に対して分散モードで実行する場合、ODBC ドライバーは IBM SPSS Modeler Server がインストールされたコンピューターにインストールする必要があります。UNIX システムの IBM SPSS Modeler Server を使用している場合は、このセクションの『UNIX システムの ODBC ドライバーの設定』も参照してください。
- IBM SPSS Modeler と IBM SPSS Modeler Server の両方から同じデータ・ソースにアクセスする必要がある場合、ODBC ドライバーは両方のコンピューターにインストールする必要があります。
- 端末サービスを介して IBM SPSS Modeler を実行する場合、ODBC ドライバーは IBM SPSS Modeler がインストールされた端末サービス・サーバーにインストールする必要があります。

UNIX システムの ODBC ドライバーの設定

デフォルトでは、DataDirect Driver Manager は UNIX システムの IBM SPSS Modeler Server 向けには設定されていません。DataDirect Driver Manager ロードするよう UNIX を設定するには、次のコマンドを入力します。

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

これにより、デフォルトのリンクが削除され、DataDirect Driver Manager へのリンクを作成します。

注: 一部のデータベースでは、SAP HANA ドライバーまたは IBM Db2 CLI ドライバーを使用するために UTF16 ドライバー ラッパーが必要です。DashDB には、IBM Db2 CLI ドライバーが必要です。UTF16 ドライバー ラッパーのリンクを作成するために、代わりに以下のコマンドを入力します。

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

SPSS Modeler Server を構成するには、以下の手順を実行します。

1. `modelersrv.sh` に以下の行を追加することにより、IBM SPSS Data Access Pack の `odbc.sh` 環境ファイルをソースとして使用するように SPSS Modeler Server の起動スクリプト `modelersrv.sh` を構成します。

   ```
   . /<pathtoSDAPinstall>/odbc.sh
   ```

   `<pathtoSDAPinstall>` は、IBM SPSS Data Access Pack のインストール済み環境の絶対パスです。
2. SPSS Modeler Server を再起動します。

さらに、SAP HANA および IBM Db2 の場合にのみ、接続中にバッファーがオーバーフローしないように、`odbc.ini` ファイル内の DSN に以下のパラメーター定義を追加します。

```
DriverUnicodeType=1
```

注: `libspssodbc_datadirect_utf16.so` ラッパーは、SPSS Modeler Server がサポートする他の ODBC ドライバーにも対応しています。

データベースで発生する可能性のある問題

使用するデータベースに応じて、いくつかの問題が発生する可能性があるため、注意が必要です。

IBM Db2

Db2 データベースからデータを読み取るストリーム内のノードをキャッシュに入れようとすると、次のエラー・メッセージが表示される場合があります。

```
A default table space could not be found with a pagesize of at least 4096 that authorization
 ID TEST is authorized to use
```

SPSS Modeler でデータベース内キャッシングが正常に機能するように Db2 を構成するには、データベース管理者が「ユーザー一時」テーブルスペースを作成し、このテーブルスペースへのアクセス権限を関連する Db2 アカウントに付与する必要があります。

新しいテーブルスペースでは、ページサイズとして 32768 を使用することをお勧めしています。このサイズを使用すると、正常にキャッシュできるフィールド数の制限が増加するからです。

IBM Db2 for z/OS

- 確信度を有効にし、生成された SQL を使用してアルゴリズムのサブセットをスコアリングすると、実行時にエラーが返される可能性があります。これは、Db2 for z/OS に固有の問題です。この問題を修正するには、SPSS Modeler Server Scoring Adapter for Db2 on z/OS を使用します。
- Db2 for z/OS に対してストリームを実行する場合、アイドル・データベース接続のタイムアウトを有効にしていて、その設定値が小さすぎると、データベース・エラーが発生することがあります。Db2 for z/OS バージョン 8 では、デフォルト設定がタイムアウトなしから 2 分へと変更されました。解決策として、Db2 システム・パラメーター `IDLE THREAD TIMEOUT (IDTHTOIN)` の値を増やすか、または 0 にリセットします。

Oracle

集計ノードを含むストリームを実行する場合、Oracle データベースに SQL をプッシュバックするときに第 1 四分位数と第 3 四分位数について返される値が、ネイティブ・モードで返される値と異なる場合があります。

データベース接続のプリセット値の指定

一部のデータベースで、データベース接続のデフォルト設定を指定できます。設定はすべてデータベース・エクスポートに適用されます。

この機能をサポートするデータベースの種類は次のとおりです。

- SQL Server Enterprise Edition および Developer Edition。詳しくは、トピック『SQL Server の設定』を参照してください。
- Oracle Enterprise Edition または Personal Edition。詳しくは、トピック 24 ページの『Oracle の設定』を参照してください。
- IBM Db2 for z/OS と Teradata は、いずれも同じ方法でデータベースやスキーマに接続します。詳しくは、トピック 24 ページの『IBM Db2 for z/OS、IBM Db2 LUW、および Teradata の設定』を参照してください。

この機能をサポートしていないデータベースやスキーマに接続すると、「このデータベース接続に設定できるプリセットはありません」というメッセージが表示されます。

SQL Server の設定

これらの設定は、SQL Server Enterprise Edition および Developer Edition に表示されます。

圧縮を使用: 選択した場合は、圧縮によるエクスポート用のテーブルを作成します。

圧縮レベル: 圧縮のレベルを選択します。

- 行: 行レベルの圧縮を有効にします (例えば、SQL の CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW); と同等)。
- ページ: ページ・レベルの圧縮を有効にします (例えば、SQL の CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);)。

Oracle の設定

Oracle の設定 - 基本オプション

これらの設定は、基本オプションを使用する Oracle の Enterprise Edition または Personal Edition に表示されます。

圧縮を使用: 選択した場合は、圧縮によるエクスポート用のテーブルを作成します。

圧縮レベル: 圧縮のレベルを選択します。

- デフォルト: デフォルトの圧縮を有効にします (例えば SQL の CREATE TABLE MYTABLE(...) COMPRESS;)。この場合、「基本」 オプションと同じ効果があります。
- 基本 : このビューには、予測値間の関係に加え、対象値と最も重要な予測値間の関係を表示するノードのネットワーク グラフが含まれています。基本的な圧縮を有効にします (例えば、SQL の CREATE TABLE MYTABLE(...) COMPRESS BASIC;)。

Oracle の設定 - 高度なオプション

これらの設定は、高度なオプションを使用する Oracle の Enterprise Edition または Personal Edition に表示されます。

圧縮を使用: 選択した場合は、圧縮によるエクスポート用のテーブルを作成します。

圧縮レベル: 圧縮のレベルを選択します。

- デフォルト: デフォルトの圧縮を有効にします (例えば SQL の CREATE TABLE MYTABLE(...) COMPRESS;)。この場合、「基本」 オプションと同じ効果があります。
- 基本 : このビューには、予測値間の関係に加え、対象値と最も重要な予測値間の関係を表示するノードのネットワーク グラフが含まれています。基本的な圧縮を有効にします (例えば、SQL の CREATE TABLE MYTABLE(...) COMPRESS BASIC;)。
- **OLTP**: OLTP の圧縮を有効にします (例えば、SQL の CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP;)。
- クエリー低/高: (Exadata サーバーのみ) クエリーの Hybrid Columnar Compression を有効にします (例えば、SQL の CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW; や CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH;)。データ・ウェアハウス環境では QUERY の圧縮が役に立ちます。HIGH は、LOW より圧縮率が高くなります。
- アーカイブ低/高: (Exadata サーバーのみ) アーカイブの Hybrid Columnar Compression を有効にします (例えば、SQL の CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW; や CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH;)。長期間格納されるデータの圧縮には ARCHIVE の圧縮が役に立ちます。HIGH は、LOW より圧縮率が高くなります。

IBM Db2 for z/OS、IBM Db2 LUW、および Teradata の設定

IBM Db2 for z/OS、IBM Db2 LUW、または Teradata のプリセットを指定する際に、以下の項目を選択するためのプロンプトが表示されます。

「**Server Scoring Adapter** データベースを使用」または「**Server Scoring Adapter** データベースを使用」。どちらかを選択すると、「**Server Scoring Adapter** データベース」オプションまたは「**Server Scoring Adapter** スキーマ」オプションが有効になります。

「**Server Scoring Adapter** データベース」または「**Server Scoring Adapter** スキーマ」: ドロップダウン リストで、必要な接続を選択します。

さらに、Teradata の場合は、クエリー バンド化の詳細を設定して追加のメタデータを提供し、ワークロード管理、クエリーの照合、識別、解決、およびデータベース使用状況の追跡などの項目を支援することもできます。

クエリー バンド化のスペル: Teradata データベース接続の処理全体で 1 回だけクエリー バンド化を設定するか (「セッション」)、ストリームを実行するたびに設定するか (「トランザクション」) を選択します。

注**:** ストリームでクエリー バンド化を設定した場合は、そのストリームを別のマシンにコピーすると、バンド化が失われます。これを防ぐために、スクリプトを使用してストリームを実行し、スクリプトでキーワード *querybanding* を使用して必要な設定を適用することができます。

必要なデータベース権限

SPSS Modeler データベース機能を正しく機能させるには、以下の項目に対するアクセス権限を、使用されるすべてのユーザー ID に付与します。

Db2 LUW

SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABCONSTPKC

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSCAT.TABLESPACES

SYSCAT.SCHEMATA

Db2/z SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABCONSTPKC

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSIBM.SYSDUMMYU

SYSIBM.SYSPACKSTMT

Teradata

DBC.Functions

DBC.USERS

データベース・テーブルの選択

データ・ソースに接続したら、特定のテーブルまたはビューからフィールドをインポートできます。「データベース」ダイアログ・ボックスの「データ」タブで、「テーブル名」フィールドにテーブル名を入力するか、「選択」をクリックして「表/ビューの選択」ダイアログ・ボックスを開き、使用可能なテーブルとビューのリストを表示します。

テーブル所有者の表示 **:** 指定したデータ・ソースがテーブルにアクセスするためにテーブルの所有者を指定する必要がある場合に、選択します。指定する必要がないデータ・ソースの場合は、このオプションの選択を解除してください。

注**:** 通常、SAS データベースと Oracle データベースでは、テーブル所有者を表示する必要があります。

テーブル**/**ビュー **:** インポートするテーブルまたはビューを選択します。

表示 **:** 現在接続しているデータ・ソースの列の一覧を表示します。次のいずれかのオプションを選択して、利用できるテーブルのビューをカスタマイズできます。

- データベース・ユーザーが作成した通常のデータベース・テーブルを表示する場合は、「ユーザー テーブル」 をクリックします。
- システムが所有するテーブル (インデックス詳細などのデータベースに関する情報を提供するテーブル) を表示するには、「システム テーブル」をクリックします。このオプションは、Excel データベースで

使用されるタブを表示する場合に必要になります(独立した Excel ソース・ノードも利用可能です)。 詳しくは、トピック 48 ページの『Excel ソース・ノード』を参照してください。

- 1 つまたは複数の通常のテーブルを含むクエリーに基づいた仮想テーブルを表示するには、「表示」をクリックします。
- 既存のテーブルに対してデータベースで作成されたシノニムを表示するには、「シノニム」をクリックします。

名前/所有者フィルター : これらのフィールドを使用すると、名前または所有者で表示されたテーブルのリストにフィルターを適用できます。例えば、SYS と入力すると、この所有者のテーブルのみがリストに表示されます。ワイルドカード検索では、アンダースコア (_) を任意の 1 文字に、パーセント記号 (%) を 0 個以上の文字の並びに表現するために使用できます。

デフォルト値に設定 : 現在の設定を現在のユーザーのデフォルトとして保存します。この設定は、後でユーザーが新しいテーブル選択ダイアログ・ボックスを開いたときに、同じデータ・ソース名とユーザー・ログインの場合に限り 復元されます。

データベースの照会

データ・ソースに接続したら、SQL クエリーを使用してフィールドをインポートできます。メイン・ダイアログ・ボックスから、接続モードとして「**SQL** クエリー」を選択します。ダイアログ・ボックスにクエリー・エディター・ウィンドウが追加されます。クエリー・エディターを使用して、結果セットがデータ・ストリームに読み込まれる SQL クエリーの作成やロードなどの作業を行うことができます。

複数の SQL クエリーを指定する場合、セミコロン (;) で区切って、複数の SELECT 文がないようにします。

クエリー・エディター・ウィンドウをキャンセルして閉じるには、接続モードとして「テーブル」を選択します。

SQL クエリに SPSS Modeler ストリーム・パラメーター (ユーザ定義変数の一種) を含めることができます。 詳しくは、トピック 28 ページの『SQL クエリーのストリーム・パラメーターの使用』を参照してください。

クエリーのロード: 以前保存したクエリーをロードするファイル・ブラウザーを表示する場合にクリックします。

クエリーの保存: 現在のクエリーを保存できる「クエリーの保存」ダイアログ・ボックスを表示する場合にクリックします。

デフォルト値のインポート: これをクリックすると、ダイアログ・ボックスで選択されているテーブルと列を使用して自動的に作成されたサンプルの SQL SELECT ステートメントがインポートされます。

クリア: 作業領域の内容を消去します。作業を最初からやり直す場合に使用します。

テキスト分割: デフォルト・オプション「なし **(Never)**」を使用すると、クエリー全体がデータベースに送信されます。あるいは、「必要に応じて」を選択することもできます。これを使用すると、SPSS Modeler は、クエリーを構文解析して、1 つずつデータベースに送信する必要がある SQL ステートメントが存在するかどうかを識別しようとします。

SQL クエリーのストリーム・パラメーターの使用

フィールドをインポートする SQL クエリーを作成する場合、以前定義された SPSS Modeler ストリーム・パラメーターを含めることができます。すべての種類のパラメーターをサポートできます。

次の表では、SQL クエリーのストリーム・パラメーターのいくつかの例をどのように解釈するかについて示しています。

表 3. ストリーム・パラメーターの例：

ストリーム・パラメーター名 (例)	ストレージ	ストリーム・パラメーター値	解釈
PString	文字列	ss	'ss'
PInt	整数	5	5
PReal	実数	5.5	5.5
PTime	時間	23:05:01	t{'23:05:01'}
PDate	日付	2011-03-02	d{'2011-03-02'}
PTimeStamp	タイムスタンプ	2011-03-02 23:05:01	ts{'2011-03-02 23:05:01'}
PColumn	不明	IntValue	IntValue

SQL クエリーでは、主に '$P-<parameter_name>' を使用して、CLEM 式と同じ方法でストリーム・パラメーターを指定します。<parameter_name> は、ストリーム・パラメーター用に定義された名前です。

フィールドを参照する場合、ストレージ・タイプは Unknown に指定、パラメーター値は必要に応じて引用符で囲む必要があります。SQL クエリーを入力した場合の例は以下のとおりです。

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

上記は以下のように評価されます。

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

PColumn パラメーターを使用して IntValue フィールドを参照する場合、以下のようにクエリーを指定して同じ結果を取得する必要があります。

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

可変長ファイル・ノード

可変長ファイル・ノードを使用すると、区切りテキスト・ファイルとも呼ばれる可変長フィールド・テキスト・ファイル (フィールド数は一定だが各フィールド中の文字数が異なるレコードを含むファイル) からデータをインポートできます。このノードは、固定長のヘッダー・テキストや特定の種類の注釈があるファイルにも使用できます。レコードは 1 回に 1 つずつ読み込まれ、ファイル全体が読み込まれるまでストリームを通過します。

地理空間データの読み込みに関する注意

ノードが地理空間データを含み、ノードがフラット ファイルからのエクスポートによって作成された場合は、追加のステップを実行して地理空間メタデータをセットアップする必要があります。詳しくは、 31 ページの『可変長ファイル ノードへの地理空間データのインポート』を参照してください。

区切りテキスト データの読み込み時の注意

- レコードは、各行の終わりで改行文字で区切る必要があります。改行文字は、(フィールド名または値内など) 他の目的で使用することはできません。理想的には前後のスペースを除去してスペースを少なくする必要がありますが、それほど重要ではありません。オプションで、ノードによってこれらのスペースを除去することができます。
- フィールドは、カンマまたは区切り文字としてのみ使用されるそのほかの文字によって区切られる必要があります。区切り文字としてのみ使用される文字はフィールド名または値には使用されません。このようにすることができない場合、フィールド名またはテキスト値に二重引用符が含まれていなければ、すべてのテキスト フィールドを二重引用符で囲むことができます。フィールド名または値に二重引用符が使用されている場合、こちらも同様に、値の中で他に単一引用符が使用されていなければ、代わりに単一引用符でテキスト フィールドを囲むことができます。単一引用符および二重引用符のいずれも使用することができない場合、テキスト値を修正して区切り文字、単一引用符または二重引用符のいずれかを削除または置換する必要があります。
- ヘッダー行を含む各行には、同じ数のフィールドが含まれています。
- 最初の行にはフィールド名が含まれています。含まれていない場合、 「ファイルからフィールド名を取得」を選択解除すると、各フィールドに「フィールド 1」、「フィールド 2」などの一般名が付けられます。
- 2 行目にはデータの最初のレコードが含まれている必要があります。空白の行またはコメントはありません。
- 千単位の区切り文字やグループ化記号を数値で使用することはできません。例えば、「3,000.00」のようなカンマは使用しないでください。小数点 (米国や英国におけるピリオドやフルストップ) の使用は、必要な場合に限ってください。
- 日付や時間の値には、`DD/MM/YYYY` または `HH:MM:SS` など [ストリーム・オプション」ダイアログ・ボックスで認識される形式のいずれかの形式を使用する必要があります。ファイル内のすべての日付/時間フィールドは同じ形式を使用するのが理想的です。また、日付を含むフィールドでは該当するフィールド内のすべての値に同じ形式を使用する必要があります。

可変長ファイル・ノードのオプションの設定

「可変長ファイル・ノード」ダイアログ・ボックスの「ファイル」タブでオプションを設定します。

ファイル ファイルの名前を指定します。ファイル名を入力するか、省略符号ボタン (「**...**」) をクリックしてファイルを選択できます。ファイルを選択すると、ファイル パスが表示され、下のパネルに区切り文字付きでファイルの内容が表示されます。

データ・ソースから表示されたサンプル テキストは、次のコントロールを使用してコピー、貼り付けすることができます。EOL コメント文字およびユーザー指定の区切り文字。コピーと貼り付けには、それぞれ Ctrl-C および Ctrl-V を使用します。

ファイルからフィールド名を取得 デフォルトで選択されているこのオプションは、データ・ファイル中の最初の行を列のラベルとして取り扱います。最初の行が見出しではない場合は、このオプションを解除すると、データ・セット中の各フィールドにフィールド *1*、フィールド *2* のような数字の付けられた一般名が与えられます。

フィールド数を指定**:** 各レコードのフィールドの数を指定します。レコードが改行文字で終了していれば、フィールド数が自動的に検出されます。フィールド数を自分で設定することもできます。

ヘッダー文字をスキップ**:** 最初のレコードの先頭で無視する文字数を指定します。

EOL コメント文字**:** データ内で注釈を示す文字 (# や !など) を指定します。データ・ファイル内でこれらの文字がある場所から次の改行文字のある場所までは、すべて注釈になります。ただし、その改行文字は注釈に含まれずに無視されます。

前後のスペースを除去 **:** インポート時に文字列の前後のスペースを破棄する場合に選択します。

注*:* SQL プッシュバックを使用する文字列と使用しない文字列と比較すると、接尾空白を含むさまざまな結果を生成する場合があります。

不正な文字**:** データ入力から不正な文字を削除する場合に、「破棄」 を選択します。不正な文字を指定した記号 (1 文字だけ) で置換する場合は、「置換値」 を選択します。ヌル (0) 文字または指定されたエンコード方法に存在しない任意の文字が不正な文字になります。

エンコード**:** 使用するテキストのエンコード方法を指定します。サーバー・デフォルト、システム・デフォルト、UTF-8 から選択できます。

- システム・デフォルトは、Windows のコントロール・パネル (分散モードで実行している場合はサーバー・コンピューター) で指定できます。
- デフォルトは、「ストリーム・プロパティー」ダイアログ・ボックスで指定されます。

小数点記号 データ ソースで使用する小数点区切り文字の種類を選択します。「ストリームのデフォルト」は、「ストリームのプロパティー」ダイアログ・ボックスの「オプション」タブで選択された文字です。これを使用しない場合は、「ピリオド **(.)**」 または 「カンマ **(,)**」 を選択すると、その文字を小数点区切り文字として、このダイアログ・ボックス中のすべてのデータを読み込みます。

行区切り文字は改行文字です フィールド区切りの代わりに、改行文字を行の区切り文字として使用するには、このオプションを選択します。例えば、行が折り返して表示される行に奇数の区切り文字がある場合役立つ場合があります。このオプションを選択した場合、「区切り文字」リストの 「改行」 は選択できません。

注**:** このオプションを選択した場合、データ行の末尾の空白値は除去されます。

区切り文字**:** このコントロール用に表示されたチェック・ボックスを使用して、カンマ (,) などの、ファイル内のフィールドの境界を定義する文字を指定できます。複数の区切り文字を使用するレコードの場合、「, |」のように複数の区切り文字を指定することもできます。デフォルトの区切り文字はカンマです。

注 : カンマが桁区切り文字としても定義されている場合、ここでのデフォルト設定は使用されません。この場合、カンマはフィールドの区切り文字と桁区切り記号の両方であるため、区切り文字リストから 「その他」 を選択します。次に、手動で入力フィールドにカンマを指定します。

隣接する複数の空白文字を単一の区切り文字として認識する場合は、「複数の空白区切り文字を許可」を選択します。例えば、あるデータ値の後に 4 つのスペースが続き、その後に別のデータ値が続いている場合は、5 つのフィールドではなく、2 つのフィールドとして扱われます。

列およびデータ型についてスキャンする行 指定したデータ型をスキャンする行および列数を指定します。

自動的に日付と時間を認識します IBM SPSS Modeler がデータ項目を自動的に日付または時刻として認識できるようにするには、このチェック ボックスを選択します。例えば、07-11-1965 などのエントリーを日付として識別し、02:35:58 を時刻として認識します。ただし、07111965 や 023558 のようなあいまいなエントリーは、数値の間に区切り文字がないため、整数として表示されます。

注**:** 以前のバージョンの IBM SPSS Modeler のデータ・ファイルを使用する場合に考えられるデータ上の問題を回避するために、13 より前のバージョンで保存された情報についてはデフォルトでこのボックスがオフになります。

大括弧をリストとして扱う このチェック ボックスを選択すると、左大括弧と右大括弧で囲まれたデータにコンマや二重引用符などの区切り文字が含まれていても、データが単一の値として扱われます。例えば、2 次元または 3 次元の地理空間データにおいて、大括弧で囲んだ座標を単一のリスト項目として処理する場合が該当します。詳しくは、『可変長ファイル ノードへの地理空間データのインポート』を参照してください。

引用符。ドロップダウン・リストを使用して、インポート時に単一引用符および二重引用符をどのように取り扱うかを指定できます。すべての引用符を 「破棄」、フィールド値として引用符を 「テキストとして含める」、または 「ペアで破棄」 を選択して、引用符のペアを組み合わせて破棄することができます。対応する引用符がない場合は、エラー・メッセージが表示されます。「破棄」 と 「ペアで破棄」 では、フィールド値を文字列として (引用符なしで) 保存します。

注**:** 「ペアで破棄」を使用すると、スペースは保持されます。「破棄」を使用すると、引用符の内側と外側の後続スペースは除去されます (例えば、**' " ab c" , "d ef " , " gh i " '** は、**'ab c, d ef, gh i'**) となります。「テキストとして含む」を使用すると、引用符は正規文字として扱われるため、前後のスペースは必然的に除去されます。

このダイアログ・ボックスで作業中は、任意の時点で 「リフレッシュ」 をクリックすると、フィールドがデータ・ソースから再ロードされます。これは、ソース・ノードへのデータ接続を変更したり、ダイアログ・ボックス内のタブ間を行き来して作業を行うような場合に役立ちます。

可変長ファイル ノードへの地理空間データのインポート

フラット ファイルからのエクスポートとして作成されたノード内に地理空間データが存在し、作成元のストリームでこのノードが使用されている場合、このノードには地理空間メタデータが存在しているため、これ以上の構成ステップは必要ありません。

しかし、ノードをエクスポートした後に別のストリームで使用している場合は、地理空間リスト データが自動的に文字列の形式に変換されています。この場合は、追加のステップを実行して、リスト ストレージ タイプおよび関連する地理空間メタデータを復元する必要があります。

リストについて詳しくは、 12 ページの『リストス トレージおよび関連する尺度』を参照してください。

地理空間メタデータとして設定できる詳細情報については、 148 ページの『地理空間のサブ尺度』を参照してください。

地理空間メタデータをセットアップするには、以下のステップを行います。

1. 可変長ファイル ノードの「ファイル」タブで、「大括弧をリストとして扱う」チェック ボックスを選択します。このチェック ボックスを選択すると、左大括弧と右大括弧で囲まれたデータにコンマや二重引用符などの区切り文字が含まれていても、データが単一の値として扱われます。このチェック ボックスを選択しなかった場合、データが文字列のストレージ タイプとして読み込まれ、フィールドに含まれるコンマがすべて区切り文字として処理され、データ構造が正しく解釈されないことになります。
2. データに単一引用符や二重引用符が含まれる場合は、「単一引用符」フィールドと「二重引用符」フィールドで適宜「ペアで破棄」オプションを選択します。

3. 可変長ファイル ノードの「データ」タブで、地理空間データ フィールドの「上書き」チェック ボックスを選択し、「ストレージ」タイプを文字列からリストに変更します。
4. デフォルトでは、リストの「ストレージ」タイプは「実数のリスト」に設定され、リスト フィールドの下位の値のストレージ タイプは「実数」に設定されます。下位の値のストレージ タイプまたは深さを変更するには、「指定...」をクリックし、「ストレージ」サブダイアログ ボックスを表示します。
5. 「ストレージ」サブダイアログ ボックスでは以下の設定を変更できます。
   - ストレージ データ フィールド全体のストレージ タイプを指定します。デフォルトでは、ストレージ タイプが「リスト」に設定されます。ただし、ドロップダウン リストには、他のすべてのストレージ タイプ (文字列、整数、実数、日付、時間、タイムスタンプ) が含まれています。リスト以外のストレージ タイプを選択する場合は、「ストレージの値の設定」オプションおよび「ツリーの深さ」オプションを使用できません。
   - ストレージの値の設定 フィールド全体ではなく、リストの要素のストレージ タイプを指定します。地理空間フィールドをインポートするときに関連するストレージ タイプは実数と整数のみです。デフォルト設定は実数です。
   - ツリーの深さ リストフィールドの深さを指定します。必要な深さは地理空間フィールドのタイプによって異なり、以下に示す基準に従います。
     - ポイント – 0
     - 行ストリング – 1
     - 多角形 – 1
     - 複数点 – 1
     - 複数行ストリング – 2
     - 多角形群 – 2

     注: リストに変換し直す地理空間フィールドのタイプと、その種類のフィールドに必要な深さを把握しておく必要があります。この情報の設定が誤っていると、フィールドを使用できません。
6. 可変長ファイル ノードの「タイプ」タブで、地理空間データ フィールドの「尺度」セルに正しい尺度が入っていることを確認します。尺度を変更するには、「尺度」セルで「指定...」をクリックし、「値」ダイアログボックスを表示します。
7. 「値」ダイアログボックスに、リストの「尺度」、「ストレージ」、および「ツリーの深さ」が表示されます。「値とラベルを指定」オプションを選択し、「タイプ」ドロップダウン リストから「尺度」に適したタイプを選択します。「タイプ」によっては、データの表現が 2 次元と 3 次元のいずれであるかや使用されている座標系などの詳細情報を求めるプロンプトが出される場合があります。

固定長ファイル・ノード

固定長ファイル・ノードを使用して、固定長フィールド・テキスト・ファイル (各フィールドは区切られていないが、同じ位置から始まる固定長であるファイル) からデータをインポートすることができます。コンピューターによって生成されたデータや旧来のシステムのデータなどは、固定長フィールドの形式で保存されていることがよくあります。固定長ファイル・ノードの「ファイル」タブを使用すると、データ中の列の位置や長さを簡単に指定することができます。

固定長ファイル・ノードのオプションの設定

固定長ファイル・ノードの「ファイル」タブを使用して、列の位置やレコードの長さを指定し、データを IBM SPSS Modeler に取り込むことができます。ダイアログ・ボックスの中央にあるプレビュー領域では、クリックしてフィールド間の区切りを指定する矢印を追加できます。

ファイル**:** ファイルの名前を指定します。ファイル名を入力するか、省略符号ボタン (「**...**」) をクリックしてファイルを選択できます。また、ファイルの内容が、下のパネルに区切り文字付きで表示されます。

データ・プレビュー領域を使用して、列の位置と長さを指定することができます。プレビュー・ウィンドウの上部にあるルーラーは、変数の長さを測定し、フィールドの区切りを指定するために役立ちます。区切り点を指定するには、フィールドの上部にあるルーラーの部分をクリックします。区切り点を移動するには、それをドラッグします。区切り点を削除するには、データ・プレビュー領域外にそれをドラッグ・アンド・ドロップします。

- 各区切り点により、下のテーブルに自動的に新規フィールドが追加されます。
- 矢印で示されている開始点は、下のテーブルの「開始位置」列に自動的に追加されます。

改行文字のスキップ**:** 各レコードの末尾の改行文字をスキップする場合に選択します。

ヘッダー行をスキップ**:** 最初のレコードの先頭で無視する行数を指定します。列見出しを無視する場合などに役立ちます。

レコード長**:** 各レコードの文字数を指定します。

フィールド**:** このデータ・ファイルに対して定義されているすべてのフィールドをリスト表示します。フィールドを定義するには、次の 2 つの方法があります。

- 上のデータ・プレビュー領域を使用して、対話的にフィールドを指定する。
- 下のテーブルに空のフィールド行を追加して、フィールドを手作業で指定する。フィールドの右側にあるボタンをクリックして、新規フィールドを追加します。次に、空のフィールドに、フィールド名、開始位置、および長さを入力します。これらのオプションを指定すると、データ・プレビュー領域に矢印が自動的に追加され、それを使用して簡単に調整することができます。

以前に定義したフィールドを削除するには、削除するフィールドを一覧から選択し、赤い削除ボタンをクリックします。

開始**:** フィールド内の最初の文字の位置を指定します。例えば、レコードの 2 番目のフィールドが 16 文字目から始まる場合は、16 を開始位置として入力します。

長さ**:** 各フィールドの最も長い値の文字数を指定します。これによって、次のフィールドとの分割点が決まります。

前後のスペースを除去**:** インポート時に文字列の前後のスペースを破棄する場合に選択します。

注: SQL プッシュバックを使用する文字列と使用しない文字列と比較すると、接尾空白を含むさまざまな結果を生成する場合があります。

不正な文字**:** データ入力から不正な文字を削除する場合に、「破棄」 を選択します。不正な文字を指定した記号 (1 文字だけ) で置換する場合は、「置換値」 を選択します。ヌル (0) 文字または現在のエンコード中に存在しない任意の文字が不正な文字になります。

エンコード**:** 使用するテキストのエンコード方法を指定します。サーバー・デフォルト、システム・デフォルト、UTF-8 から選択できます。

- システム・デフォルトは、Windows のコントロール・パネル (分散モードで実行している場合はサーバー・コンピューター) で指定できます。
- デフォルトは、「ストリーム・プロパティー」ダイアログ・ボックスで指定されます。

小数点記号**:** データ・ソースで使用する小数点記号の種類を選択します。「ストリームのデフォルト」は、「ストリーム・プロパティー」ダイアログ・ボックスの「オプション」タブで選択されている文字です。これを使用しない場合は、「ピリオド **(.)**」 または 「カンマ **(,)**」 を選択すると、その文字を小数点区切り文字として、このダイアログ・ボックス中のすべてのデータを読み込みます。

自動的に日時を認識**:** IBM SPSS Modeler がデータ項目を自動的に日付または時刻として認識できるようにするには、このチェック・ボックスを選択します。例えば、07-11-1965 などのエントリーを日付として識別し、02:35:58 を時刻として認識します。ただし、07111965 や 023558 のようなあいまいなエントリーは、数値の間に区切り文字がないため、整数として表示されます。

注 : 以前のバージョンの IBM SPSS Modeler のデータ・ファイルを使用する場合に考えられるデータ上の問題を回避するために、13 より前のバージョンで保存された情報についてはデフォルトでこのボックスがオフになります。

スキャン行**:** 指定したデータ型をスキャンする行数を指定します。

このダイアログ・ボックスで作業中は、任意の時点で「リフレッシュ」をクリックすると、フィールドがデータ・ソースから再ロードされます。これは、ソース・ノードへのデータ接続を変更したり、ダイアログ・ボックス中のタブ間を行き来して作業を行うような場合に役立ちます。

Statistics ファイル・ノード

Statistics ファイル・ノードを使用すると、保存された IBM SPSS Statistics ファイル (*.sav* または *.zsav*) からデータを直接読み込むことができます。この形式は、旧バージョンの IBM SPSS Modeler からキャッシュ・ファイルを置換するために使用されます。保存されているキャッシュ・ファイルをインポートする場合は、IBM SPSS Statistics ファイル・ノードを使用します。

インポート・ファイル。ファイルの名前を指定します。ファイル名を入力するか、省略符号ボタン (「**...**」) をクリックしてファイルを選択できます。ファイルを選択すると、ファイル・パスが表示されます。

ファイルはパスワードで暗号化されています。ファイルがパスワード保護されていることがわかっている場合に、このボックスを選択します。「パスワード」に入力を要求するプロンプトが出されます。ファイルがパスワード保護されている場合は、パスワードを入力しないと、別のタブへの移動時、データのリフレッシュ時、ノード内容のプレビュー時、またはノードを含むストリームの実行時に警告メッセージが表示されます。

注**:** パスワード保護されたファイルを開くことができるのは、IBM SPSS Modeler バージョン 16 以降のみです。

変数名。IBM SPSS Statistics の *.sav* または *.zsav* ファイルからインポートする場合の変数名とラベルの処理方法を選択します。ここで指定したメタデータは、IBM SPSS Modeler での作業中は保存され、IBM SPSS Statistics で使用するために再びエクスポートすることができます。

- 名前とラベルを読み込む。変数名とラベルの両方を IBM SPSS Modeler に読み込むために選択します。デフォルトでは、このオプションが選択され、変数名がデータ型ノードに表示されます。ラベルは、「ストリームのプロパティー」ダイアログ・ボックスで指定したオプションに応じて、グラフやモデル・ブラウザー、その他のタイプの出力中に表示できます。デフォルトでは、出力中のラベル表示は無効になっています。
- ラベルを名前として読み取る。短いフィールド名ではなく、IBM SPSS Statistics の *.sav* または *.zsav* ファイルから詳細な変数ラベルを読み取り、そのラベルを IBM SPSS Modeler で変数名として使用する場合に選択します。

値。IBM SPSS Statistics の *.sav* または *.zsav* ファイルからインポートする場合の値とラベルの処理方法を選択します。ここで指定したメタデータは、IBM SPSS Modeler での作業中は保存され、IBM SPSS Statistics で使用するために再びエクスポートすることができます。

- データとラベルを読み込む。実際の値と値ラベルの両方を IBM SPSS Modeler に読み込むために選択します。デフォルトでは、このオプションが選択され、値自体がデータ型ノードに表示されます。値ラベルは、「ストリームのプロパティー」ダイアログ・ボックスで指定したオプションに応じて、式ビルダー、グラフ、モデル・ブラウザー、その他の種類の出力中に表示できます。
- ラベルをデータとして読み込み。値を表すために使用される数値コードまたはシンボリック・コードではなく、*.sav* ファイルまたは *.zsav* ファイルの値ラベルを使用する場合に選択します。例えば、「1」と「2」の値が実際には「男性」と「女性」を表す性別フィールドを持つデータについてこのオプションを選択すると、性別フィールドが文字列に変換され、「男性」と「女性」が実際の値としてインポートされます。

  このオプションを選択する前に、IBM SPSS Statistics データ中の欠損値を検討しておくことが大切です。例えば、数値フィールドで欠損値についてのみラベルを使用している場合 (0=回答なし、99=不明)、このオプションを選択すると、「回答なし」と「不明」という値ラベルだけがインポートされ、数値フィールドが文字列に変換されます。このような場合は、値自体をインポートして、データ型ノードに欠損値を設定する必要があります。

フィールド形式情報を使用して、ストレージを指定します。このボックスを選択解除すると、 *.sav* ファイル内で整数形式のフィールド値 (例えば、IBM SPSS Statistics の変数ビューで F*n*.0 として指定されているフィールド) は、整数ストレージを使用してインポートされます。文字列を除くすべてのフィールド値は、実数としてインポートされます。

このボックスを選択すると (デフォルト)、文字列以外のフィールド値はすべて、 *.sav* ファイル内で整数形式であるかないかに関係なく、実数としてインポートされます。

複数回答設定。ストリームに定義された複数の回答セットは、IBM SPSS Statistics ファイルがエクスポートされると自動的に保存されます。「フィルター」タブで、ノードの複数の回答セットを表示および編集することができます。 詳しくは、トピック 161 ページの『複数回答セット編集』を参照してください。

Data Collectionノード

Data Collection ソース・ノードは、Data Collection 製品に付属の Survey Reporter Developer Kit に基づいた調査データをインポートします。この形式は、調査中に収集された質問に対する実際の回答であるケース データを、ケース データ が収集され整理されたメタデータ と区別します。メタデータは、質問テキスト、変数名とその説明、複数の回答の変数定義、テキスト文字列の翻訳、ケース・データの構造の定義などの情報から構成されます。

注**:** このノードには、Survey Reporter Developer Kit (Data Collection 製品に付属) が必要です。Developer Kit のインストールとは別に、追加の設定を行う必要はありません。

コメント

- 調査データは、フラットな表形式の VDATA 形式から、またはメタデータが含まれている場合、階層式の HDATA 形式のソースから読み込まれます。
- データ型は、メタデータの情報を使用して、自動的にインスタンス化されます。
- 調査データが SPSS Modeler へインポートされると、各回答者を 1 レコードにして、質問がフィールドとして提供されます。

Data Collection インポート・ファイルのオプション

Data Collection ノードの「ファイル」タブを使用して、インポートするメタデータとケース・データのオプションを指定できます。

メタデータの設定

注**:** 使用できるプロバイダのファイルの種類についての完全なリストを表示するには、Survey Reporter Developer Kit (Data Collection ソフトウェア製品に付属) がインストールされている必要があります。

メタデータ プロバイダ。調査データは、Data Collection Survey Reporter Developer Kit でサポートされている多くの形式からインポートできます。利用できるプロバイダの種類には、次のものがあります。

- **DataCollectionMDD**。質問定義ファイル (*.mdd*) からメタデータを読み込みます。これは、標準の Data Collection Data Model 形式です。
- **ADO** データベース。ADO ファイルからケース・データとメタデータを読み取ります。メタデータが含まれた *.adoinfo* の名前と場所を指定します。この DSC の内部名は *mrADODsc* です。
- **In2data** データベース。In2data のケース・データとメタデータを読み取ります。この DSC の内部名は *mrI2dDsc* です。
- **Data Collection** ログ・ファイル。標準の Data Collection ログ・ファイルからメタデータを読み込みます。通常、ログ・ファイルにはファイル名拡張子の *.tmp* が付いています。ただし、別のファイル名拡張子が付いているログ・ファイルがある場合もあります。必要な場合にはこのファイルの名前を変更して、ファイル名拡張子の *.tmp* を付けることもできます。この DSC の内部名は *mrLogDsc* です。
- **Quancept** 定義ファイル。メタデータを Quancept スクリプトに変換します。Quancept *.qdi* ファイルの名前を指定します。この DSC の内部名は *mrQdiDrsDsc* です。
- **Quanvert Database**。 Quanvert のケース・データとメタデータを読み込みます。*.qvinfo* または *.pkd* ファイルの名前と場所を指定します。この DSC の内部名は *mrQvDsc* です。
- **Data Collection Participation** データベース。 プロジェクトの Sample テーブルと History Table テーブルを読み込み、これらのテーブル内の列に対応する派生カテゴリー変数を作成します。この DSC の内部名は *mrSampleReportingMDSC* です。
- **Statistics** ファイル。IBM SPSS Statistics の *.sav* ファイルからケース・データとメタデータを読み込みます。IBM SPSS Statistics 内での分析用に、IBM SPSS Statistics *.sav* ファイルへケース・データを書き出します。ta from an IBM SPSS Statistics *.sav* ファイルからのメタデータを *.mdd* ファイルへ書き出します。この DSC の内部名は *mrSavDsc* です。
- **Surveycraft** ファイル。SurveyCraft のケース・データとメタデータを読み込みます。SurveyCraft *.vq* ファイルの名前を指定します。この DSC の内部名は *mrSCDsc* です。
- **Data Collection** スクリプト ファイル。mrScriptMetadata ファイル内のメタデータを読み取ります。通常、これらのファイルには、ファイル名拡張子の *.mdd* または *.dms* が付いています。この DSC の内部名は *mrScriptMDSC* です。
- **Triple-S XML** ファイル。XML 形式の Triple-S ファイルのメタデータを読み込みます。この DSC の内部名は *mrTripleSDsc* です。

メタデータのプロパティー。オプションで、「プロパティー」 を選択して、インポートする調査のバージョンや、使用する言語、コンテキスト、ラベルの種類を指定します。 詳しくは、トピック 38 ページの『Data Collection インポート・メタデータのプロパティー』を参照してください。

ケース・データの設定

注**:** 使用できるプロバイダのファイルの種類についての完全なリストを表示するには、Survey Reporter Developer Kit (Data Collection ソフトウェア製品に付属) がインストールされている必要があります。

ケース データ設定の取得。*.mdd* ファイルだけからメタデータを読み取る場合は、「ケース データ設定の取得」をクリックして、特定のソースにアクセスするために必要な具体的な設定とともに、選択したメタデータに関連付けるケース・データ・ソースを決定します。このオプションは、*.mdd* ファイルに対してだけ利用できます。

ケース データ プロバイダー。次のプロバイダの種類がサポートされます。

- **ADO** データベース。Microsoft ADO インターフェースを使用して、ケース・データを読み取ります。ケース・データには「OLE-DB UDL」を選択して、「ケース・データの URL」フィールドに接続文字列を指定します。詳しくは、トピック 39 ページの『データベース接続文字列』を参照してください。このコンポーネントの内部名は *mrADODsc* です。
- 区切りテキスト・ファイル **(Excel)**。 Excel で出力できるような、カンマ区切り形式 (.CSV) ファイルからケース・データを読み込みます。内部名は *mrCsvDsc* です。
- **Data Collection** データ ファイル。 ネイティブの Data Collection データ形式ファイルからケース・データを読み取ります。内部名は *mrDataFileDsc* です。
- **In2data** データベース。 In2data データベース・ファイル (*.i2d*) からケース・データとメタデータを読み込みます。内部名は *mrI2dDsc* です。
- **Data Collection** ログ・ファイル。 標準の Data Collection ログ・ファイルからケース・データを読み込みます。通常、ログ・ファイルにはファイル名拡張子の *.tmp* が付いています。ただし、別のファイル名拡張子が付いているログ・ファイルがある場合もあります。必要な場合にはこのファイルの名前を変更して、ファイル名拡張子の *.tmp* を付けることもできます。内部名は *mrLogDsc* です。
- **Quantum** データ・ファイル。Quantum 形式の ASCII ファイル (*.dat*)からケース・データを読み込みます。内部名は *mrPunchDsc* です。
- **Quancept** データ・ファイル。Quancept *.drs*、*.drz*、または *.dru* ファイルからケース・データを読み込みます。内部名は *mrQdiDrsDsc* です。
- **Quanvert** データベース。Quanvert の *qvinfo* または *.pkd* ファイルからケース・データを読み込みます。内部名は *mrQvDsc* です。
- **Data Collection** データベース **(MS SQL Server)**。ケース・データをリレーショナルな Microsoft SQL Server データベースへ読み込みます。詳しくは、トピック 39 ページの『データベース接続文字列』を参照してください。内部名は *mrRdbDsc2* です。
- **Statistics** ファイル。IBM SPSS Statistics の *.sav* ファイルからケース・データを読み込みます。内部名は *mrSavDsc* です。
- **Surveycraft** ファイル。SurveyCraft の *.qdt* ファイルからケース・データを読み込みます。*.vq* ファイルと *.qdt* ファイルの両方とも、両ファイルへの読み取りと書き込みアクセス権限付きで、同じディレクトリー内に存在している必要があります。これは、SurveyCraft の使用時にデフォルトで 2 つのファイルが作成される方法ではありません。したがって、ファイルの 1 つは SurveyCraft データをインポートするために移動する必要があります。内部名は *mrScDsc* です。
- **Triple-S Data** ファイル。固定長またはカンマ区切り形式の Triple-S データ・ファイルから、ケース・データを読み込みます。内部名は *mr TripleDsc* です。
- **Data Collection XML**。Data Collection の XML データ・ファイルからケース・データを読み込みます。通常、この形式はある場所から別の場所へケース・データを転送するのに使用できます。内部名は *mrXmlDsc* です。

ケース・データ・タイプ。ケース・データがファイル、フォルダー、OLE-DB UDL、または ODBC DSN から読み込まれたかどうかを指定し、それに従ってダイアログ・ボックスのオプションが更新されます。有効なオプションは、プロバイダの種類によって異なります。データベース・プロバイダの場合は、OLE-DB または ODBC 接続にオプションを指定できます。 詳しくは、トピック 39 ページの『データベース接続文字列』を参照してください。

ケース・データ・プロジェクト。Data Collection データベースからケース・データを読み込むときに、プロジェクトの名前を入力できます。その他のケース・データのデータ型については、この設定を空白のままにしておく必要があります。

変数インポート

システム変数インポート。面談の状態 (進行中、終了、終了日など) を示す変数など、システム変数をインポートするかどうかを指定します。「なし」、「すべて」、または「標準」を選択することができます。

インポート **"Codes"** 変数。カテゴリー変数の、自由回答形式の「その他」の回答に使用されるコードを示す変数のインポートを制御します。

インポート**"SourceFile"** 変数。スキャンされた回答の画像のファイル名を含む変数のインポートを制御します。

複数回答変数をインポート。複数回答変数を複数のフラグ型フィールド (複合二分セット) としてインポートできます。これは新しいストリームのデフォルトの方法です。IBM SPSS Modeler 12.0 以前のリリースで作成されたストリームでは、値をカンマで区切った複数の回答を単一のフィールドにインポートしていました。古い方法は現在もサポートされており、以前と同じように既存のストリームを実行できますが、古いストリームを更新して新しい方法を使用することをお勧めします。 詳しくは、トピック 39 ページの『複数プロパティー設定のインポート』を参照してください。

Data Collection インポート・メタデータのプロパティー

Data Collection の調査データのインポート時に、「メタデータ・プロパティー」ダイアログ・ボックスで、インポートする調査のバージョンのほかに、使用する言語、コンテキスト、ラベル・タイプを指定することができます。一度に 1 つの言語、コンテキスト、およびラベルの種類のみをインポートできます。

「バージョン」。調査の各バージョンは、ケース・データの特定のセットを収集するために使用されるメタデータのスナップショットと見なすことができます。質問には変更が加えられるので、複数のバージョンが作成されることがあります。最新バージョン、すべてのバージョン、または特定のバージョンをインポートできます。

- すべてのバージョン。利用可能なすべてのバージョンの組み合わせ (スーパーセット) を使用する場合に、このオプションを選択します。(これは、スーパーバージョンと呼ばれることもあります。)バージョン間に矛盾がある場合は、最新バージョンが優先されます。例えば、カテゴリーのラベルがバージョン間で異なる場合、最新バージョン内のテキストが使用されます。
- 最新のバージョン。最新のバージョンを使用する場合に、このオプションを使用します。
- バージョンの指定。特定の調査バージョンを使用する場合に、このオプションを使用します。

すべてのバージョンを選択することが、役立つ場合があります。例えば、複数のバージョンのケース・データをエクスポートする予定で、あるバージョンで収集されたケース・データが別のバージョンで有効でないことになる、変数とカテゴリーの定義に変更があるときなどです。ケース・データをエクスポートするバージョンすべてを選択することは、バージョンの違いが原因の有効性のエラーが発生することなく、異なるバ

ージョンで収集されたケース・データを同時にエクスポートできる、ということです。ただし、バージョンの変更に応じて、何らかの有効性のエラーが引き続き発生する可能性はあります。

言語。質問と関連テキストは、メタデータ内に複数の言語で格納できます。調査にデフォルトの言語を使用することも、特定の言語を指定することもできます。ある項目が指定された言語で利用できない場合、デフォルトが使用されます。

コンテキスト。使用するユーザー・コンテキストを選択します。ユーザー・コンテキストで、表示されるテキストが制御されます。例えば、質問のテキストを表示するには「質問」を選択し、データを分析するときの表示に適した短いテキストを表示するには「分析」を選択します。

ラベル型。定義されたラベルの種類を一覧表示します。デフォルトは、「質問] ユーザー・コンテキストで質問のテキストに使用され、「分析] ユーザー・コンテキストで変数の説明に使用される「ラベル」です。その他のラベルの種類は、指示、説明などのために定義されます。

データベース接続文字列

OLE-DB または ODBC 経由でデータベースからケース・データをインポートするために Data Collection ノードを使用する場合は、「ファイル」タブから「編集」を選択して「接続文字列」ダイアログ・ボックスを利用します。このダイアログ・ボックスで、接続を微調整するために、プロバイダに渡す接続文字列をカスタマイズできます。

アドバンス プロパティー

明示してログインすることが必要なデータベースからケース・データをインポートするために Data Collection ノードを使用する場合は、「詳細設定」を選択して、データ・ソースにアクセスするためのユーザー ID とパスワードを提示します。

複数プロパティー設定のインポート

複数回答変数は、変数の各値に個別のフラグ型フィールドを持つ複合二分セットとして、Data Collection からインポートできます。例えば、回答者が訪れたことのある博物館をリストから選択するよう質問された場合、セットには表示されたそれぞれの博物館に個別のフラグ型フィールドが含まれます。

データをインポートした後、「フィルター」タブを含むノードの複数回答のセットを追加または編集することができます。 詳しくは、トピック 161 ページの『複数回答セット編集』を参照してください。

単一フィールドへの複数回答のインポート (以前のリリースで作成されたストリームの場合)

SPSS Modeler の以前のリリースでは、前述のように複数回答をインポートするのではなく、カンマで値を区切って単一フィールドにインポートしていました。この方法は現在もサポートされており、既存のストリームがサポートされていますが、既存のストリームを更新して新しい方法を使用することをお勧めします。

Data Collection 列インポート・ノード

Data Collection データの列が、次の表に要約したように、SPSS Modeler へ読み込まれます。

表 *4. Data Collection* 列インポートの要約

Data Collection 列のタイプ	**SPSS Modeler** ストレージ	測定水準
ブール型フラグ (yes/no)	文字列	フラグ型 (0 と 1 の値)
カテゴリー	文字列	名義

表 4. *Data Collection* 列インポートの要約 (続き)

Data Collection 列のタイプ	**SPSS Modeler** ストレージ	測定水準
日付またはタイム・スタンプ	タイムスタンプ	連続
倍精度 (指定された範囲内の浮動小数点数値)	実数	連続
長形 (指定された範囲内の整数値)	整数	連続
テキスト (自由形式のテキスト記述)	文字列	不明
レベル (質問内のグリッドまたはループを示す)	VDATA 内にはなく、SPSS Modeler へインポートされない	
オブジェクト (走り書きのテキストを示すファクシミリや音声の録音などのバイナリ データ)	SPSS Modeler へインポートされません	
なし (不明なデータ型)	SPSS Modeler へインポートされません	
Respondent.Serial 列 (一意の ID を各回答者に関連付ける)	整数	不明

メタデータから読み込んだ値のラベルと実際の値の間に潜在する矛盾を避けるために、すべてのメタデータ値が小文字に変換されます。例えば、値のラベル *E1720_years* は *e1720_years* へ変換されます。

IBM Cognos ソース・ノード

IBM Cognos ソース・ノードを使用すると、Cognos データベース・データまたは 1 つの表のレポートをデータ・マイニング セッションに取り込むことができます。このように、Cognos のビジネス・インテリジェンス機能を IBM SPSS Modeler の予測分析機能とを統合できます。関連するディメンションを使用してモデル化されたリレーショナル (DMR) データおよび OLAP データをインポートできます。

Cognos サーバー接続から、まずデータまたはレポートをインポートする場所を選択します。場所には Cognos モデルと、モデルに関連するすべてのフォルダー、クエリー、レポート、ビュー、ショートカット、URL、ジョブ定義が含まれます。Cognos モデルはビジネス・ルール、データの説明、データの関係、ビジネスの次元のおよび階層、その他の管理タスクを定義します。

データをインポートしている場合、選択したパッケージからインポートするオブジェクトを選択します。インポートできるオブジェクトには、クエリーの件名 (データベース・テーブルを示す) または各クエリーの項目 (テーブルの列を示す) があります。 詳しくは、 41 ページの『Cognos オブジェクトのアイコン』を参照してください。

パッケージのフィルターが定義されている場合、フィルターをインポートできます。インポートするフィルターがインポートされるデータを関連している場合、データがインポートされる前にフィルターが適用されます。インポートするデータは、UTF-8 形式になっている必要があります。

レポートをインポートする場合、1 つまたは複数のレポートを含む、パッケージ、またはパッケージ内のフォルダーを選択します。インポートするここのレポートを選択します。1 つのリストのレポートだけをインポートできます。複数のリストはサポートされていません。

パラメーターが定義されている場合、データ・オブジェクトまたはレポートに対しては、オブジェクトまたはレポートをインポートする前にこれらのパラメーターの値を指定することができます。

注: Cognos ソース・ノードでは、Cognos CQM パッケージのみがサポートされます。DQM パッケージは、サポートされません。

Cognos オブジェクトのアイコン

次の表に示されているように、Cognos Analytics データベースからインポートできるさまざまなタイプのオブジェクトは、異なるアイコンによって表されます。

表 5. *Cognos* オブジェクトのアイコン:

アイコン	オブジェクト
	パッケージ
	名前空間
	クエリーの件名
	クエリーの項目
	数値データ・ディメンション
	指標
	次元
	レベルの階層
	レベル
	フィルター
	レポート
	スタンドアロンの計算

Cognos データのインポート

IBM Cognos Analytics (バージョン 11 以降がサポート対象) のデータベースからデータをインポートするには、IBM Cognos のダイアログ・ボックスの「データ」タブで「モード」 が「データ」に設定されていることを確認します。

接続: 「編集」をクリックするとダイアログ ボックスが表示され、データまたはレポートのインポート元となる新規 Cognos 接続の詳細情報を定義することができます。IBM SPSS Modeler 経由ですでに Cognos サーバーにログインしている場合、現在の接続の詳細を編集することもできます。詳しくは、トピック 43 ページの『Cognos の接続』を参照してください。

位置: Cognos サーバー接続を確立したら、フィールドの隣の「編集」をクリックし、コンテンツをインポートするパッケージのリストを表示します。詳しくは、 44 ページの『Cognos の場所の選択』を参照してください。

内容。選択したパッケージの名前が、パッケージに関連する名前空間とともに表示されます。名前空間をダブルクリックすると、インポートできるオブジェクトが表示されます。さまざまなオブジェクト・タイプが、異なるアイコンで示されます。 詳しくは、 41 ページの『Cognos オブジェクトのアイコン』を参照してください。

インポートするオブジェクトを選択するには、オブジェクトを選択して、2 つの右側の矢印の上部をクリックし、「インポートするフィールド」 ペインにオブジェクトを移動します。クエリーの件名を選択すると、そのクエリー項目がすべてインポートされます。クエリーの件名をダブルクリックすると展開され、各クエリー項目を 1 つまたは複数選択できます。Ctrl キー (各項目を選択)、Shift キー (項目のブロックを選択) および Ctrl + A キー (すべての項目を選択) で複数の項目を選択できます。

(パッケージにフィルターが定義されている場合)適用するフィルターを選択するには、「内容」ペインのフィルターに移動し、右側の 2 つの矢印のうち下の矢印をクリックして 「適用するフィルター」 ペインにフィルターを移動します。Ctrl キー (各フィルターを選択)、Shift キー (フィルターをまとめて選択) で複数の項目を選択できます。

インポートするフィールド。IBM SPSS Modeler にインポートして処理することを選択したデータベース・オブジェクトがリストされます。特定のオブジェクトを必要としなくなった場合、そのフィールドをクリックして左方向矢印をクリックすると、「内容」 ペインに移動します。「内容」 と同じ方法で複数の項目を選択できます。

適用するフィルター。インポートされる前にデータに適用することを選択したフィルターが一覧表示されます。特定のフィルターを必要としなくなった場合、そのフィールドをクリックして左方向矢印をクリックすると、「内容」 ペインに移動します。「内容」 と同じ方法で複数の項目を選択できます。

パラメーター: このボタンが有効な場合は、選択したオブジェクトにパラメーターが定義されています。パラメーターを使用して、データをインポートする前に調整を行うことができます (例えば、パラメーター化された計算を実行できます)。パラメーターが定義されていてもデフォルトが指定されていない場合、ボタンは警告の三角形を表示します。ボタンをクリックして、パラメーターを表示し、必要に応じて編集します。ボタンが無効になっている場合、レポートにはパラメーターが定義されません。

インポートする前にデータを集計する。未処理のデータではなく集計済みのデータをインポートする場合、このボックスをオンにします。

Cognos レポートのインポート

IBM Cognos のデータベースからレポートをインポートするには、IBM Cognos のダイアログ・ボックスの「データ」タブで「モード」が「レポート」に設定されていることを確認します。1 つのリストのレポートだけをインポートできます。複数のリストはサポートされていません。

接続**:** 「編集」をクリックするとダイアログ ボックスが表示され、データまたはレポートのインポート元となる新規 Cognos 接続の詳細情報を定義することができます。IBM SPSS Modeler 経由ですでに Cognos サーバーにログインしている場合、現在の接続の詳細を編集することもできます。詳しくは、トピック『Cognos の接続』を参照してください。

位置**:** Cognos サーバー接続を確立したら、フィールドの隣の「編集」をクリックし、コンテンツをインポートするパッケージのリストを表示します。詳しくは、 44 ページの『Cognos の場所の選択』を参照してください。

内容。レポートを含む、選択したパッケージやフォルダーの名前を表示します。特定のレポートに移動して選択し、右側の矢印をクリックして 「インポートするレポート」 フィールドにレポートを移動します。

インポートするレポート。IBM SPSS Modeler へのインポートに選択されたレポートを示します。レポートを必要としなくなった場合、そのフィールドをクリックして左方向矢印をクリックすると、「内容」 ペインに移動し、または異なるレポートをこのフィールドに移動します。

パラメーター**:** このボタンが有効な場合、選択したレポートにパラメーターが定義されます。パラメーターを使用して、レポートをインポートする前に調整を行うことができます (例えば、レポート・データの開始日および終了日を指定するなど)。パラメーターが定義されていてもデフォルトが指定されていない場合、ボタンは警告の三角形を表示します。ボタンをクリックして、パラメーターを表示し、必要に応じて編集します。ボタンが無効になっている場合、レポートにはパラメーターが定義されません。

Cognos の接続

「Cognos 接続」ダイアログ・ボックスで、データベース・オブジェクトのインポート元またはエクスポート先の Cognos Analytics サーバー (バージョン 11 以降がサポート対象) を選択できます。

Cognos サーバー **URL (Cognos server URL)**: インポートまたはエクスポートの対象となる Cognos Analytics サーバーの URL を入力します。これは、Cognos サーバーの IBM Cognos Configuration の「外部ディスパッチャー URI」環境プロパティーの値です。使用する URL が分からない場合、Cognos システム管理者にお問い合わせください。

モード: 特定の Cognos 名前空間、ユーザー名、およびパスワードを使用して (例えば、管理者として) ログインする場合、「資格情報を設定」を選択します。ユーザー資格情報を使用せずにログインする場合は「匿名接続を使用」 を選択します。この場合、他のフィールドには入力しません。

または、IBM SPSS Collaboration and Deployment Services リポジトリに格納された IBM Cognos 資格情報がある場合は、ユーザーの名前およびパスワードの情報の入力や匿名接続の作成を行う代わりに、この資格情報を使用できます。既存の資格情報を使用するには、「保管されている資格情報」を選択し、「資格情報名」にその名前を入力するか、それを探します。

Cognos 名前空間は、IBM SPSS Collaboration and Deployment Services 内のドメインによりモデル化されます。

「名前空間 **ID**」: サーバーへのログオンに使用される Cognos セキュリティー認証プロバイダーを指定します。認証プロバイダーを使用して、ユーザー、グループ、役割を定義および維持し、認証プロセスを制御します。これは名前空間 ID であり、名前空間の名前ではないことに注意してください (ID は必ずしも名前と同じではありません)。

ユーザー名: サーバーへのログオンに使用する Cognos ユーザー名を入力します。

パスワード: 指定したユーザー名に関連付けられているパスワードを入力します。

デフォルトとして保存 :このボタンをクリックすると、これらの設定がデフォルトとして保存され、ノードを開くたびに再度入力する必要がなくなります。

Cognos の場所の選択

「場所の指定」ダイアログ・ボックスを使用すると、データをインポートする Cognos パッケージ、またはレポートをインポートするパッケージを選択できます。

公開フォルダー**:** データをインポートする場合、選択したサーバから使用可能なパッケージとフォルダーが一覧表示されます。使用するパッケージを選択し、「**OK**」 をクリックします。Cognos ソース・ノードごとに選択できるパッケージは 1 つだけです。

レポートをインポートする場合、これは、選択したサーバーから使用可能なレポートを含むフォルダーやパッケージが一覧表示されます。パッケージまたはレポート・フォルダーを選択し、「**OK**」 をクリックします。レポート・フォルダーには各レポートのほか他のレポート・フォルダーも含まれますが、Cognos ソース・ノードごとにパッケージまたはレポート・フォルダーを 1 つだけ選択できます。

データまたはレポートのパラメーターの指定

Cognos Analytics にパラメーターが定義されている場合、データ・オブジェクトまたはレポートに対しては、オブジェクトまたはレポートをインポートする前にこれらのパラメーターの値を指定できます。レポートのパラメーターの例は、レポートの内容の開始日および終了日です。

名前。Cognos データベースで指定されるパラメーター名。

タイプ: パラメーターの説明。

値**:** パラメーターに割り当てる値。値を入力または編集するには、テーブルのセルをダブルクリックします。値はここでは検証されていないため、無効な値が実行時に検出されます。

無効なパラメーターを表から自動的に削除。このオプションはデフォルトで選択され、データ・オブジェクトまたはレポート内で見つかった無効なパラメーターが削除されます。

IBM Cognos TM1 ソース・ノード

IBM Cognos TM1 ソース・ノードを使用すると、Cognos TM1 データをデータ・マイニング・セッションに取り込むことができます。これにより、Cognos のエンタープライズ計画機能を IBM SPSS Modeler の予測分析機能と統合できます。多次元 OLAP キューブ・データのフラット化バージョンをインポートできます。

注**:** TM1 ユーザーは、次の権限が必要です: キューブの書き込み権限、ディメンションの読み取り権限、ディメンション要素の書き込み権限。また、SPSS Modeler が Cognos TM1 データのインポートとエクス

ポートを行うには、IBM Cognos TM1 10.2 フィックスパック 3 以降が必要です。以前のバージョンに基づいていた既存のストリームは、依然として機能します。

このノードに対して管理者の資格情報は不要です。しかし、17.1 より前の古いレガシー TM1 ノードを引き続き使用している場合、管理者の資格情報は今までどおり必要です。

データをインポートする前に、TM1 でデータを変更する必要があります。

注: インポートするデータは、UTF-8 形式になっている必要があります。

IBM Cognos TM1 管理ホスト接続から、最初にデータのインポート元の TM1 サーバーを選択します。サーバーには 1 つ以上の TM1 キューブが含まれています。次に、必要なキューブを選択し、そのキューブ内でインポートする列と行を選択します。

注: SPSS Modeler で TM1 ソース ノードまたは TM1 エクスポート ノードを使用するには、事前に `tm1s.cfg` ファイル内の一部の設定を確認する必要があります。このファイルは、TM1 サーバーのルート ディレクトリーにある TM1 サーバー構成ファイルです。

- HTTPPortNumber - 有効なポート番号を設定します。通常は、1 から 65535 です。これは、後で接続時にノードで指定するポート番号ではなく、TM1 で使用される内部ポート (デフォルトでは無効) であることに注意してください。必要な場合は、TM1 管理者に問い合わせて、このポートの有効な設定を確認してください。
- UseSSL - これを *True* に設定すると、HTTPS がトランスポート プロトコルとして使用されます。この場合、TM1 認証を SPSS Modeler Server JRE にインポートする必要があります。

IBM Cognos TM1 データのインポート

IBM Cognos TM1 データベースからデータをインポートするには、IBM Cognos TM1 のダイアログ ボックスの「データ」タブで、該当する TM1 管理ホストと、関連するサーバー、キューブ、およびデータの詳細を選択します。

注: データをインポートする前に、TM1 内で前処理を実行して、データを IBM SPSS Modeler が認識できる形式にしておく必要があります。このために、サブセット・エディターを使用してデータをフィルター操作し、ビューをインポートに適したサイズと形状にします。

ゼロ (0) の値を TM1 からインポートする場合は「ヌル」値と見なされる (TM1 では空白とゼロの値を区別しない) ことに注意してください。また、標準ディメンション の非数値データ (メタデータ) を IBM SPSS Modeler にインポートできることにも注意してください。しかし、数値でない測定 のインポートは現在サポートされていません。

管理ホスト: 接続先の TM1 サーバーがインストールされている管理ホストの URL を入力します。管理ホストは、すべての TM1 サーバーに対する単一の URL として定義されます。この URL から、ご使用の環境にインストールされ、稼働しているすべての IBM Cognos TM1 サーバーをディスカバーし、これらのサーバーにアクセスすることができます。

TM1 サーバー: 管理ホストへの接続が確立したら、インポートするデータが含まれているサーバーを選択し、「ログイン」をクリックします。このサーバーに以前に接続していない場合は、「ユーザー名」と「パスワード」の入力を要求するプロンプトが出されます。または、以前に入力し、「保管されている資格情報」として保存したログイン詳細を検索できます。

インポート対象の **TM1** キューブ・ビューの選択。データのインポート元の TM1 サーバー内にあるキューブの名前が表示されます。キューブをダブルクリックして、インポート可能なデータを表示します。

注**:**

IBM SPSS Modeler にインポートできるのは、ディメンションを持つキューブだけです。

TM1 キューブ内の要素に対してエイリアスが定義されている場合 (例えば、23277 の値が Sales というエイリアスを持つ場合) は、エイリアスではなく値がインポートされます。

インポートするデータを選択するには、ビューを選択して右矢印をクリックし、ビューを「インポート対象のビュー」ペインに移動します。必要なビューが表示されない場合は、キューブをダブルクリックしてビューのリストを展開します。共有ビューまたは専用ビューを選択できます。

行ディメンション。インポート用に選択したデータの行ディメンションの名前が表示されます。レベルのリストをスクロールして、必要な項目を選択します。

列ディメンション。インポート用に選択したデータの列ディメンションの名前が表示されます。レベルのリストをスクロールして、必要な項目を選択します。

コンテキスト・ディメンション。表示のみ。選択した列と行に関連するコンテキスト・ディメンションが表示されます。

TWC ソース・ノード

TWC ソース・ノードは、IBM ビジネスの 1 つである The Weather Company から気象データをインポートします。これを使用して、ある場所の過去または予報の気象データを取得できます。これにより、使用可能な最も正確で高精度の気象データを利用して、意思決定を向上させるための気象主導のビジネス・ソリューションの開発に役立てることができます。

このノードを使用して、気象関連データ (latitude、longitude、time、day_ind (夜または昼を示す)、temp、dewpt (露点)、rh (相対湿度)、feels_like 気温、heat_index、wc (風速冷却)、wx_phrase (ほぼ曇り、晴れ時々曇りなど)、pressure、clds (雲)、vis (視界)、wspd (風速)、gust、wdir (風向)、wdir_cardinal (NW、NNW、N など)、uv_index (紫外線指数)、uv_desc (低、高など) など) を入力できます。

TWC ソース・ノードは以下の API を使用します。

- TWC Historical Observations Airport (http://goo.gl/DplOKj)。過去の気象データを取得します
- TWC Hourly Forecast (http://goo.gl/IJhhvZ)。予測された気象データを取得します

場所

緯度: 気象データの取得対象とする場所の緯度の値を、[-90.0~90.0] の形式で入力します。

経度: 気象データの取得対象とする場所の経度の値を、[-180.0~180.0] の形式で入力します。

その他

ライセンス・キー: ライセンス・キーが必要です。The Weather Company から入手したライセンス・キーを入力します。キーがない場合は、管理者または IBM 担当員に連絡してください。

すべてのユーザーにキーを発行する代わりに、管理者はキーを IBM SPSS Modeler Server 上の新しい config.cfg ファイルに指定している場合があります。その場合は、このフィールドを空白のままにできます。両方の場所に指定した場合は、このダイアログのキーが優先されます。管理者への注記: サーバーにライセンス・キーを追加するには、config.cfg という新規ファイルを場所

<ModelerServerInstallation>¥ext¥bin¥pasw.twcdata に「LicenseKey=<LICENCEKEY>」(ここで <LICENSEKEY> はライセンス・キーです) という内容で作成してください。

単位: 使用する測定単位を、「英語」、「メトリック」、「**Hybrid**」の中から選択します。デフォルトは「メトリック」です。

データ型

過去: 過去の気象データをインポートする場合は、「過去」を選択し、YYYYMMDD の形式で開始日と終了日を指定します (例えば、2012 年 1 月 1 日の場合 20120101)。

予測: 予報の気象データをインポートする場合は、「予測」を選択して、予測する時間を指定します。

SAS ソース・ノード

この機能は SPSS Modeler Professional および SPSS Modeler Premium で使用できます。

SAS ソース・ノードを使用すると、SAS データをデータ・マイニング セッションに取り込むことができます。次の 4 種類のファイルをインポートできます。

- SAS for Windows/OS2 (*.sd2*)
- SAS for UNIX (*.ssd*)
- SAS 移送ファイル (*.tpt*)
- SAS バージョン 7/8/9 (*.sas7bdat*)

データのインポート時は、変数はすべて保持され、変数の型は変更されません。また、すべてのケースが選択されます。

SAS ソース・ノードのオプションの設定

インポート**:** 移送する SAS ファイルの種類を選択します。「**SAS for Windows/OS2 (.sd2)**」、「**SAS for UNIX (.SSD)**」、「**SAS** トランスポート・ファイル **(.tpt)**」、または「**SAS** バージョン **7/8/9 (.sas7bdat)**」を選択することができます。

インポート・ファイル**:** ファイルの名前を指定します。ファイル名を入力するか、省略符号ボタン (「**...**」) をクリックしてファイルの場所を参照できます。

メンバー**:** 上で選択した SAS 転送ファイルからインポートするメンバーを選択します。メンバー名を入力するか、または「選択」をクリックしてファイル内のメンバーを指定します。

SAS データ・ファイルからユーザー形式を読み込む**:** ユーザー形式を読み込む場合に選択します。SAS ファイルでは、データとデータ形式 (変数ラベルなど) を別々のファイルに保存します。通常は、形式もインポートします。データ・セットのサイズが大きい場合は、このオプションの選択を解除すると、メモリーを節約することができます。

形式ファイル**:** 形式ファイルが必要な場合は、このテキスト・ボックスが使用可能になります。ファイル名を入力するか、省略符号ボタン (「**...**」) をクリックしてファイルの場所を参照できます。

変数名**:** SAS ファイルからのインポート時に変数名とラベルを処理する方法を選択します。ここで含めることを選択したメタデータは、IBM SPSS Modeler 内での作業を通じて維持され、SAS 内での使用のために再びエクスポートされる可能性があります。

- 名前とラベルを読み込む: 変数名とラベルの両方を IBM SPSS Modeler に読み込むために選択します。デフォルトでは、このオプションが選択され、変数名がデータ型ノードに表示されます。ラベルは、「ストリームのプロパティー」ダイアログ・ボックスで指定したオプションに応じて、式ビルダー、グラフ、モデル・ブラウザー、その他の種類の出力中に表示できます。
- ラベルを名前として読み込む: 短いフィールド名ではなく、SAS ファイルの詳細な変数ラベルを使用する場合に選択します。このラベルは IBM SPSS Modeler の変数名として使用します。

Excel ソース・ノード

Excel ソース・ノードでは、Microsoft Excel から .xlsx ファイル形式でデータをインポートできます。

ファイルの種類。インポートする Excel ファイルを選択します。

インポート・ファイル。インポートするスプレッドシートの名前と場所を指定します。

名前付き範囲を使用。Excel ワークシート内で定義された名前付きのセルの範囲を指定できるようになります。省略記号ボタン (「...」) をクリックして、利用できる範囲のリストから選択します。名前付きの範囲が使用されると、その他のワークシートとデータ範囲の設定が以後適用不能になり、結果として無効になります。

ワークシートを選択。インデックスまたは名前のどちらかで、インポートするワークシートを指定します。

- インデックスによる。インポートするワークシートのインデックス値を、最初のワークシートの 0 から始まり、2 番目のワークシートは 1 というように指定します。
- 名前順。インポートするワークシートの名前を指定します。省略記号ボタン (「...」) をクリックして、利用できるワークシートのリストから選択します。

ワークシートの範囲。最初の空白でない行から、または明示したセルの範囲のデータをインポートできます。

- 範囲の始点は最初に値を含む行。最初の空白でないセルに位置決めして、これをデータ範囲の左上隅として使用します。
- セルの明示的な範囲。行と列で明示した範囲を指定できるようになります。例えば、Excel の範囲 `A1:D5` を指定するには、最初のフィールドで `A1` と、2 番目のフィールドで `D5` (または、`R1C1` および `R5C4`) と入力します。指定した範囲内のすべての行が、空白行も含めて、返されます。

空白行。複数の空白行があると、そこで 「読み込みの停止」 することを選択するか、その空白行も含めてワークシートの最後まですべてのデータを読み込むために「空白行を返す」を選択することができます。

最初の行に列名を指定。指定された範囲の最初の行がフィールド (列) 名として使用されることを示します。選択されないと、フィールド名が自動的に生成されます。

フィールドのストレージと尺度

Excel から値を読み込むときに、数値のストレージ付きのフィールドの尺度は連続型として読み取られ、文字列フィールドは名義型として読み取られます。「データ型」タブで尺度 (連続型と名義型) を変更できますが、ストレージは自動的に決定されます。ただし、必要に応じて、置換ノードまたはフィールド作成ノードで `to_integer` などの変換関数を使用して、ストレージを変更することもできます。 詳しくは、トピック 9 ページの『フィールドのストレージと形式の設定』を参照してください。

デフォルトでは、数値と文字列値の混在は数字として読み込まれます。つまり、すべての文字列値は、IBM SPSS Modeler 内でヌル (システム欠損値) に設定されます。Excel とは異なり、IBM SPSS Modeler では

フィールド内でのストレージ・タイプの混在が許可されていないため、この現象が発生します。これを避けるには、Excel スプレッドシート内でセルの形式を手動で「**テキスト**」に設定します。そのようにすると、すべての値 (数字も含めて) が文字列として読み込まれます。

XML ソース・ノード

この機能は SPSS Modeler Professional および SPSS Modeler Premium で使用できます。

XML ソース・ノードを使用して、ファイルから IBM SPSS Modeler ストリームにデータを XML 形式でインポートできます。 XML はデータ交換に使用する標準言語で、多くの組織がこうした目的で選択する形式です。例えば国税庁は、オンラインで送信された XML 形式の確定申告書のデータを分析する必要があります (http://www.w3.org/standards/xml/を参照してください)。

XML データを IBM SPSS Modeler ストリームにインポートすると、ソースに幅広い予測分析機能を実行できます。XML データがテーブル形式で解析されます。それぞれの列は、XML の要素および属性の入れ子のレベルに対応します。XML 項目は XPath 形式で表示されます (http://www.w3.org/TR/xpath20/を参照)。

重要: XML ソース・ノードは、名前空間宣言を考慮しません。そのため、例えば、XML ファイルでは、コロン (:) 文字を `name` タグに含めることはできません。含めると、実行時に不正な文字に関するエラーが発生します。

単一ファイルの読み取り: デフォルトでは、SPSS Modeler は単一のファイルを読み取ります。このファイルは「**XML** データ ソース」フィールドで指定します。

ディレクトリー内のすべての **XML** ファイルを読み込む: 特定のディレクトリーのすべての XML ファイルを読み込む場合、このオプションを選択します。表示される 「ディレクトリー」 フィールドで場所を指定します。「サブディレクトリーを含める」 チェック ボックスをオンにし、指定したディレクトリーのすべてのサブディレクトリーから XML ファイルを読み込みます。

XML データ ソース: インポートする XML ソース・ファイルの完全パスとファイル名を入力するか、「参照」ボタンを使用してファイルを検索します。

XML スキーマ: (オプション) XML 構造を読み込む XSD または DTD の完全パスまたはファイル名を指定するか、「参照」ボタンを使用してこのファイルを検索します。このフィールドを空白にすると、構造が XML 入力ファイルから読み込まれます。XSD ファイルまたは DTD ファイルには複数のルート要素があります。この場合、フォーカスを異なるフィールドに変更すると、使用するルート要素を選択するダイアログが表示されます。 詳しくは、トピック 50 ページの『複数のルート要素からの選択』を参照してください。

注: XSD インディケーターは SPSS Modeler では無視されます。

XML 構造: XML ソース・ファイル (「**XML** スキーマ」 フィールドで指定している場合はスキーマ) の構造を示す階層ツリー。レコードの境界を定義する場合、要素を選択して右矢印をクリックし、項目を「レコード」 フィールドにコピーします。

表示属性: 「**XML** 構造」 フィールドの XML 要素の属性を、表示または非表示にします。

レコード **(XPath** 式**):**「XML 構造」フィールドからコピーした要素の XPath シンタックスを示します。この要素は XML 構造内で強調表示され、レコードの境界を定義します。入力ファイルにこの要素が出現するごとに、新しいレコードが作成されます。このフィールドが空白である場合、ルート下の最初の子要素がレコードの境界として使用されます。

すべてのデータを読み込む**:** デフォルトでは、ソース・ファイルのすべてのデータがストリームに読み込まれます。

読み込むデータを指定**:** 各要素、属性、または両方をインポートする場合、このオプションを選択します。このオプションを選択すると、「フィールド」テーブルで、インポートするデータを指定できます。

フィールド**:**「読み込むデータを指定」 オプションを選択している場合、このテーブルには、インポートするよう選択された要素と属性が表示されます。要素または属性の XPath シンタックスを XPath 列に直接入力することも、XML 構造の要素または属性を選択して、右方向矢印ボタンをクリックして、項目をテーブルにコピーすることもできます。要素のすべての子要素および属性をコピーするには、XML 構造の要素と以下を選択し、二重矢印のボタンをクリックします。

- **XPath:** インポートする項目の XPath シンタックス。
- 場所**:** インポートする項目の XML 構造での位置。「固定パス」 には、XML 構造で強調表示された要素に関連する項目のパス (または強調表示される要素がない場合、ルートの下の最初の子要素) が表示されます。「任意の位置」 は、XML 構造の任意の場所の指定された名前の項目を示します。「ユーザー指定」 は、XPath 列に位置を直接入力する場合に表示されます。

複数のルート要素からの選択

正しいŒ`式の XML ファイルが設定できるルート要素は 1 つだけですが、XSD または DTD ファイルには複数のルートを含むことができます。いずれかのルートが XML ソース・ファイルと一致する場合、そのルートが使用されます。一致しない場合は、使用するルートを選択する必要があります。

表示するルートを選択**:** 使用するルート要素を選択します。デフォルトは、XSD 構造または DTD 構造の最初のルート要素です。

XML ソース・データの不要なスペースの削除

XML ソース・データ内での改行は、[CR][LF] 文字を組み合わせて行うことができます。これらの改行は、次のように文字列の途中で発生する場合があります。

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

Web ブラウザーなど、いくつかのアプリケーションでファイルが開いている場合、これらの改行が表示できない場合があります。ただし、データを XML ソース・ノード経由でストリームに読み込むと、改行が一連のスペース文字に変換されます。

置換ノードを使用して不要なスペースを削除することにより、これを修正することができます。

スペースを削除する例を次に示します。

1. 置換ノードを XML ノードに接続します。
2. 置換ノードを開いてフィールド・ピッカーを使用し、不要なスペースのあるフィールドを選択します。
3. 「置換」 を 「条件を指定」 に設定、「条件」 を 「**true**」 に設定します。
4. 「置換後の文字列」フィールドに `replace(" ","",@FIELD)` と入力して「OK」をクリックします。
5. テーブル・ノードを置換ノードに接続してストリームを実行します。

テーブル・ノード出力に、スペースが追加されていない状態のテキストが表示されます。

ユーザー入力ノード

ユーザー入力ノードを利用すれば、最初から、または既存のデータを変更して、合成データを簡単に作成できます。これは、モデル作成用の検定データ・セットを作成する場合などに役立ちます。

最初からデータを作成

ユーザー入力ノードは「入力」パレットにあり、直接ストリーム領域に追加することができます。

1. ノード・パレットの「入力」タブをクリックします。
2. ユーザー入力ノードをストリーム領域上にドラッグ・アンド・ドロップするか、またはダブルクリックします。
3. 追加したユーザー入力ノードをダブルクリックしてダイアログ・ボックスを表示し、フィールドと値を指定します。

注：「入力」パレットから選択されたユーザー入力ノードには、フィールドやデータ情報などは何も指定されていません。そのため、合成データをすべて最初から作成することができます。

既存のデータ・ソースからのデータの生成

ストリーム中の任意の非ターミナル・ノードから、ユーザー入力ノードを生成することもできます。

1. ストリーム内の、ノードを置換する位置を決めます。
2. データをユーザー入力ノードに取り込むノードを右クリックして、メニューから「ユーザー入力ノードの生成」を選択します。
3. ユーザー入力ノードがそのノードに関連付けられているすべての下流プロセスとともに表示され、そこに存在していた既存のノードと置換されます。生成されたユーザー入力ノードは、メタデータからデータ構造およびフィールドのデータ型情報 (利用できる場合) をすべて継承します。

注：まだデータがストリーム中のすべてのノードを通過していない場合、ノードは完全にインスタンス化されていないため、ユーザー入力ノードで置換する際にストレージおよびデータ値を利用できない可能性があります。

ユーザー入力ノードのオプションの設定

ユーザー入力ノードのダイアログ・ボックスには、合成データの値を入力したり、データ構造を定義するために使用できるさまざまなツールが用意されています。生成されたノードの「データ」タブのテーブルには、元のデータ・ソースのフィールド名が表示されます。「入力」パレットから追加したノードの場合、テーブルには何も表示されません。このテーブルのオプションを使用して、次のような作業を実行できます。

- テーブルの右にある、「新規フィールドの追加」ボタンを使用したフィールドの追加。
- 既存のフィールド名の変更。
- 各フィールドのデータ・ストレージの指定。
- 値の指定。
- 表示されているフィールドの順序を変更。

データの入力

各フィールドに対して値を指定したり、テーブルの右側にある値ピッカー・ボタンを使用して、オリジナルのデータ・セットから値を挿入することができます。値の指定方法の詳細は、次の規則を参照してください。フィールドをブランクのままにしておくこともできます。ブランクのままのフィールドには、システムのヌル値 ($null$) が設定されます。

文字列値を指定するには、以下のように、単にスペースで区切って「値」列に入力します。

```
Fred Ethel Martin
```

スペースを含む文字列は、次のように二重引用符で囲まれます。

```
"Bill Smith" "Fred Martin" "Jack Jones"
```

数値型フィールドの場合は、次のように複数の値を同じ方法 (間にスペースを入れて一覧表示する方法) で入力できます。

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

または、同じ一連の数値を、範囲 (10, 20) とその間のステップ (2) を使用して指定することもできます。この方法を使用する場合は、次のように入力します。

```
10,20,2
```

上記の 2 通りの方法は、次のように一方を他方に埋め込んで組み合わせることができます。

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

この入力では、次の値が生成されます。

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

日付と時間の値は、「ストリーム・プロパティー」ダイアログ・ボックスで選択した現在のデフォルト形式を使用して、次のように入力することができます。

```
11:04:00 11:05:00 11:06:00
```

```
2007-03-14 2007-03-15 2007-03-16
```

日付と時間の両方のコンポーネントを含むタイムスタンプ値の場合は、次のように二重引用符を使用する必要があります。

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

詳細は、後述のデータ・ストレージに関するコメントを参照してください。

データの生成 **:** ストリームを実行するとレコードがどのようにして生成されるかを指定できます。

- すべての組み合わせ :フィールド値の考えられるあらゆる組み合わせを含む記録を生成するため、各フィールド値は複数のレコードに表示されます。これにより、予想以上に多くのデータが生成されることがあるので、このノードにサンプル・ノードを付け加えるとよいでしょう。
- 順番に :データ・フィールド値が指定されている順序でレコードを作成します。各フィールド値のみが 1 件のレコードに表示されます。生成されるレコード数は、1 つのフィールドの最大の値に等しくなります。フィールドの値がその最大数未満の場合、未定義 ($null$) の値が挿入されます。

例の表示

例えば次のように入力すると、以下の表で例示する 2 件のレコードが生成されます。

- 年齢**:** 30,60,10
- **BP:** LOW (低)

- コレステロール: NORMAL HIGH (正常 高)
- 薬品: (空白のまま)

表 6. 「データの生成」フィールドを「すべての組み合わせ」に設定した場合：

Age	BP	Cholesterol (コレステロール)	Drug (薬品)
30	LOW	NORMAL	$null$
30	LOW	HIGH	$null$
40	LOW	NORMAL	$null$
40	LOW	HIGH	$null$
50	LOW	NORMAL	$null$
50	LOW	HIGH	$null$
60	LOW	NORMAL	$null$
60	LOW	HIGH	$null$

表 7. 「データの生成」フィールドを「順序どおり」に設定した場合：

Age	BP	Cholesterol (コレステロール)	Drug (薬品)
30	LOW	NORMAL	$null$
40	$null$	HIGH	$null$
50	$null$	$null$	$null$
60	$null$	$null$	$null$

データ・ストレージ

ストレージは、フィールド中へのデータの格納方法を表しています。例えば、1 と 0 の値をとるフィールドは整数データを格納します。これはデータの使用法を記述する測定の尺度とは異なり、ストレージに影響を与えません。例えば、値 1 と 0 をとる整数フィールドの測定の尺度をフラグ型に設定することができます。通常は、1 = *True*、0 = *False* を示します。ストレージはソースで確定する必要がありますが、測定の尺度はストリームのどこでもデータ型ノードを使用して変更できます。詳しくは、トピック 145 ページの『尺度』を参照してください。

指定できるストレージ・タイプを次に示します。

- 文字列 非数値データ (別名、英数字データ) を含むフィールドに使用されます。文字列には、*fred*、*Class 2*、または *1234* など、任意の文字のシーケンスを含めることができます。注意を要するのは、文字列内の数字は計算には使えないことです。
- 整数 値が整数で示されるフィールドです。
- 実数 値は数字で示され、小数点を含むことがあります (整数に限定されません)。表示形式は「ストリーム プロパティー」ダイアログ ボックスで指定し、データ型ノード(「形式」タブ) の各フィールドでオーバーライドできます。
- 日付 年、月、日など、標準形式で指定された日付の値です (2007-09-26 など)。「ストリーム・プロパティー」ダイアログ・ボックスで特定の形式を指定します。
- 時間 期間として測定される時間です。例えば、「ストリーム プロパティー」ダイアログ ボックスで指定した現在の時間形式に応じて、1 時間 26 分 38 秒続くサービスコールを「01:26:38」と表現することができます。

- タイムスタンプ 例えば 2007-09-26 09:04:00 のように、日付と時刻の両方の構成要素を含む値です。この場合も、「ストリーム・プロパティー」ダイアログボックスの現在の日付と時間の形式に従います。日付と時刻が別々の値ではなく 1 つの値として解釈されるようにするには、タイム・スタンプ値を二重引用符で囲む必要があります (この規則は、ユーザー入力ノードで値を入力する場合などに適用されます)。
- リスト SPSS Modeler バージョン 17 で、地理空間および集合の新しい尺度とともに導入されました。リストのストレージ フィールドには、単一のレコードに対する複数の値が入ります。その他すべてのストレージ タイプのリスト版が存在します。

表 8. リストのストレージ タイプを示すアイコン

アイコン	ストレージ・タイプ
[A]	文字列のリスト
[◇]	整数のリスト
[#]	実数のリスト
[🕒]	時間のリスト
[📅]	日付のリスト
[📅]	タイムスタンプのリスト
[[]]	0 よりも大きな深さを持つリスト

さらに、集合の尺度とともに使用する場合は、以下の尺度のリスト版もあります。

表 9. リストの尺度を示すアイコン

アイコン	尺度
[✎]	連続のリスト
[▮]	カテゴリのリスト
[●]	フラグのリスト

表 9. リストの尺度を示すアイコン (続き)

アイコン	尺度
[]	名義のリスト
[]	順序のリスト

リストは、Analytic Server、地理空間、または可変長ファイルのいずれかのソース・ノードで SPSS Modeler にインポートするか、フィールド作成ノードまたは置換のフィールド操作ノードを使用してストリーム内で作成することができます。

リストと、集合および地理空間の尺度との相互作用について詳しくは、 12 ページの『リストス トレージおよび関連する尺度』を参照してください。

ストレージの変換: 置換ノードで `to_string` や `to_integer` などのさまざまな変換関数を使用して、フィールドのストレージを変換することができます。詳しくは、トピック 170 ページの『置換ノードを使ったストレージの変換』を参照してください。 変換関数および日付や時刻の値のような、入力に特別な型が必要なその他の関数は、「ストリームのプロパティー」ダイアログ・ボックスに指定されている現在の形式に依存します。例えば、値が *Jan 2003*、*Feb 2003* などの文字列フィールドを日付ストレージへ変換する場合、ストリームのデフォルトの日付形式として「**MON YYYY**」を選択します。変換関数は、フィールド作成ノードのフィールド作成計算時の一時変換でも、利用できます。また、フィールド作成ノードを使用して、カテゴリー値を含む文字列フィールドの読み取りなど、他の操作も実行できます。詳しくは、トピック 169 ページの『フィールド作成ノードを使用して値を再コード化する』を参照してください。

混在データの読み込み: 注意を要するのは、数値ストレージ (整数、実数、時間、タイムスタンプ、または日付のいずれか) を含むフィールドで読み込む場合、数値以外の値は、ヌル値またはシステム欠損値に設定されることです。これは、一部のアプリケーションと異なり、IBM SPSS Modeler では、1 つのフィールド内でストレージ・タイプが混在することは許されないためです。これを回避するには、ソース・ノードでストレージ・タイプを変更するか、必要な場合は外部アプリケーションで、データが混ざり合ったフィールドを文字列として読み込む必要があります。

注 : インスタンス化されている場合、生成されたユーザー・ソース・ノードには、すでにストレージ情報が入力ノードから収集されて存在していることもあります。インスタンス化されていないノードには、ストレージ・タイプまたは使用タイプの情報はありません。

値の指定規則

シンボル値フィールドの場合は、次のように複数の値の間にスペースを入れる必要があります。

```
HIGH MEDIUM LOW
```

数値型フィールドの場合は、次のように複数の値を同じ方法 (間にスペースを入れて一覧表示する方法) で入力できます。

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

または、同じ一連の数値を、範囲 (10, 20) とその間のステップ (2) を使用して指定することもできます。この方法を使用する場合は、次のように入力します。

```
10,20,2
```

上記の 2 通りの方法は、次のように一方を他方に埋め込んで組み合わせることができます。

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

この入力では、次の値が生成されます。

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

シミュレーション生成ノード

シミュレーション生成ノードは、シミュレーション・データを生成する簡単な方法を提供します。データは、ユーザーが指定した統計分布を使用して履歴データを使用せずに生成されるか、または既存の履歴データでシミュレーション適合ノードを実行して得られた分布を自動的に使用することにより生成されます。シミュレーション・データの生成が役に立つのは、モデルの入力内に不確定性が存在する予測モデルの結果を評価する場合です。

履歴データを使用しないデータの作成

シミュレーション生成ノードは「入力」パレットにあり、ストリーム・キャンバスに直接追加できます。

1. ノード・パレットの「入力」タブをクリックします。
2. シミュレーション生成ノードをストリーム・キャンバス上にドラッグ・アンド・ドロップするか、またはダブルクリックします。
3. 追加したシミュレーション生成ノードをダブルクリックしてダイアログ・ボックスを表示し、フィールド、ストレージ・タイプ、統計分布、および分布パラメーターを指定します。

注: 「入力」パレットから選択されたシミュレーション生成ノードには、フィールドや分布情報などは何も指定されていません。これにより、履歴データを使用しないシミュレーション・データを作成できます。

既存の履歴データを使用したシミュレーション・データの生成

シミュレーション生成ノードは、シミュレーション適合ターミナル・ノードを実行して作成することもできます。

1. シミュレーション適合ノードを右クリックして、メニューから「実行」を選択します。
2. シミュレーション生成ノードが、シミュレーション適合ノードへの更新リンクと共にストリーム・キャンバスに表示されます。
3. シミュレーション生成ノードは、生成時に、シミュレーション適合ノードのフィールド、ストレージ・タイプ、および統計分布情報をすべて継承します。

シミュレーション適合ノードへの更新リンクの定義

シミュレーション生成ノードとシミュレーション適合ノードの間にリンクを作成できます。このリンクは、履歴データへの適合により決定された、最適な適合分布情報で 1 つ以上のフィールドを更新する際に役に立ちます。

1. シミュレーション生成ノードを右クリックします。
2. メニューから「更新リンクの定義」を選択します。カーソルがリンク・カーソルに変わります。

3. 別のノードをクリックします。このノードがシミュレーション適合ノードである場合、リンクが作成されます。このノードがシミュレーション適合ノードではない場合、リンクは作成されず、カーソルが通常のカーソルに戻ります。

シミュレーション適合ノードのフィールドがシミュレーション生成ノードのフィールドとは異なる場合は、その内容を示すメッセージが表示されます。

シミュレーション適合ノードを使用してリンク先のシミュレーション生成ノードを更新する場合、両方のノードに同じフィールドが存在するかどうか、およびシミュレーション生成ノードのフィールドがロック解除されているかどうかに応じて、結果が変わります。シミュレーション適合ノードの更新結果を次の表に示します。

表 *10.* シミュレーション適合ノードの更新結果

	シミュレーション 適合ノードのフィールド	
シミュレーション生成ノードのフィールド	存在する	欠損値
存在してロック解除されている。	フィールドが上書きされます。	フィールドが削除されます。
存在しない。	フィールドが追加されます。	変更なし。
存在してロックされている。	フィールドの分布は上書きされません。「適合の詳細」ダイアログ・ボックスの情報および相関が更新されます。	フィールドは上書きされません。相関がゼロに設定されます。
「再適合する場合は最小値および最大値を消去しない」チェック・ボックスが選択されている。	フィールドは上書きされます。	「最小値」および「最大値」列の値はそのままです。
「再適合する場合は相関を再計算しない」チェック・ボックスが選択されている。	フィールドがロック解除されている場合は上書きされます。	相関は上書きされません。

シミュレーション適合ノードへの更新リンクの削除

以下の手順を実行することにより、シミュレーション生成ノードとシミュレーション適合ノードの間のリンクを削除できます。

1. シミュレーション生成ノードを右クリックします。
2. メニューから「更新リンクの削除」を選択します。リンクが削除されます。

シミュレーション生成ノードのオプションの設定

シミュレーション生成ノードのダイアログ・ボックスの「データ」タブのオプションを使用して、以下の操作を行うことができます。

- 各フィールドの統計分布情報の表示、指定、および編集。
- フィールド間の相関の表示、指定、および編集。
- シミュレーションする反復およびケースの数の指定。

項目の選択: シミュレーション生成ノードの 3 つのビュー (「シミュレーションしたフィールド」、「相関」、「詳細オプション」) を切り替えることができます。

「シミュレーションしたフィールド」ビュー

シミュレーション生成ノードが、履歴データを使用してシミュレーション適合ノードから生成または更新された場合、「シミュレーションしたフィールド」ビューで、各フィールドの統計分布情報を表示して編集することができます。各フィールドに関する以下の情報が、シミュレーション適合ノードからシミュレーション生成ノードの「タイプ」タブにコピーされます。

- 尺度
- 値
- 欠損値
- チェック
- 役割

履歴データがない場合にフィールドを定義してその分布を指定するには、ストレージ・タイプと分布タイプを選択して、必須パラメーターを入力します。この方法でデータを生成すると、例えば、「タイプ」タブまたはタイプ・ノードでデータがインスタンス化されるまで、各フィールドの尺度に関する情報は使用できません。

「シミュレーションしたフィールド」ビューには複数のツールが表示されており、以下のタスクを実行するのに使用できます。

- フィールドの追加または削除。
- 表示されているフィールドの順序を変更。
- 各フィールドのストレージ・タイプの指定。
- 各フィールドの統計分布の指定。
- 各フィールドの統計分布のパラメーター値の指定。

シミュレーションしたフィールド: 「入力」パレットからシミュレーション生成ノードがストリーム・キャンバスに追加された場合、この表には空白の行が 1 行表示されます。この行を編集すると、表の最後に空白の行が 1 行新しく追加されます。シミュレーション適合ノードからシミュレーション生成ノードが作成された場合、この表には、履歴データのフィールドごとに 1 つの行が表示されます。表に行を追加するには、「新規フィールドの追加」アイコンをクリックします。

「シミュレーションしたフィールド」表は、以下の列から構成されます。

- フィールド: フィールドの名前が入ります。フィールド名を編集する場合は、セルに入力します。
- ストレージ: この列のセルには、ストレージ・タイプのドロップダウン・リストが表示されます。ストレージ・タイプとしては、文字列、整数、実数、時間、日付、またはタイムスタンプを利用できます。ストレージ・タイプを選択すると、「分布」列で使用可能な分布が決定されます。シミュレーション適合ノードからシミュレーション生成ノードが作成された場合は、シミュレーション適合ノードからストレージ・タイプがコピーされます。

  注: ストレージ・タイプが日時のフィールドでは、分布パラメーターを整数値として指定する必要があります。例えば、基準となる日付 1970 年の 1 月 1 日を指定する場合は、整数値 0 を使用します。符号付き整数値は、1970 年 1 月 1 日の午前 0 時以降 (または、以前) の秒数を表します。
- ステータス: 「ステータス」列のアイコンは、各フィールドの適合のステータスを示します。

フィールドに対して分布が指定されていないか、または 1 つ以上の分布パラメーターが欠落しています。シミュレーションを実行するには、このフィールドに対して分布を指定し、パラメーターに有効な値を入力する必要があります。

フィールドは最も適合する分布に設定されています。
注: このアイコンが表示されるのは、シミュレーション生成ノードがシミュレーション適合ノードから作成された場合のみです。

最も適合する分布が、「適合の詳細」サブダイアログ・ボックスの代替の分布で置き換えられています。詳しくは、トピック 63 ページの『適合の詳細』を参照してください。

分布が手動で指定または編集されています。複数のレベルで指定したパラメーターが含まれている可能性があります。

- ロック済み**:** シミュレーションしたフィールドをロックする (ロック・アイコンがある列のチェック・ボックスを選択する) と、リンク先のシミュレーション適合ノードによる自動更新の対象から除外されます。これが最も有効になるのは、分布を手動で指定して、リンク先のシミュレーション適合ノードの実行時に、自動分布適合によって影響を受けないようにする必要がある場合です。
- 分布**:** この列のセルには、統計分布のドロップダウン・リストが表示されます。ストレージ・タイプを選択すると、指定したフィールドのこの列で使用可能な分布が決定されます。詳しくは、トピック 66 ページの『分布』を参照してください。

  注**:** すべてのフィールドに「固定」分布を指定することはできません。生成されたデータ内のすべてのフィールドを固定する場合、ユーザー入力ノードに続けてバランス・ノードを使用します。
- パラメーター**:** 適合する各分布に関連する分布パラメーターがこの列に表示されます。パラメーターの複数の値はカンマで区切られています。パラメーターに対して複数の値を指定すると、シミュレーションで複数の反復が生成されます。詳しくは、トピック 66 ページの『反復回数』を参照してください。パラメーターが欠落している場合は、「ステータス」列のアイコンで示されます。パラメーターに値を指定するには、該当するフィールドに対応する行でこの列をクリックして、リストから「指定」を選択します。「パラメーターの指定」サブダイアログ・ボックスが開きます。詳しくは、トピック 64 ページの『パラメーターの指定』を参照してください。「分布」列で「経験的」が選択されている場合、この列は無効です。
- 最小値、最大値**:** この列では、一部の分布に対して、シミュレーション・データの最小値、最大値、またはその両方を指定できます。シミュレーション・データのうち、最小値より小さいデータおよび最大値より大きいデータは、指定した分布で有効な場合でも拒否されます。最小値および最大値を指定するには、該当するフィールドに対応する行でこの列をクリックして、リストから「指定」を選択します。「パラメーターの指定」サブダイアログ・ボックスが開きます。詳しくは、トピック 64 ページの『パラメーターの指定』を参照してください。「分布」列で「経験的」が選択されている場合、この列は無効です。

最も近い適合を使用**:** シミュレーション生成ノードが、履歴データを使用してシミュレーション適合ノードから自動的に作成された場合、および「シミュレーションしたフィールド」表で単一の行が選択されている

場合にのみ有効です。選択した行のフィールドの情報を、フィールドに最適な適合分布の情報で置き換えます。選択した行の情報が編集されている場合、このボタンを押すと、現在の情報がシミュレーション適合ノードで決定された最適な適合分布にリセットされます。

適合の詳細: シミュレーション生成ノードがシミュレーション適合ノードから自動的に作成された場合にのみ有効です。「適合の詳細」サブダイアログ・ボックスが開きます。詳しくは、トピック 63 ページの『適合の詳細』を参照してください。

「シミュレーションしたフィールド」ビューの右側にあるアイコンを使用すると、いくつかの便利なタスクを実行できます。次の表で、これらのアイコンを説明します。

表 *11.* 「シミュレーションしたフィールド」ビューのアイコン:

アイコン	ツールチップ	説明
	分布パラメーターの編集	「シミュレーションしたフィールド」表で単一の行が選択されている場合にのみ有効です。選択した行の「パラメーターの指定」サブダイアログ・ボックスを開きます。詳しくは、トピック 64 ページの『パラメーターの指定』を参照してください。
	新規フィールドの追加	「シミュレーションしたフィールド」表で単一の行が選択されている場合にのみ有効です。「シミュレーションしたフィールド」表の最後に空白の行を 1 行新しく追加します。
	複数コピーの作成	「シミュレーションしたフィールド」表で単一の行が選択されている場合にのみ有効です。「フィールドの複製」サブダイアログ・ボックスを開きます。詳しくは、トピック 63 ページの『フィールドの複製』を参照してください。
	選択したフィールドの削除	「シミュレーションしたフィールド」表から選択した行を削除します。
	一番上に移動	選択した行が、まだ「シミュレーションしたフィールド」表の一番上の行にない場合にのみ有効です。選択した行を「シミュレーションしたフィールド」表の一番上に移動します。このアクションは、シミュレーション・データのフィールドの順序に影響を与えます。
	上へ移動	選択した行が、「シミュレーションしたフィールド」表の一番上の行にない場合にのみ有効です。選択した行を「シミュレーションしたフィールド」表の 1 つ上の位置に移動します。このアクションは、シミュレーション・データのフィールドの順序に影響を与えます。

表 11. 「シミュレーションしたフィールド」ビューのアイコン (続き) :

アイコン	ツールチップ	説明
	下へ移動	選択した行が、「シミュレーションしたフィールド」表の一番下の行にない場合にのみ有効です。選択した行を「シミュレーションしたフィールド」表の 1 つ下の位置に移動します。このアクションは、シミュレーション・データのフィールドの順序に影響を与えます。
	一番下に移動	選択した行が、まだ「シミュレーションしたフィールド」表の一番下の行にない場合にのみ有効です。選択した行を「シミュレーションしたフィールド」表の一番下に移動します。このアクションは、シミュレーション・データのフィールドの順序に影響を与えます。

再適合する場合は最小値および最大値を消去しない。 選択すると、接続したシミュレーション適合ノードを実行して分布が更新された際に、最小値および最大値が上書きされません。

「相関」ビュー

予測モデルの入力フィールドは、多くの場合、身長と体重などのように、相関があることがわかっています。シミュレーション値でこれらの相関を保持するには、シミュレーションするフィールド間の相関を示す必要があります。

シミュレーション生成ノードが、履歴データを使用してシミュレーション適合ノードから生成または更新された場合、「相関」ビューで、フィールドのペア間で計算された相関を表示して編集することができます。履歴データがない場合は、複数のフィールド間にどのような相関があるかという認識に基づいて、相関を手動で指定できます。

注: データが生成される際には、半正定値であるかどうか、および反転可能かどうかを判別するために相関行列が自動的にチェックされます。列に線形独立性がある場合、行列は反転可能です。相関行列が反転できない場合、反転が可能になるように自動的に調整されます。

行列形式またはリスト形式を選択して、相関を表示することができます。

相関行列。フィールドのペア間の相関を行列で表示します。フィールド名は、アルファベット順に行列の左上側から表示されます。対角より下のセルのみが編集可能です。-1.000 以上 1.000 以下の値を入力する必要があります。対角より上のセルは、対角より下の対称の位置にあるセルからフォーカスが外れた場合に更新されます。これにより、両方のセルに同じ値が表示されます。対角のセルは常に無効で、相関の値はすべて 1.000 です。その他のすべてのセルのデフォルト値は 0.000 です。値 0.000 は、フィールドの関連するペアの間に相関がないことを示します。行列には、連続型および順序型のフィールドのみが含まれています。名義型、カテゴリー型、フラグ型、および「固定」分布が割り当てられている各フィールドは、表には表示されません。

相関リスト。フィールドのペア間の相関を表で表示します。表の各行は、フィールドのペア間の相関を示しています。行は追加も削除もできません。見出しが「フィールド 1」および「フィールド 2」である列には

フィールド名が表示されており、編集はできません。「相関」列には相関が表示されており、編集が可能です。-1.000 以上 1.000 以下の値を入力する必要があります。すべてのセルのデフォルト値は 0.000 です。リストには、連続型および順序型のフィールドのみが含まれています。名義型、カテゴリー型、フラグ型、および「固定」分布が割り当てられている各フィールドは、リストには表示されません。

相関のリセット。「相関のリセット」ダイアログ・ボックスが開きます。履歴データが使用可能な場合は、以下の 3 つのオプションのいずれかを選択できます。

- 適合。現在の相関を、履歴データを使用して計算した相関で置き換えます。
- ゼロ。現在の相関をゼロで置き換えます。
- キャンセル。ダイアログ・ボックスを閉じます。相関は変更されません。

履歴データを使用できない状況で相関に変更を加えた場合は、現在の相関をゼロで置き換えるか、またはキャンセルするかを選択できます。

表示形式。相関を行列として表示する場合は、「表」を選択します。相関をリストとして表示する場合は、「一覧」を選択します。

再適合する場合は相関を再計算しない。相関を手動で指定して、シミュレーション適合ノードと履歴データを使用した分布の自動適合時に上書きされないようにする場合は、このオプションを選択します。

カテゴリー分布の入力に適合多元分割表を使用。デフォルトでは、カテゴリー型分布を持つすべてのフィールドが分割表に表示されています (または、カテゴリー型分布を持つフィールドの数によっては、多元分割表)。分割表は、相関と同様に、シミュレーション適合ノードの実行時に作成されます。分割表は表示できません。このオプションを選択すると、分割表の実際のパーセンテージを使用して、カテゴリー型分布を持つフィールドがシミュレーションされます。つまり、名義型フィールド間のアソシエーションが、新規シミュレーション・データ内に再作成されます。このオプションを選択解除すると、分割表の予測パーセンテージを使用して、カテゴリー型分布を持つフィールドがシミュレーションされます。フィールドを変更すると、そのフィールドは分割表から削除されます。

「詳細オプション」ビュー

シミュレーションするケース数。シミュレーションするケース数、および反復に名前を付ける方法を指定するオプションが表示されます。

- 最大ケース数。生成するシミュレーション・データの最大ケース数と、関連する目標値を指定します。デフォルト値は 100,00、最小値は 1000、最大値は 2,147,483,647 です。
- 反復回数。この数値は自動的に計算され、編集できません。反復は、分布パラメーターに複数の値が指定されるたびに自動的に作成されます。
- 合計行数。反復回数が 1 より大きい場合にのみ有効です。この数値は表示される式を使用して自動的に計算され、編集できません。
- 反復フィールドの作成。反復回数が 1 より大きい場合にのみ有効です。選択すると、「名前」フィールドが有効になります。詳しくは、トピック 66 ページの『反復回数』を参照してください。
- 名前。「反復フィールドの作成」チェック・ボックスが選択されており、反復回数が 1 より大きい場合にのみ有効です。反復フィールドの名前を編集する場合は、このテキスト・フィールドに入力します。詳しくは、トピック 66 ページの『反復回数』を参照してください。

ランダム シード。ランダム・シードを設定すると、シミュレーションを再現することができます。

- 結果の再現。選択すると、「生成」ボタンと「ランダム シード」フィールドが有効になります。
- ランダム シード。「結果を再現」チェック・ボックスが選択されている場合にのみ有効です。このフィールドでは、ランダム・シードとして使用する整数を指定できます。デフォルト値は 629111597 です。

- 生成。「結果を再現」チェック・ボックスが選択されている場合にのみ有効です。「ランダム・シード」フィールドに、1 以上 999999999 以下の整数の擬似乱数を作成します。

フィールドの複製

「フィールドの複製」ダイアログ・ボックスで、選択したフィールドのコピーの作成数と、各コピーの命名方法を指定することができます。複雑な効果を調査する場合、フィールドのコピーを複数指定すると役に立ちます。例えば、さまざまな継続期間にわたる金利や成長率などを指定します。

ダイアログ・ボックスのタイトル・バーには、選択したフィールドの名前が表示されます。

作成するコピー数**:** 作成するフィールドのコピー数が表示されます。作成するコピーの数を選択するには、矢印をクリックします。作成するコピーの最小数は 1 で、最大数は 512 です。コピー数の初期値は 10 に設定されています。

接尾辞の文字のコピー**:** 各コピーのフィールド名の末尾に追加される文字が表示されます。これらの文字により、フィールド名とコピー数が分離されます。このフィールドに接尾辞の文字を入力して編集することができます。このフィールドは空にすることができます。その場合、フィールド名とコピー数の間に文字は挿入されません。デフォルトの文字は下線 (_) です。

初期コピー数**:** 初期コピー用の接尾辞の数値が表示されます。初期コピー数を選択するには、矢印をクリックします。初期コピー数の最小値は 1 で、最大値は 1000 です。デフォルトの初期コピー数は 1 です。

コピー数ステップ**:** 接尾辞の数値の増分が表示されます。増分を選択するには、矢印をクリックします。増分の最小値は 1 で、最大値は 255 です。増分の初期値は 1 に設定されています。

フィールド**:** コピーのフィールド名のプレビューが表示されます。「フィールドの複製」ダイアログ・ボックスのいずれかのフィールドが編集されると、このプレビューが更新されます。このテキストは自動的に生成され、編集することはできません。

OK: ダイアログ・ボックスで指定されたすべてのコピーを生成します。このコピーは、「シミュレーション生成ノード (Simulation Generate node)」ダイアログ・ボックスの「シミュレーションしたフィールド」テーブルの、コピーされたフィールドが含まれている行のすぐ下に追加されます。

キャンセル**:** ダイアログ・ボックスを閉じます。行われた変更はすべて破棄されます。

適合の詳細

「適合の詳細」ダイアログ・ボックスを使用できるのは、シミュレーション適合ノードを実行してシミュレーション生成ノードが作成または更新された場合だけです。このダイアログ・ボックスには、選択されたフィールドについて、自動的な分布の適合結果が表示されます。分布は適合度順に配置され、最も適合した分布が最初にリストされます。このダイアログ・ボックスでは、以下のタスクを実行することができます。

- 過去のデータに対して適合された分布を確認する。
- 適合されたいずれかの分布を選択する。

フィールド**:** 選択したフィールドの名前が表示されます。このテキストを編集することはできません。

処理 **(**指標**):** 選択したフィールドの尺度タイプが表示されます。この尺度タイプは、「シミュレーション生成ノード (Simulation Generate node)」ダイアログ・ボックスの「シミュレーションしたフィールド」テーブルから取得されます。尺度タイプを変更するには、矢印をクリックして、尺度タイプをドロップダウン・リストから選択します。3 つのオプション (「連続型」、「名義型」、「順序型」) があります。

分布**:** 「分布」テーブルには、尺度タイプに適したすべての分布が表示されます。過去のデータに対して適合された分布は、適合度を基準として、最も適合する分布から最も適合しない分布の順に配列されます。適合度は、シミュレーション適合ノードで選択された適合度統計によって決定されます。過去のデータに対して適合されていない分布は、適合された分布の下のテーブルにアルファベット順に表示されます。

分布テーブルには、以下の列が表示されます。

- 使用**:** 選択されたラジオ・ボタンは、フィールド用に現在選択されている分布を示します。「使用」列の該当する分布のラジオ・ボタンを選択することにより、最も適合する分布を上書きすることができます。「使用」列のラジオ・ボタンを選択すると、選択したフィールドの過去のデータのヒストグラム (または棒グラフ) に重ね合わせた分布の散布図も表示されます。一度に選択できる分布は 1 つだけです。
- 分布: 分布の名前が表示されます。この列は編集できません。
- 適合度統計**:** 分布について計算された適合度統計が表示されます。この列は編集できません。セルの内容は、以下のフィールドの尺度タイプによって異なります。
  - 連続**:** Anderson-Darling 検定と Kolmogorov-Smirnoff 検定の結果が表示されます。検定に関連する p 値も表示されます。シミュレーション適合ノードの適合度基準として選択された適合度統計が最初に表示されます。これを使用して、分布が並び替えられます。Anderson-Darling 統計量は、A=*aval* P=*pval* として表示されます。Kolmogorov-Smirnoff 統計量は、K=*kval* P=*pval* として表示されます。統計量を計算できない場合、数値の代わりに点が表示されます。
  - 名義型および順序型**:** カイ 2 乗検定の結果が表示されます。検定に関連する p 値も表示されます。統計量は、Chi-Sq=*val* P=*pval* として表示されます。分布が適合されていない場合、**「適合しない」**が表示されます。分布を数学的に適合させることができない場合、**「適合できません」**が表示されます。

  注: 経験分布のセルは常に空になります。
- パラメーター: 適合された各分布に関連する分布パラメーターが表示されます。各パラメーターは 1 つのスペースで区切られて、*parameter_name = parameter_value* として表示されます。カテゴリー型分布の場合、カテゴリーがパラメーター名になり、関連する確率がパラメーター値になります。分布が過去のデータに対して適合されていない場合、セルは空になります。この列は編集できません。

ヒストグラムのサムネール**:** 選択したフィールドの過去のデータのヒストグラムに重ね合わせた、選択した分布の散布図が表示されます。

分布のサムネール**:** 選択した分布の説明と図が表示されます。

OK: ダイアログ・ボックスを閉じ、選択したフィールドの「シミュレーションしたフィールド」テーブルの「尺度」列、「分布」列、「パラメーター」列、「最小値」列、「最大値」列の値が、選択した分布の情報で更新されます。また、「状態」列のアイコンも更新され、選択した分布がデータに最も適合する分布かどうかが反映されます。

キャンセル**:** ダイアログ・ボックスを閉じます。行われた変更はすべて破棄されます。

パラメーターの指定

「パラメーターの指定」ダイアログ・ボックスでは、選択したフィールドの分布に対してパラメーター値を手動で指定できます。また、選択したフィールドに対して別の分布を選択することもできます。

「パラメーターの指定」ダイアログ・ボックスは、以下の 3 つの方法で開くことができます。

- シミュレーション生成ノードのダイアログ・ボックスの「シミュレーションしたフィールド」表にあるフィールド名をダブルクリックします。
- 「シミュレーションしたフィールド」表の「パラメーター」または「最小値」、「最大値」列をクリックして、リストから「指定」を選択します。
- 「シミュレーションしたフィールド」表で行を選択して、「分布パラメーターの編集」アイコンをクリックします。

フィールド**:** 選択したフィールドの名前が表示されます。このテキストを編集することはできません。

分布: 選択したフィールドの分布が表示されます。これは、「シミュレーションしたフィールド」表から取得されます。分布を変更するには、矢印をクリックして、ドロップダウン・リストから分布を選択します。使用可能な分布は、選択したフィールドのストレージ・タイプに応じて異なります。

面。このオプションは、「分布」フィールドでダイス分布を選択した場合にのみ有効です。矢印をクリックして、フィールドを分割する面またはカテゴリーの数を選択します。面数の最小値は 2 で、最大値は 20 です。面数の初期値は 6 に設定されています。

分布パラメーター。「分布パラメーター」表には、選択した分布のパラメーターごとに 1 つの行が表示されます。

注**:** 分布では、形状パラメーター $\alpha = k$ および逆スケール パラメーター $\beta = 1/\theta$ の比率パラメーターを使用します。
テーブルには、次の 2 つの列があります。

- パラメーター。パラメーターの名前が入ります。この列は編集できません。
- 値。パラメーターの値が表示されます。シミュレーション適合ノードからシミュレーション生成ノードが作成または更新された場合、この列のセルには、分布を履歴データに適合させることによって決定されたパラメーター値が表示されます。「ソース・ノード」パレットからシミュレーション生成ノードがストリーム・キャンバスに追加された場合、この列のセルは空白です。値を編集する場合は、セルに入力します。各分布で必要なパラメーターおよび使用可能なパラメーター値について詳しくは、トピック 66 ページの『分布』を参照してください。

  パラメーターに複数の値を指定する場合は、カンマで区切る必要があります。パラメーターに対して複数の値を指定すると、シミュレーションで複数の反復が定義されます。1 つのパラメーターに対してのみ複数の値を指定できます。

  注**:** ストレージ・タイプが日時のフィールドでは、分布パラメーターを整数値として指定する必要があります。例えば、基準となる日付 1970 年の 1 月 1 日を指定する場合は、整数値 0 を使用します。

注**:** ダイス分布を選択した場合、「分布パラメーター」表は多少異なります。表には、面 (または、カテゴリー) ごとに 1 つの行が表示されます。また、「値」列と「確率」列も表示されています。「値」列には、各カテゴリーのラベルが表示されます。ラベルのデフォルト値は、1 から N までの整数値です。ここで、N は面数です。ラベルを編集する場合は、セルに入力します。セルには任意の値を入力できます。数値以外の値を使用する必要があり、ストレージ・タイプがまだ文字列に設定されていない場合は、データ・フィールドのストレージ・タイプを文字列に変更する必要があります。「確率」列には、各カテゴリーの確率が表示されます。確率は編集することはできず、1/N として計算されます。

プレビュー: 指定したパラメーターに基づいて、分布のサンプル・プロットを表示します。1 つのパラメーターに 2 つ以上の値が指定されている場合は、パラメーターのそれぞれの値に対してサンプル・プロットが表示されます。選択したフィールドで履歴データが使用できる場合は、履歴データのヒストグラム上に分布のプロットが重ね合わせられます。

オプション設定。シミュレーション・データに対して、最小値、最大値、またはその両方を指定する場合は、このオプションを使用します。シミュレーション・データのうち、最小値より小さいデータおよび最大値より大きいデータは、指定した分布で有効な場合でも拒否されます。

- 最小値を指定。選択すると、「次の値より小さい値を拒否」フィールドが有効になります。「経験的」分布を選択した場合、このチェック・ボックスは無効にされます。
- 次の値より小さい値を拒否。「最小値を指定」を選択した場合にのみ有効になります。シミュレーション・データの最小値を入力します。この値より小さいシミュレーション値はすべて拒否されます。
- 最大値を指定。選択すると、「次の値より大きい値を拒否」フィールドが有効になります。「経験的」分布を選択した場合、このチェック・ボックスは無効にされます。
- 次の値より大きい値を拒否。「最大値を指定」を選択した場合にのみ有効になります。シミュレーション・データの最大値を入力します。この値より大きいシミュレーション値はすべて拒否されます。

OK: ダイアログ・ボックスを閉じ、選択したフィールドの「シミュレーションしたフィールド」表の「分布」、「パラメーター」、「最小値」、および「最大値」の各列の値を更新します。「ステータス」列のアイコンも更新され、選択した分布を反映します。

キャンセル**:** ダイアログ・ボックスを閉じます。行われた変更はすべて破棄されます。

反復回数

固定フィールドまたは分布パラメーターに対して複数の値を指定した場合は、それぞれの値に対してシミュレーションされたケースのセット (事実上、別のシミュレーション) が個別に生成されます。これにより、フィールドまたはパラメーターを変更した際の影響を調べることができます。シミュレーションされたケースの各セットは、反復と呼ばれます。シミュレーション・データ内では、反復が積み重ねられます。

シミュレーション生成ノードのダイアログの「詳細オプション」ビューで「反復フィールドの作成」チェック・ボックスが選択されている場合、数値ストレージを持つ名義型フィールドとして、シミュレーション・データに反復フィールドが追加されます。このフィールドの名前を編集する場合は、「詳細オプション」ビューの「名前」フィールドに入力します。このフィールドには、シミュレーションされた各ケースがどの反復に属しているかを示すラベルが付けられています。ラベルの形式は、反復のタイプによって異なります。

- 固定フィールドの反復**:** ラベルは、フィールドの名前の後に等号、さらにその反復のフィールドの値が続きます。つまり、以下のようになります。

  field_name = field_value
- 分布パラメーターの反復**:** ラベルは、フィールドの名前の後にコロン、反復パラメーターの名前、等号、その反復のパラメーターの値が順に続きます。つまり、以下のようになります。

  field_name:parameter_name = parameter_value
- カテゴリー型分布または範囲型分布の分布パラメーターの反復**:** ラベルは、フィールドの名前の後にコロン、「反復」、反復数が順に続きます。つまり、以下のようになります。

  field_name: 反復 *iteration_number*

分布

任意のフィールドの確率分布を手動で指定するには、そのフィールドの「パラメーターの指定」ダイアログ・ボックスを開いて必要な分布を「分布」リストから選択し、分布パラメーターを「分布パラメーター」テーブルに入力します。特定の分布に関するいくつかの注意事項を以下に示します。

- カテゴリー: カテゴリー分布は、カテゴリーと呼ばれる、固定された数の数値がある入力フィールドを示します。各カテゴリーには確率が関連付けられ、すべてのカテゴリーの確率の合計は 1 となります。

注**:** 合計が 1 にならないカテゴリーの確率を指定すると、警告が表示されます。

- 負の **2** 項 **-** 失敗: 指定した数の成功が確認される前の、一連の試行回数のうちの失敗数の分布を示します。パラメーター *Threshold* は指定された成功数で、パラメーター *Probability* は、指定された試行回数のうち成功する確率です。
- 負の **2** 項 **-** 試行回数: 指定した数の成功が確認される前に必要な試行回数の分布を示します。パラメーター *Threshold* は指定された成功数で、パラメーター *Probability* は、指定された試行回数のうち成功する確率です。
- 範囲: この分布は一連の区間で構成されます。各区間には確率が割り当てられ、すべての区間の確率の合計は 1 となります。特定の区間内の値は、その区間に定義された一様分布から抽出されます。区間の指定は、最小値、最大値、関連付けられた確率を入力することにより行います。

  例えば、原料のコストが単位あたり $10 から $15 の範囲に収まる確率が 40%、$15 から $20 の範囲に収まる確率が 60% であるとします。この場合、「10 - 15」と「15 - 20」の 2 つの区間で構成される範囲の分布を使用してコストをモデル化し、最初の区間に関連付けられる確率を 0.4、2 番目の区間に関連付けられる確率を 0.6 に設定します。区間は連続させる必要はなく、重複しても構いません。例えば、区間として $10 から $15 と $20 から $25 や、$10 から $15 と $13 から $16 を指定することができます。
- **Weibull**: パラメーター *Location* は、分布の原点の位置を指定するオプションのロケーション・パラメーターです。

以下の表に、カスタムの分布の適合で使用できる分布と、パラメーターで許可される値を示します。これらの分布の一部は、シミュレーション適合ノードによって特定のストレージ・タイプに自動的に適合されない場合でも、それらのストレージ・タイプに対するカスタムの適合で使用することができます。

表 *12.* カスタムの適合で使用できる分布

分布	カスタムの適合でサポートされるストレージ・タイプ	パラメーター	パラメーターの制限	注
ベルヌーイ	整数、実数、日時	Probability	0 ≤ *Probability* ≤ 1	
ベータ	整数、実数、日時	Shape 1 Shape 2 Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Minimum と Maximum はオプションです。
2 項	整数、実数、日時	試行回数 (n) Probability Minimum Maximum	> 0、整数 0 ≤ *Probability* ≤ 1 < *Maximum* > *Minimum*	繰り返し回数は、整数でなければなりません。Minimum と Maximum はオプションです。
カテゴリー型	整数、実数、日時、文字列	カテゴリー名 (またはラベル)	0 ≤ *Value* ≤ 1	値はカテゴリーの確率です。値の合計は 1 でなければなりません。そうでない場合、警告が生成されます。

表 12. カスタムの適合で使用できる分布 (続き)

分布	カスタムの適合でサポートされるストレージ・タイプ	パラメーター	パラメーターの制限	注
ダイス	整数、文字列	Sides	2 ≤ *Sides* ≤ 20	各カテゴリー (サイド) の確率は 1/N として計算されます。N はサイド数です。確率を編集することはできません。
経験的	整数、実数、日時			経験分布を編集したり、経験分布をタイプとして選択したりすることはできません。 経験分布は、過去のデータが存在する場合のみ使用できます。
指数	整数、実数、日時	Scale Minimum Maximum	> 0 < *Maximum* > *Minimum*	Minimum と Maximum はオプションです。
固定	整数、実数、日時、文字列	Value		すべてのフィールドに「固定」分布を指定することはできません。生成されたデータ内のすべてのフィールドを固定する場合、ユーザー入力ノードに続けてバランス・ノードを使用します。
ガンマ	整数、実数、日時	Shape Scale Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Minimum と Maximum はオプションです。 分布では、形状パラメーター $\alpha = k$ および逆スケール パラメーター $\beta = 1/\theta$ の比率パラメーターを使用します。
対数正規	整数、実数、日時	Shape 1 Shape 2 Minimum Maximum	≥ 0 ≥ 0 < *Maximum* > *Minimum*	Minimum と Maximum はオプションです。
負の 2 項 - 失敗回数	整数、実数、日時	Threshold Probability Minimum Maximum	≥ 0 0 ≤ *Probability* ≤ 1 < *Maximum* > *Minimum*	Minimum と Maximum はオプションです。

表 12. カスタムの適合で使用できる分布 (続き)

分布	カスタムの適合でサポートされるストレージ・タイプ	パラメーター	パラメーターの制限	注
負の 2 項 - 試行回数	整数、実数、日時	Threshold Probability Minimum Maximum	≥ 0 0 ≤ *Probability* ≤ 1 < *Maximum* > *Minimum*	Minimum と Maximum はオプションです。
正常	整数、実数、日時	Mean Standard deviation Minimum Maximum	≥ 0 > 0 < *Maximum* > *Minimum*	Minimum と Maximum はオプションです。
ポアソン	整数、実数、日時	Mean Minimum Maximum	≥ 0 < *Maximum* > *Minimum*	Minimum と Maximum はオプションです。
範囲	整数、実数、日時	Begin(X) End(X) Probability(X)	0 ≤ *Value* ≤ 1	X は、各ビンのインデックスです。確率値の合計は 1 でなければなりません。
三角	整数、実数、日時	Mode Minimum Maximum	*Minimum* ≤ *Value* ≤ *Maximum* < *Maximum* > *Minimum*	
一様	整数、実数、日時	Minimum Maximum	< *Maximum* > *Minimum*	
ワイブル	整数、実数、日時	Rate Scale Location Minimum Maximum	> 0 > 0 ≥ 0 < *Maximum* > *Minimum*	Location、Maximum、Minimum はオプションです。

拡張インポート・ノード

拡張インポート・ノードを使用すると、R スクリプトまたは Python for Spark スクリプトを実行して、データをインポートできます。

拡張インポート・ノード - 「シンタックス」タブ

シンタックスのタイプ (**R** または **Python for Spark**) を選択します。次に、データをインポートするためのカスタム・スクリプトを入力するか、貼り付けます。シンタックスの準備ができたら、「実行」をクリックして、拡張インポート・ノードを実行できます。

拡張インポート・ノード - 「コンソール出力」タブ

「コンソール出力」タブには、「シンタックス」タブの R スクリプトまたは Python for Spark スクリプトが実行されたときに受信するすべての出力が含まれます (例えば、R スクリプトを使用する場合、「シンタックス」タブの「**R** シンタックス」フィールドにある R スクリプトが実行されたときに R コンソールから受信する出力が表示されます)。この出力には、R スクリプトまたは Python スクリプトの実行時に生成される R または Python のエラー・メッセージや警告が含まれる場合があります。出力は、主にスク

リプトをデバッグするために使用できます。「コンソール出力」タブには、「**R** シンタックス」フィールドまたは「**Python** シンタックス」フィールドのスクリプトも表示されます。

拡張インポート・スクリプトが実行されるたびに、R コンソールまたは Python for Spark から受信した出力で「コンソール出力」タブの内容が上書きされます。出力を編集することはできません。

フィールドのフィルタリングまたは名前の変更

ストリーム内の任意の場所でフィールドの名前変更またはフィールドを除外することができます。例えば、患者 (レコード レベル・データ) の カリウム値 (フィールド レベル・データ) を重要視していない場合、「K」 (カリウム値) フィールドを除外できます。個々のフィルター・ノード、または入力ノードか出力ノードの「フィルター」タブを使用して実行することができます。どのノードからアクセスしているかに関係なく、機能は同じです。

- 可変長ファイル、固定長ファイル、Statistics ファイル、XML、または拡張インポートなどの入力ノードから、IBM SPSS Modeler にデータを読み込むフィールドの名前を変更したり、フィルタリングを行うことができます。
- フィルター・ノードを使用して、ストリーム中の任意の場所でフィールドの名前を変更したり、フィルタリングすることができます。
- Statistics エクスポート・ノード、Statistics 変換ノード、Statistics モデル・ノードおよび Statistics 出力ノードから、IBM SPSS Statistics の命名規則に従ったフィールドをフィルタリングしたり、名前を変更することができます。 詳しくは、トピック 388 ページの『IBM SPSS Statistics 用のフィールドの名前変更またはフィルタリング』を参照してください。
- 上記のノードいずれかの「フィルター」タブを使用して、複数の回答セットを定義または編集することができます。 詳しくは、トピック 161 ページの『複数回答セット編集』を参照してください。
- 最後にフィルター・ノードを使用して、ある入力ノードから別のノードへフィールドをマップすることができます。

データ・ビュー・ノード

データ・ビュー・ノードは、IBM SPSS Collaboration and Deployment Services 分析データ ビューで定義されたデータをストリームに含めるために使用します。分析データ ビューは、予測モデルおよびビジネス・ルールで使用するエンティティーを記述するデータにアクセスするための構造を定義します。ビューは、データ構造を分析の物理データ・ソースに関連付けます。

予測分析には、予測を行うエンティティーに各行が対応するテーブルに編成されたデータが必要です。テーブルの各列は、エンティティーの測定可能な属性を表します。一部の属性は、別の属性の値を集計することで導き出すことができます。例えば、テーブルの行が顧客を表し、列が顧客の名前、性別、郵便番号、および過去 1 年間に 5 万円超の購入を行った回数に対応するとします。最後の列は、顧客注文履歴から導き出します。通常、注文履歴は関連した 1 つ以上のテーブルに格納されています。

予測分析プロセスでは、モデルのライフサイクル全体を通じてさまざまなデータのセットを使用する必要があります。予測モデルの初期の開発では、多くの場合は予測するイベントの既知の結果である履歴データを使用します。モデルの有効性および精度を評価するには、候補となるモデルを別のデータで検証します。モデルを検証した後、実動使用に展開して、バッチ処理で複数のエンティティーのスコアを生成するか、リアルタイム処理で単一のエンティティーを生成します。意思決定管理プロセスでモデルをビジネス・ルールと結合する場合は、シミュレーションしたデータを使用して結合の結果を検証します。ただし、使用するデータはモデル開発プロセスのステージによって異なりますが、各データ・セットには同じモデル属性群が存在する必要があります。属性群は変化しません。変化するのは、分析するデータ・レコードです。

ビューは以下の要素から構成され、各要素が予測分析の専門の要件に対応します。

- データにアクセスするための論理インターフェースを、関連したテーブルに編成された一連の属性として定義する、データ・ビュー・スキーマまたはデータ・モデル。モデルの属性は他の属性から導き出すことができます。
- データ・モデル属性に物理値を提供する、1 つ以上のデータ・アクセス計画。データ・モデルに使用可能なデータを制御するには、特定のアプリケーションに対してどのデータ・アクセス計画をアクティブにするかを指定します。

重要**:** データ・ビュー・ノードを使用するには、まず IBM SPSS Collaboration and Deployment Services Repository をサイトにインストールして設定する必要があります。ノードによって参照される分析データ ビューは、通常、IBM SPSS Deployment Manager を使用して作成され、リポジトリーに格納されます。

データ・ビュー・ノードのオプション設定

データ・ビュー・ノードのダイアログ・ボックスの「データ」タブにあるオプションは、IBM SPSS Collaboration and Deployment Services Repository から選択した分析データ ビューのデータ設定を指定するために使用します。

分析データ ビュー。省略符号ボタン (**...**) をクリックして「分析データ ビュー」を選択します。現在リポジトリー・サーバーに接続していない場合は、「リポジトリー: サーバー」ダイアログ・ボックスでサーバーの URL を指定して「**OK**」をクリックし、「リポジトリー: 資格情報」ダイアログ・ボックスで接続資格情報を指定します。リポジトリーへのログインおよびオブジェクトの取得について詳しくは、「IBM SPSS Modeler ユーザーズ・ガイド」を参照してください。

テーブル名。分析データ ビューのデータ・モデルからテーブルを選択します。データ・モデルの各テーブルは予測分析プロセスに関連する 1 つの概念 (エンティティー) を表します。テーブルのフィールドは、テーブルによって表されるエンティティーの属性に対応します。例えば、顧客からの注文を分析する場合は、顧客のテーブルと注文のテーブルがデータ・モデルに入ることになります。顧客のテーブルには、顧客 ID、年齢、性別、配偶者の有無、および在住国の属性が入ります。注文のテーブルには、注文 ID、注文された商品の数、総コスト、および注文を出した顧客の ID の属性が入ります。顧客 ID 属性は、顧客テーブルにある顧客を注文テーブルの注文に関連付けるために使用します。

データ アクセス計画。分析データ ビューからデータ・アクセス計画を選択します。データ・アクセス計画は、分析データ ビューのデータ・モデル・テーブルを物理データ・ソースに関連付けます。通常、分析データ ビューは複数のデータ・アクセス計画を含みます。使用中のデータ・アクセス計画を変更すると、ストリームで使用されるデータも変更されます。例えば、モデルに学習させるためのデータ・アクセス計画とモデルをテストするためのデータ・アクセス計画が分析データ ビューに含まれる場合は、使用するデータ・アクセス計画を変更することで、学習用データからテスト用データに切り替えることができます。

オプションの属性。分析データ ビューを使用するアプリケーションが特定の属性を必要としない場合は、オプションであるというマークをその属性に付けることができます。必須の属性と異なり、オプションの属性はヌル値を含む可能性があります。アプリケーションを調整して、オプション属性のヌル値の処理を組み込む必要がある場合があります。例えば、IBM Operational Decision Manager で作成したビジネス・ルールを呼び出すときには、 IBM Analytical Decision Management がルール・サービスを照会し、必要な入力を判別します。スコアリングされるレコードで、ルール・サービスに必要なフィールドのいずれかにヌル値が入っている場合は、そのルールが呼び出されず、ルールの出力フィールドにデフォルト値が入力されます。オプション・フィールドがヌル値を含む場合、ルールは呼び出されません。ルールは、ヌル値かどうかを検査して処理を制御することができます。

属性をオプションと指定するには、「オプションの属性」をクリックし、オプションにする属性を選択します。

フィールドに **XML** データを含める。データの各行について実行可能オブジェクト・モデルの XML データを含むフィールドを作成するには、このオプションを選択します。この情報は、データを IBM Operational Decision Manager で使用する場合に必要です。この新規フィールドの名前を指定します。

地理空間ソース・ノード

データ マイニング セッションにマップまたは地理空間データを導入するには、地理空間ソース・ノードを使用します。データのインポートは、以下の 2 つの方法のいずれかによって行えます。

- シェープファイル (.shp) を使用する方法
- マップ ファイルが存在する階層ファイル システムを含む ESRI サーバーに接続する方法

  注: パブリック・マップ・サービスにのみ接続できます。

時空間予測 (STP) モデルの予測には、マップまたは空間要素を含めることができます。このモデルの詳細については、「Modeler Modeling Nodes guide」(ModelerModelingNodes.pdf) の『Time Series Models』セクションに記載されているトピック『Spatio-Temporal Prediction modeling node』を参照してください。

地理空間入力ノードのオプションの設定

データ ソース タイプ データのインポートは、「シェープファイル」 (.shp) から行うか、「マップ サービス」に接続して行うことができます。

「シェープファイル」を使用する場合は、ファイル名とファイルのパスを入力するか、参照してファイルを選択します。ファイルは、ローカル・ディレクトリーに置くか、またはマップされたドライブからアクセスする必要があります。統一命名規則 (UNC) パスを使用してファイルにアクセスすることはできません。

注: 形状データには .shp ファイルと .dbf ファイルの両方が必要です。 この 2 つのファイルは、同じ名前を持ち、同じフォルダーに格納されている必要があります。.shp ファイルを選択すると、.dbf ファイルが自動的にインポートされます。また、形状データの座標系を指定する .prj ファイルが存在する場合もあります。

「マップ サービス」を使用する場合は、サービスの URL を入力して「接続」をクリックします。サービスに接続すると、サービス内の層が、ダイアログ ボックスの下部にある「使用可能なマップ」ペインにツリー構造で表示されます。このツリーを展開して、必要な層を選択します。

注: パブリック・マップ・サービスにのみ接続できます。

地理空間データの自動定義

デフォルトでは、SPSS Modeler は、可能な場合、正しいメタデータを使用してソース・ノードの地理空間データ フィールドをすべて自動的に定義します。メタデータに含まれる情報には、地理空間フィールドの尺度 (ポイントや多角形など)、フィールドで使用している座標系 (原点 (例えば緯度 0、経度 0) や測定単位などの詳細を含みます) などがあります。尺度について詳しくは、 148 ページの『地理空間のサブ尺度』を参照してください。

シェープファイルを構成する .shp ファイルと .dbf ファイルには、キーとして使用する共通の ID フィールドが入っています。例えば、.shp ファイルには国と国名フィールド (ID として使用) が含まれ、.dbf ファイルにはそれらの国に関する情報と国名 (ID としても使用) が含まれる場合があります。

注: 座標系が、SPSS Modeler のデフォルトの座標系と異なっている場合、必要な座標系を使用するために、データの再投影が必要になる場合があります。詳しくは、 196 ページの『再投影ノード』を参照してください。

共通のソース・ノード・タブ

次のオプションは、すべてのソース・ノードで、それぞれ適切なタブをクリックして指定することができます。

- 「データ」タブ: デフォルトのストレージ・タイプを変更するために使用されます。
- 「フィルター」タブ: データ・フィールドの除外や名前の変更を行うために使用されます。このタブは、フィルター・ノードと同じ機能を提供します。 詳しくは、トピック 159 ページの『フィルタリング・オプションの設定』を参照してください。
- 「データ型」タブ: 尺度の設定に使用します。このタブは、データ型ノードと同じ機能を提供します。
- 「注釈」タブ: すべてのノードで使用されるこのタブには、ノード名の変更、カスタム・ツールヒントの提供、および長い注釈の保存などのオプションが用意されています。

ソース・ノードの尺度の設定

フィールドのプロパティーは、ソース・ノードまたは、個別のデータ型ノードで指定できます。どちらのノードでも機能は同じです。次のプロパティが使用できます。

- フィールド IBM SPSS Modeler でデータの値とフィールド ラベルを指定するには、そのフィールド名をダブルクリックします。例えば、IBM SPSS Statistics からインポートされるフィールド メタデータを表示したり、データ型ノードで変更したりできます。同様に、フィールドの新しいラベルとそれらの値を作成できます。データ型ノードで指定したラベルは、「ストリームのプロパティー」ダイアログ・ボックスの選択内容に応じて、IBM SPSS Modeler 全体にわたって表示されます。
- 尺度 測定の尺度で、特定フィールドのデータの特性を記述するために使用します。フィールドの詳細がすべてわかっている場合には、完全にインスタンス化済みとされます。詳しくは、 145 ページの『尺度』を参照してください。

  注: フィールドの尺度はストレージ タイプとは異なります。ストレージ タイプは、データが文字列、整数、実数、日付、時間、タイムスタンプ、リストのいずれで格納されているかを示します。
- 値 この列を使用して、データセットからデータ値を読み込むオプションを指定したり、「指定」オプションを使用して別のダイアログ ボックスで尺度および値を指定したりすることができます。値を読み込まないでフィールドを渡すこともできます。詳しくは、 150 ページの『データ値』を参照してください。

  注: 対応する「フィールド」エントリーがリストを含む場合、この列のセルは修正できません。
- 欠損値 フィールドの欠損値の処理方法を指定するために使用します。詳しくは、 155 ページの『欠損値の定義』を参照してください。

  注: 対応する「フィールド」エントリーがリストを含む場合、この列のセルは修正できません。
- 検査 この列には、フィールドの値が指定された値または範囲内に収まっているかどうかを検査するオプションを設定できます。詳しくは、 155 ページの『データ型の値の検査』を参照してください。

注: 対応する「フィールド」エントリーがリストを含む場合、この列のセルは修正できません。

- 役割 フィールドがマシン学習プロセスの「入力」(予測フィールド) または「対象」(予測されるフィールド) のどちらになるかをモデル作成ノードに指示するために使用します。「両方」 および 「なし」 も役割として利用できます。さらに、レコードを学習用、検定用、および検証用の独立したサンプルに分割するために使用されるフィールドを示す 「データ区分」 も利用できます。値 「分割」 は各モデルがフィールドの可能な値それぞれに作成されるように設定します。 詳しくは、 156 ページの『フィールドの役割の設定』を参照してください。

詳しくは、トピック 144 ページの『データ型ノード』を参照してください。

ソース・ノードでインスタンス化する時期

フィールドのデータ・ストレージと値を学習するには、2 種類の方法があります。このインスタンス化は、最初に IBM SPSS Modeler にデータを取り込んだ時に入力ノードで、またはデータ・ストリームにデータ型ノードを挿入した時に行われます。

ソース・ノードでのインスタンス化は、次のような場合に役立ちます。

- データ・セットが小さい場合。
- 式ビルダーを使用して新規フィールドの作成を計画している場合 (インスタンス化により、式ビルダーでフィールドの値を利用できるようになります)。

一般的に、データ・セットがさほど大きくなく、後でストリームにフィールドを追加する予定がない場合は、ソース・ノードでインスタンス化するのが便利です。

ソース・ノードからのフィールドのフィルタリング

「ソース・ノード」ダイアログ・ボックスの「フィルター」タブで、初期のデータ調査結果に基づいて下流の操作から一部のフィールドを除外することができます。このことは、例えば、データ中に重複するフィールドがある場合や、すでにデータをよく理解しており、不要なフィールドを除外したい場合などに役立ちます。または、後の時点で別個のフィルター・ノードをストリームに追加することもできます。どちらの場合でも、機能は同じです。 詳しくは、トピック 159 ページの『フィルタリング・オプションの設定』を参照してください。

第 3 章 レコード設定ノード

レコード設定の概要

レコード設定ノードは、レコード レベルでデータを変更するために使用されます。これらの操作は、特定のビジネス・ニーズに合わせてデータを調整できるので、データ・マイニングのデータの理解およびデータの準備フェーズ中に重要です。

例えば、データ検査ノード (出力パレット) を使用したデータ検査の結果に基づいて、過去 3 か月の顧客購入レコードを結合するように決定できます。レコード結合ノードを使用して、*Customer ID* (顧客 ID) などのキー・フィールドの値を基準にしてレコードを結合できます。または、Web サイトのアクセス件数に関する情報を含むデータベースに 100 万件以上のデータが蓄積されていて管理不能になっていることもあります。その場合、サンプリング・ノードを使用して、モデリング用のデータのサブセットを選択することもできます。

「レコード設定」パレットには、次のノードがあります。

条件抽出ノードで、特定の条件に基づいて、データ・ストリームからレコードのサブセットを選択したり破棄したりできます。例えば、特定の営業地域に関連するレコードを選択できます。

サンプル・ノードでは、レコードのサブセットを選択します。層化サンプル、クラスター・サンプル、非無作為 (構造化) サンプルなど、さまざまなサンプルの種類がサポートされています。サンプリングは、パフォーマンスの向上、および分析のための関連するレコードまたはトランザクションのグループの選択に役に立ちます。

バランス・ノードで、データ・セットが指定した条件に合うように、データ・セットの不均衡を修正します。バランス式で、指定した比率によって条件が真 (true) の場合に、レコードの比率を調整します。

レコード集計ノードで、一連の入力レコードを要約集計された出力レコードに置き換えます。

リーセンシ、フリクエンシ、マネタリー (RFM) のレコード集計ノードを使用すると、顧客の過去のトランザクション・データを取得、未使用のデータを削除、残りのトランザクション・データをすべて単一行に結合することができます。これにより、最後のトランザクションの時期、トランザクション数、これらのトランザクションの合計金額が一覧表示されます。

ソート・ノードで、1 つまたは複数のフィールド値に基づいて、レコードを昇順または降順にソートします。

レコード結合ノードは、複数の入力レコードを取得し、入力フィールドの全部または一部を含む 1 つの出力レコードを作成します。この機能は、内部顧客データと購入人口データのような、異なるソースからのデータを結合する場合に役立ちます。

レコード追加ノードで、レコードのセットを連結します。レコード追加ノードは、構造が似ていながらデータが異なるデータ・セットを組み合せる場合に役立ちます。

重複レコード・ノードで、重複レコードを削除します。その場合、最初の重複するレコードをデータ・ストリームに渡すか、または、最初のレコードを破棄して、その後の重複レコードをデータ・ストリームに渡します。

ストリーミング時系列ノードは、1 つのステップで時系列モデルを作成してスコアリングします。ローカル環境または分散環境のどちらのデータでもこのノードを使用できます。分散環境の場合、IBM SPSS Analytic Server の機能を活用できます。

Space-Time-Box (STB) は、Geohash の空間的な場所を拡張したものです。具体的には、STB は英数字の文字列で、空間および時間を規則的に分割した領域です。

ストリーミング TCM ノードは、1 つのステップで時間的因果モデルを作成してスコアリングします。

CPLEX の最適化ノードにより、OPL (Optimization Programming Language) モデル・ファイルを介した複雑な数学 (CPLEX) ベースの最適化の機能が提供されます。この機能は IBM Analytical Decision Management 製品で使用可能ですが、IBM Analytical Decision Management の必要なしに、SPSS Modeler でも CPLEX ノードを使用できるようになりました。

CPLEX の最適化および OPL について詳しくは、IBM Analytical Decision Management の資料を参照してください。

「レコード設定」パレットの多くのノードでは、CLEM 式を使用する必要があります。CLEM に精通している場合は、フィールドに式を入力できます。ただし、すべての式フィールドには CLEM 式ビルダーを開くボタンが用意されていて、これを使用すると、自動的に式が作成されるようになります。

図 *1.* 「*Clem* 式ビルダー」ボタン

条件抽出ノード

条件抽出ノードを使用すると、BP (血圧) = "HIGH" (高) などの特定の条件に基づいて、データ・ストリームからレコードのサブセットを選択したり破棄したりすることができます。

モード: 条件を満たすレコードを、データ・ストリームに入れるか、データ・ストリームから除外するかを指定します。

- 含める: 選択条件を満たすレコードを入れる場合に選択します。
- 破棄: 選択条件を満たすレコードを除外する場合に選択します。

条件: 各レコードのテストに使用される選択条件が表示されます。CLEM 式を使用して指定します。ウィンドウに式を入力するか、またはウィンドウの右側にある計算機 (Clem 式ビルダー) ボタンをクリックして表示される Clem 式ビルダーを使用します。

次のような条件に基づいてレコードを破棄する場合、

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

デフォルトでは、条件抽出ノードがすべての選択フィールドに Null 値を持つレコードも破棄します。こうしたレコードが破棄されないよう、次の条件を元の条件に追加します。

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

条件抽出ノードは、レコードの一部を選択するためにも使用されます。通常、この操作にはサンプリング・ノードなど、別のノードを使用します。ただし、用意されているパラメーターよりも複雑な条件を指定する場合は、条件抽出ノードを使用して独自の条件を作成します。例えば、次のような条件を作成できます。

```
BP = "高" および無作為 (10) <= 4
```

この条件では、高血圧を示すレコードの約 40% が選択され、詳細な分析のために下流に渡されます。

サンプル・ノード

サンプル・ノードを使用して、分析のためにレコードのサブセットを選択、または破棄するレコードの割合を指定することができます。層化サンプル、クラスター・サンプル、非無作為 (構造化) サンプルなど、さまざまなサンプルの種類がサポートされています。サンプリングを使用する理由は、次のとおりです。

- データのサブセットのモデルを推定してパフォーマンスを向上する。サンプルから推定されたモデルは、完全なデータ・セットから取得したモデルと同じくらい正確で、向上したパフォーマンスによってこれまで試すことがなかったさまざま方法を試すことができる場合、より正確になります。
- オンライン ショッピングのカートのすべてのアイテムを選択または特定の隣接地域のすべての資産を選択するなど、分析のために換算するレコードまたはトランザクションのグループを選択する。
- 品質評価、不正防止、またはセキュリティーの対象となる無作為検査の単位またはケースを識別する。

注 : 検証の目的でデータを学習サンプルおよび検定サンプルに分割する場合、データ区分ノードを代わりに使用することができます。 詳しくは、トピック 185 ページの『データ区分ノード』を参照してください。

サンプルの種類

クラスター化サンプル**:** 個々の単位ではない、サンプル グループまたはクラスター。例えば、生徒ごとに 1 つのレコードを持つデータ・ファイルがあるとします。学校ごとにクラスター化し、標本サイズが 50% の場合、学校の 50% が選択され、選択されたそれぞれの学校からすべての生徒が取得されます。選択されない学校の生徒は却下されます。平均的には、およそ 50% の生徒が抽出されることが期待されますが、学校の規模が異なるため、割合は正確でない場合があります。同様に、トランザクション ID によってショッピング カートのアイテムをクラスター化し、選択されたトランザクションのすべてのアイテムが含まれていることを確認します。町ごとの資産をクラスター化する例については、*complexsample_property.str* のサンプル・ストリームを参照してください。

層化サンプル**:** 母集団の重複しないサブグループまたは階層内のサンプルを独立して選択します。例えば、男性および女性を等しい割合でサンプリングされ、または都市部の人口の中ですべての地域または社会経済的グループが表示されるようにすることができます。また、各階層の異なる標本サイズを指定することもできます (例えば、元のデータの 1 つのグループが実際より低く評価された場合)。町ごとの資産を層化する例については、*complexsample_property.str* のサンプル・ストリームを参照してください。

体系的または **n** 件ごとのサンプリング**:** 無作為な選択が難しい場合に、系統的に (固定間隔で) または順序に従って、単位のサンプリングを行うことができます。

抽出重み付け**:** 重みのサンプリングは、複雑なサンプルを引き出す際に自動的に計算され、サンプルされた各単位が元のデータに表示される「度数」にほとんど対応します。そのため、サンプルの重みの合計で、元のデータのサイズを推定する必要があります。

サンプリング・フレーム

サンプリング・フレームによって、サンプルまたは調査に含まれるケースの可能性のあるソースを定義します。例えば生産ラインから外れる項目のサンプリングを行う場合、母集団の各単一メンバーを識別し、サンプルにメンバーの 1 つを含めることができます。可能性のあるすべてのケースにアクセスできない場合がよくあります。例えば、選挙が実行された後まで、選挙で誰が投票するのかを確認できません。この場合、一部の人々が投票せず、名簿を確認した時点で登録されていない人々が投票する場合もありますが、サンプリング・フレームとして選挙人名簿を使用します。サンプリング・フレームに含まれない人は、サンプリングされる可能性はありません。サンプリング・フレームが評価しようとしている母集団に本質的に十分近いかどうかは、それぞれの実際のケースで処理する必要のある問題です。

サンプル・ノードのオプション

用件に応じてシンプルまたは複雑な方法を選択できます。

シンプルなサンプリングのオプション

シンプルな方法を使用すると、レコードの無作為な割合を選択、連続するレコードを選択または *n* 件ごとのレコードを選択することができます。

モード**:** 次のモードに対して、レコードを渡す (含める) か、または破棄 (除外) するかを選択します。

- サンプルを含める。データ・ストリームの選択されたレコードを含め、他のレコードをすべて破棄します。例えば、モードを 「サンプルを含める」 に設定し、「**n** 件ごと」 に 5 を指定した場合、最大標本サイズになるまで 5 件ごとに 1 つのレコードがデータ・ストリームに追加され、データ・セットが元のサイズの 5 分の 1 のサイズとなります。このモードはデータのサンプリングする際のデフォルト・モードで、複雑な方法を使用する場合に唯一使用できるモードです。
- サンプルを破棄。選択されたレコードを破棄し、他のすべてのレコードを含めます。例えば、「サンプルを破棄」 モードで 「**n** 件ごと」 を 5 に設定すると、5 件ごとに 1 つのレコードが破棄 (除外) されます。このモードはシンプルな方法でのみ使用できます。

サンプル。次のいずれかのサンプリング手法を選択します。

- 初めの **n** 件。連続したデータ・サンプリングを使用する場合に選択します。例えば、サンプルの最大サイズが 10000 に設定されている場合、最初の 10,000 件のレコードが選択されます。
- **n** 件ごと。*n* 件のレコードごとにデータを通過させるか破棄することによってデータをサンプリングする場合に選択します。例えば、*n* が 5 に設定されている場合、5 件ごとのレコードが選択されます。
- 無作為 **%**。データを任意のパーセンテージでサンプリングする場合に選択します。例えば、20% に設定すると、選択したモードに従って、データの 20% がデータ・ストリームに渡されるか、または破棄されます。このフィールドに、サンプリングのパーセンテージを指定します。「ランダム シードの設定」 から、シードの値を指定することもできます。

  ブロック レベルのサンプリングを使用 **(**データベース内のみ**)**。このオプションは、Oracle データベースまたは IBM Db2 データベースでデータベース内マイニングを実行するときに無作為パーセント抽出を選択した場合にのみ有効です。こうした環境では、ブロックレベルのサンプリングがより効果的です。

  注**:** 同じランダム・サンプル設定を実行するとしても、そのたびに返される行の数は正確ではありません。これは、各入力レコードに、サンプルに組み込まれる `N/100` の確率があり (`N` はノードに指定する **Random %**)、確率は独立しているため、結果が正確に `N%` にはならないためです。

最大サンプル数。サンプルに含めるレコードの最大数を指定します。「サンプルを含める」 および 「初めの **n** 件」 が選択されている場合、このオプションは無効になります。また、「無作為 **%**」 オプションが使用されている場合、この設定によって特定のレコードが選択されません。例えば、データ・セットに 1 千万件のレコードがあり、最大サンプル数 300 万件のレコードという設定でレコードの 50% を選択すると、最初の 600 万件のレコードだけに 50% の選択の可能性があり、残りの 400 万件のレコードからは選択されないということになります。この制限を回避するためには、複雑なサンプリング方法を選択し、クラスターまたは階層変数を指定せずに 300 万件のレコードから無作為のサンプルを要求します。

複雑なサンプリングのオプション

複雑なサンプリングのオプションを使用すると、クラスター化サンプル、層化サンプル、重み付けされたサンプルを他のオプションとともに、サンプルをより詳細に制御することができます。

クラスターと階層。必要に応じて、クラスター・フィールド、層化フィールド、および入力重みフィールドを指定することができます。詳しくは、トピック『クラスターと階層の設定』を参照してください。

サンプル・タイプ。

- 無作為。各階層内で無作為にクラスターまたはレコードを選択します。
- 体系的。固定された間隔でレコードを選択します。このオプションは、ランダム・シードに応じて最初のレコードの位置が変化することを除き、「*n* 件ごと」の方法と同様に動作します。*n* の値は、標本サイズまたは割合に基づいて自動的に決定されます。

サンプル単位。基本的なサンプル単位として割合または度数を選択することができます。

サンプル サイズ**:** 次のいくつかの方法でサンプル・サイズを指定することができます。

- 固定**:** 全体のサンプル・サイズを度数または割合として指定することができます。
- ユーザー設定**:** 各サブグループまたは階層のサンプル・サイズを指定することができます。このオプションは、層化フィールドが「クラスター」および「層化」サブ ダイアログ・ボックスで指定されている場合にのみ使用できます。
- 変数。ユーザーは、各サブグループまたは階層の標本サイズを定義するフィールドを指定することができます。このフィールドには、特定の階層内の各レコードの同じ値が含まれています。例えば、サンプルが地域ごとに層化されている場合、地域 = *Surrey* のすべてのレコードは同じ値を持つ必要があります。フィールドは数値型で、その値は選択されたサンプル単位に一致する必要があります。単位が割合の場合、値は 0 より大きく 1 より小さくなります。単位が度数の場合、最小値は 1 です。

階層ごとの最小サンプル。最小レコード数を指定します (クラスター・フィールドが指定されている場合は、最小クラスター数が指定されます)。

階層ごとの最大サンプル。レコードまたはクラスターの最大数を指定します。クラスターまたは層化フィールドを指定せずにこのオプションを選択した場合、指定されたサイズの無作為または体系的サンプルが選択されます。

ランダム・シードの設定**:** 無作為なパーセンテージに基づいてレコードをサンプリングまたはデータ区分している場合、このオプションで、別のセッションに同じ結果を複製できるようになります。乱数ジェネレータに使用される開始値を指定することで、ノードが実行されるごとに毎回同じレコードが割り当てられることが保証されます。自動的に無作為な値を生成するには、希望のシード値を入力するか、「生成」 ボタンを入力します。このオプションが選択されないと、ノードが実行されるごとに異なるサンプルが生成されます。

注 : データベースから読まれるレコードに「ランダム シードの設定」オプションを使用する場合は、ノードが実行されるごとに同じ結果を保証するために、サンプリングに先行して、ソート・ノードが必要になる可能性があります。この理由は、ランダム シードがレコードの順序に依存しているためです。各レコードがリレーショナル・データベース内で同じ位置に留まる保証はありません。詳しくは、トピック 88 ページの『ソート・ノード』を参照してください。

クラスターと階層の設定

「クラスターと階層化」ダイアログ・ボックスを使用すると、複雑なサンプルを抽出する際に、クラスター・フィールド、階レイヤー フィールド、重みフィールドを選択することができます。

クラスター: クラスター・レコードに対して使用するカテゴリー・フィールドを指定します。レコードは、含まれるクラスターと含まれないクラスターがある所属クラスターに基づいてサンプリングされます。ただし、指定のクラスターのレコードが含まれる場合、すべてのクラスターが含まれます。例えば、ショッピン

グ カート内の商品の関連を分析する場合、トランザクション ID によってアイテムをクラスター化し、選択されたトランザクションのすべてのアイテムが含まれていることを確認します。レコードをサンプリングすると、一緒に販売されるアイテムに関する情報が破棄されますが、トランザクションをサンプリングすると、選択したトランザクションのすべてのレコードを保持することができます。

階層化する**:** レコードを層化するために使用するカテゴリー・フィールドを指定し、サンプルが母集団の重複していないサブグループまたは階層内で独立して選択されるようにします。例えば、性別によって層化された 50% のサンプルを選択する場合、男女それぞれに 2 つの 50% サンプルが選択されます。例えば、階層は社会経済的なグループ、職業のカテゴリー、年齢グループ、民族グループで、対象となるサブグループの適切な標本サイズを確認することができます。元のデータ・セットで、男性の 3 倍の女性がいる場合、この比率は各グループとは別にサンプリングすることによって保存されます。複数の層化フィールドを指定することもできます (例えば、地域内の製品ラインのサンプリング、またはその逆)。

注 : 欠損値 (ヌルまたはシステム欠損値、空白文字列、空白文字、空欄またはユーザー定義の欠損値) のあるフィールドによって層化する場合、階層のカスタム標本サイズを指定することはできません。欠損値または空白値を含むフィールドによって層化する際にユーザー定義の標本サイズを使用する場合、上流でそれらの値を入力する必要があります。

入力重みのみ使用**:** サンプリングの前にレコードを重み付けするために使用するフィールドを指定します。例えば、重みフィールドに 1 ～ 5 の値が含まれる場合、5 の重み付けをされたレコードは 5 倍選択される可能性があります。このフィールドの値は、ノードに生成された最後の出力の重みに上書きされます (次の段落を参照)。

新規出力重み**:** 入力重みフィールドが指定されていない場合、最後の重みが書き込まれるフィールド名を指定します。(入力重みが指定されている場合、その値は前述のとおり最後の重みに置き換えられますが、作成される個別の出力重みフィールドはありません。)出力重みの値は、元のデータのサンプリングされた各レコードによって示されるレコード数を示します。重みの値の合計によって、標本サイズの推定が得られます。例えば、無作為の 10% のサンプルが選択された場合、出力重みはすべてのレコードに対して 10 となり、サンプリングされた各レコードが元のデータの 10 件のレコードを表すことを示します。層化サンプルまたは重み付けされたサンプルでは、出力重みの値は各階層のサンプルの割合によって異なります。

コメント

- クラスター化されたサンプルは、サンプリングする割合の詳細なリストを取得できないが、特定のグループまたはクラスターの詳細なリストを取得できる場合に役に立ちます。無作為サンプルによって連絡が取れない被験者のリストを作成する場合にも使用されます。例えば、国内のすべての地域に散在する農家を選択するよりも、1 つの地域のすべての農家を訪問することが簡単です。
- クラスター・フィールドおよび層化フィールドの両方を選択して、各階層内のクラスターを独立してサンプリングすることができます。例えば、地域によって層化し、地域内の町ごとにクラスター化した資産価値をサンプリングすることができます。これによって、町の独立したサンプルを各地域内から引き出すことができます。サンプルに含まれる町と含まれない町がある場合、含まれる町については、町内のすべての資産が含まれます。
- 各クラスター内から単位の無作為サンプルを選択するために、2 つのサンプル・ノードを結びつけることができます。例えば、前述のように地域によって層化された町を最初にサンプリングすることができます。その後 2 番目のサンプル・ノードを適用し、層化フィールドとして町を選択すると、各町ごとのレコードの割合をサンプリングすることができます。
- フィールドを組み合わせてクラスターを一意に識別する必要がある場合、フィールド作成ノードを使用して新規フィールドを生成することができます。例えば、複数の店舗でトランザクションの同じナンバリング・システムを使用している場合、店舗とトランザクション ID を連結する新規フィールドを取得できます。

階層のサンプル・サイズ

層化サンプルを引き出す場合、デフォルトのオプションは、各階層のレコードまたはクラスターの同じ割合をサンプリングすることです。あるグループが 3 の要素で別のグループを上回る場合、通常サンプルの同じ比率をサンプルに保持します。ただしそうでない場合は、各階層ごとに標本サイズを個別に指定することができます。

「階層の標本サイズ」ダイアログ・ボックスには層化フィールドの各値が一覧表示され、その階層のデフォルトを上書きすることができます。複数の層化フィールドを選択する場合、値の考えられる組み合わせがすべて表示され、例えば各市内の各民族グループまたは各地域内の町ごとのサイズを指定することができます。サンプル・ノードの現在の設定で決められたように、サイズは割合または度数として指定されています。

階層の標本サイズを指定するには

1. サンプル・ノードで 「複雑」 を選択し、1 つまたは複数の層化フィールドを選択します。 詳しくは、トピック 80 ページの『クラスターと階層の設定』を参照してください。
2. 「ユーザー設定」 を選択し、「サイズを指定」 を選択します。
3. 「階層の標本サイズ」ダイアログ・ボックスで、左下の 「値の読み込み」 ボタンをクリックして表示を指定します。上流のソースまたはデータ型ノードの値をインスタンス化する必要がある場合があります。 詳しくは、トピック 149 ページの『インスタンス化とは ?』を参照してください。
4. 行をクリックして、該当する階層のデフォルト・サイズを上書きします。

標本サイズの注意

異なる階層に異なる分散が含まれている場合、例えば標本サイズを標準偏差に比例させる場合に、ユーザー設定の標本サイズが役に立ちます。(階層内のケースがより異なる場合、より多くのケースをサンプリングして標本サンプルを取得する必要があります。)または階層が小さい場合、より高いサンプルの割合を使用して、最小観察数が含まれていることの確認が必要な場合があります。

注 : 欠損値 (ヌルまたはシステム欠損値、空白文字列、空白文字、空欄またはユーザー定義の欠損値) のあるフィールドによって層化する場合、階層のカスタム標本サイズを指定することはできません。欠損値または空白値を含むフィールドによって層化する際にユーザー定義の標本サイズを使用する場合、上流でそれらの値を入力する必要があります。

バランス・ノード

バランス・ノードを使用して、データ・セットの不均衡を修正し、指定したテスト基準を満たすことができます。例えば、データ・セットに 低 と 高 という 2 つの値しかなく、ケースの 90% が 低 で、残りの 10% が 高 である場合を考えてみましょう。このようにデータが偏っている場合、低の結果だけが学習され、より少ない高の結果は無視される傾向があるため、多くのモデリング手法で問題となります。低と高がほぼ同数で、データのバランスがとれていれば、モデルで 2 つのグループを区別するパターンを発見できる可能性が高くなります。このような場合にバランス・ノードを使用して、低 の結果を含むケースを減らすバランス式を作成します。

バランスの調整で実際に行われるのは、指定の条件に従ったレコードの複製と破棄です。適用する条件がないレコードは、すべて通過します。この処理はレコードの複製と破棄から成り立っているため、下流の操作では元のデータ・シーケンスが失われます。データ・ストリームにバランス・ノードを追加する前に、シーケンスに関連する値を必ず作成しておきます。

注 : バランス・ノードは、分布図やヒストグラムから自動的に作成することができます。例えば、データのバランスを調整して、棒グラフで表示されているように、カテゴリー型フィールドのすべてのカテゴリー全体の等しい割合を表示することができます。

例: RFM ストリームを構築して、以前のマーケティング・キャンペーンに肯定的に応答して最近の顧客を識別する場合、販売会社のマーケティング部門ではバランス・ノードを使用して、データ内の真 (true) と偽 (false) の回答間の差異のバランスを調整します。

バランス・ノードのオプション設定

レコード バランス式**:** 現在のバランス式を一覧表示します。各式には、ソフトウェアに「条件が真 (true) の場合に、指定した比率だけレコードを増やす」ことを指示する比率と条件が含まれています。1.0 より低い比率を指定している場合は、指定されたレコードの比率が減少することを表しています。例えば、処方薬が薬品 Y のレコード数を減らしたい場合、比率 0.7 で条件が `Drug = "薬品 Y"`のバランス式を作成することができます。この式では、薬品 Y が処方薬であるレコードの数が、下流のすべての操作で 70% に減らされます。

注 : 減少用のバランス比率は、小数点以下第 4 位まで指定できます。比率を 0.0001 未満に設定すると、結果が正しく算出されずエラーが発生します。

- テキスト・フィールドの右にあるボタンをクリックすると、条件が作成されます。このボタンにより、新しい条件を入力するための空の行が挿入されます。条件の CLEM 式を作成するには、Clem 式ビルダー・ボタンをクリックします。
- 式を削除するには、赤い削除ボタンをクリックします。
- 式をソートするには、赤い上矢印および下矢印ボタンを使用します。

学習データのみをバランス**:** ストリーム内にデータ区分フィールドがある場合、このオプションによって学習用データ区分のデータのみのバランスを調整します。特に、不均衡な検定用区分または検証用区分を要求する、調整された傾向スコアを生成する場合に役立ちます。 ストリーム内にデータ区分フィールドがない場合 (または複数のデータ区分フィールドが指定されている場合)、このオプションは無視され、すべてのデータのバランスが調整されます。

レコード集計ノード

レコード集計は、データ・セットのサイズを減らすために頻繁に用いられるデータ準備作業です。レコード集計を行う前に、データのクリーニングを、特に欠損値に注目して行う必要があります。レコード集計を実行すると、欠損値に関する潜在的な有益情報が失われてしまう可能性があります。

レコード集計ノードを使用すると、一連の入力レコードを要約 (集計された出力レコード) に置き換えることができます。例えば、次の表に示すような一連の入力売上レコードがあるとします。

表 *13.* 売上レコード入力の例

Age	**Sex (**性別**)**	地域	**Branch**	**Sales (**販売額**)**
23	M	S	8	4
45	M	S	16	4
37	M	S	8	5
30	M	S	5	7
44	M	N	4	9
25	M	N	2	11

表 13. 売上レコード入力の例 (続き)

Age	**Sex** (性別)	地域	**Branch**	**Sales** (販売額)
29	F	S	16	6
41	F	N	4	8
23	F	N	6	2
45	F	N	4	5
33	F	N	6	10

上記のレコードを、*Sex* と *Region* をキー・フィールドにして集計することができます。次に、「平均値」モードで*Age* フィールドを集計し、「合計」モードで *Sales* フィールドを集計します。「集計ノード」ダイアログ・ボックスで「フィールドにレコード度数を含める」を選択します。集計の出力は次の表のようになります。

表 14. 集計したレコードの例

Age (平均年齢)	**Sex** (性別)	地域	**Sales** (販売額)	レコード件数
35.5	F	N	25	4
29	F	S	6	1
34.5	M	N	20	2
33.75	M	S	20	4

例えば、このことから、北部地域の 4 名の女性販売スタッフの平均年齢が 35.5 歳で、合計販売金額が 25 単位です。

注 : 集計モードを指定しないと、*Branch* などのフィールドは自動的に無視されます。

レコード集計ノードのオプション設定

「レコード集計」ノードで以下を指定します。

- レコード集計のカテゴリーとして使用する 1 つまたは複数のキー・フィールド。
- レコード集計値を計算する 1 つまたは複数の集計フィールド。
- 各レコード集計フィールドに出力する 1 つまたは複数のレコード集計モード (集計の種類)。

新しく追加されたフィールドに使用するデフォルトの集計モードを指定することも、式 (数式に似ている) を使用して集約をカテゴリー化することもできます。

パフォーマンス上、集計操作は、並列処理を有効にすると、有利になる可能性があります。

キー・フィールド。集計にカテゴリーとして使用できるフィールドが一覧表示されます。連続型 (数値型) フィールドとカテゴリー・フィールドの両方がキーとして使用できます。複数のフィールドを選択した場合は、値が組み合されて、レコードを集計するためのキー値が生成されます。集計レコードは、それぞれ一意のキー・フィールドに対して 1 つずつ生成されます。例えば、キー・フィールドが *Sex* と *Region* の場合、一意な *M* と *F* の、および地域 *N* と *S* のそれぞれの組み合わせに対して集計レコードが作成されます (4 つの一意な組み合わせ)。キー・フィールドを追加するには、ウィンドウの右側にあるフィールド・ピッカー・ボタンをクリックします。

ダイアログ・ボックスの残りの部分は、基本集計と集計式という 2 つの主な領域に分かれています。

基本集計

集計フィールド。選択されたレコード集計のモードのほか、集計される値のフィールドを表示します。リストにフィールドを追加するには、右側にあるフィールド・ピッカー・ボタンを使用します。利用できる集計関数を次に示します。

注: 数値型以外のフィールドの適用できないフィールドがあります (例えば、日付/時刻フィールドの 「合計」)。選択した集計フィールドで使用できないモードが無効になります。

- 合計:キー・フィールドの各組み合せの合計値を返す場合に選択します。合計は、欠損値のないすべてのケースに対する変数の値の合計です。
- 平均値: キー・フィールドの各組み合せの平均値を返す場合に選択します。平均値は、中心傾向の尺度であり、算術平均です (ケース数で割った合計)。
- 最小値: キー・フィールドの各組み合わせの最小値を返す場合に選択します。
- 最大値: キー・フィールドの各組み合わせの最大値を返す場合に選択します。
- 標準偏差:キー・フィールドの各組み合わせの標準偏差を返す場合に選択します。標準偏差平均の周りの散らばり度です。変動測定の平方根に等しくなります。
- 中央値: キー・フィールドの各組み合せの中央値を返す場合に選択します。中央値は、外れ値に対して敏感でない、中心化傾向の測定値です。それに対して平均値は、いくつかの極端に大きい、または小さい値に影響されます。50 番目のパーセンタイルまたは 2 番目の四分位でもあります。
- カウント: キー・フィールドの各組み合わせの非ヌル値のカウントを返す場合に選択します。
- 分散:キー・フィールドの各組み合わせの分散値を返す場合に選択します。分散は、平均値のまわりの値の散らばりの程度。平均値からの偏差の平方和を、有効観測値の合計数から 1 を引いたもので割って求めます。
- 第一四分位:キー・フィールドの各組み合わせの第一四分位 (25 番目のパーセンタイル) を返す場合に選択します。
- 第三四分位:キー・フィールドの各組み合わせの第三四分位 (75 番目のパーセンタイル) を返す場合に選択します。

注: 集計ノードを含むストリームを実行する場合、Oracle データベースに SQL をプッシュバックするときに第一四分位数と第三四分位数について返される値が、ネイティブ・モードで返される値と異なる場合があります。

デフォルト・モード。新しく追加したフィールドに対して、デフォルトで使用する集計モードを指定します。同じ集計モードを頻繁に使用しているような場合は、ここでそれらのモードを選択し、右側にある「すべてに適用」ボタンをクリックすると、選択したモードが上記のリストに表示されているすべてのフィールドに適用されます。

新規フィールド名拡張子。重複する集計フィールドに対して、「1」や「new」などの接頭辞や接尾辞を追加する場合に選択します。例えば、接尾辞オプションを選択して「1」を指定すると、*Age* フィールドに対する最小値の集計の結果、フィールド名 *Age_Min_1* が追加されます。注 : *_Min* や *Max_* などの集計拡張子は実行された集計の種類を表し、自動的に新規フィールドに追加されます。付ける拡張子の種類に応じて、「接尾辞」 または 「接頭辞」 を選択してください。

フィールドにレコード度数を含める。各出力レコードに追加のフィールド (デフォルトでは *Record_Count*) を含める場合に選択します。このフィールドは、各集計レコードを作成するために集計された入力レコード数を示します。このフィールド名を自分で指定するには、その名前を編集フィールドに入力してください。

注**:** 集計の実行時にシステムのヌル値は除外されますが、レコードの度数には含まれます。一方、空白値は、集計レコードとレコード度数の両方に含まれます。空白値を除外するには、置換ノードを使用して空白値をヌル値に置き換えます。また、条件抽出ノードを使用して、空白値を削除することもできます。

集計式

式は、値、フィールド名、演算子、および関数から作成される数式と同じようなものです。一度に 1 つのレコードに対して機能する関数とは異なり、集計式はレコードのグループ、セット、または集合に対して働きます。

注**:** 集計式を作成できるのは、ストリームにデータベース接続が含まれている場合 (データベース・ソース・ノードによって) のみです。

新しい式は、派生フィールドとして作成されます。式を作成するには、Clem 式ビルダーから使用可能なデータベース集計 関数を使用します。

Clem 式ビルダーについて詳しくは、「IBM SPSS Modeler User's Guide」(ModelerUsersGuide.pdf) を参照してください。

集計式はキー・フィールド別にグループ化されるため、キー・フィールドと、作成する集計式との間に接続があることに注意してください。

集計結果を評価する集計式は、有効な集計式です。以下に 2 つの有効な集計式の例と規則を示します。

- スカラー関数を使用して複数の集計関数を結合し、単一の集計結果を生成できます。以下に例を示します。

  ```
  max(C01) - min(C01)
  ```

- 集計関数は、複数のスカラー関数の結果に対して働きます。以下に例を示します。

  ```
  sum (C01*C01)
  ```

集計の最適化設定

「最適化」タブで以下を指定します。

連続キー。同じキー値を持つすべてのレコードが入力にグループ化されている場合 (例えば、入力がキー・フィールドにソートされる場合)、このオプションを選択します。このオプションを選択すると、パフォーマンスが向上します。

中央値および四分位の近似値を許可。Analytic Server でデータを処理するとき、順序統計 (中央値、第 1 四分位数、第 3 四分位数) は現在サポートされていません。Analytic Server を使用する場合は、代わりにこのチェック ボックスを選択してこれらの統計の近似値を使用できます。この近似値は、データを分割してからビン全体の分布に基づいて統計の推定値を計算することで算出されます。デフォルトでは、このオプションにチェックマークは付いていません。

ビン数。「中央値および四分位の近似値を許可」チェック ボックスを選択した場合にのみ使用できます。統計値の推定時に使用するビンの数を選択します。ビン数は「最大誤差 **%**」に影響します。デフォルトのビン数は 1000 です。これは、0.1 パーセントの範囲の最大誤差に対応しています。

RFM レコード集計ノード

リーセンシ、フリクエンシ、マネタリー (RFM) のレコード集計ノードを使用すると、一意の顧客 ID をキーとして使用し、顧客の過去のトランザクション・データを取得、未使用のデータを除去、残りのトランザクション・データをすべて単一行に結合することができます。これにより、最後のトランザクションの時期 (リーセンシ)、トランザクション数 (フリクエンシ)、これらの取引の合計金額 (マネタリー) が一覧表示されます。

レコード集計を行う前には、データのクリーニングを、特に欠損値に注目して行う必要があります。

RFM レコード集計ノードを使用してデータを識別および変換すると、RFM 分析ノードを使用してより詳細な分析を実行することができます。 詳しくは、トピック 181 ページの『RFM 分析ノード』を参照してください。

データ・ファイルが RFM レコード集計ノードによって実行されると、ファイルには対象値が含まれません。そのため、データを C5.0 または CHAID などのモデル作成ノードによる詳細な分析の入力として使用する前に、(例えば顧客 ID と一致させることによって) 他の顧客のデータと結合する必要があります。詳しくは、トピック 89 ページの『レコード結合ノード』を参照してください。

IBM SPSS Modeler の RFM レコード集計ノードおよび RFM 分析ノードを設定して独立した分割を使用します。最新性、頻度、金額値の各尺度のデータを、これらの値および尺度に関係なくランク付けし、分割します。

RFM レコード集計ノードのオプション設定

RFM レコード集計ノードの「設定」タブには次のフィールドがあります。

リーセンシ基準日: トランザクションのリーセンシが計算される日付を指定します。入力する 固定日付またはシステムで設定された 今日の日付のいずれかです。今日の日付はデフォルトで入力され、ノードが実行されると自動的に更新されます。

注: 固定日付の表示は、ロケールによって異なる場合があります。例えば、値 `2007-8-10` がストリームに `Fri Aug 10 00:00:00 CST 2007` として保管されている場合、これは、時間帯「UTC+8」の日時になります。ただし、時間帯「UTC-8」では、`Thu Aug 9 12:00:00 EDT 2007` として表示されます。

連続する **ID** データ・ストリーム中で同じ ID を持つすべてのレコードが一緒に表示されるようにデータをソートしている場合、このオプションを選択すると処理を高速化することができます。データがあらかじめソートされていない場合 (またはわからない場合) は、このオプションは選択しないでください。この場合、ノードが自動的にデータをソートします。

ID 顧客およびトランザクションを識別するために使用するフィールドを選択します。選択できるフィールドを表示するには、右側のフィールド・ピッカー・ボタンを使用します。

日付: リーセンシを計算するために使用される日付フィールドを選択します。選択できるフィールドを表示するには、右側のフィールド・ピッカー・ボタンを使用します。

入力として使用するには、適切な形式の日付、時間、またはタイムスタンプのストレージのフィールドが必要です。例えば、*Jan 2007*、*Feb 2007* などのような値を持つ文字列フィールドがある場合、置換ノードおよび `to_date()` 関数を使用してこれらのフィールドを日付フィールドに変換することができます。 詳しくは、トピック 170 ページの『置換ノードを使ったストレージの変換』を参照してください。

値 顧客のトランザクションの全体の金額を計算するために使用するフィールドを選択します。選択できるフィールドを表示するには、右側のフィールド・ピッカー・ボタンを使用します。注 : これは、数値である必要があります。

新規フィールド名拡張子「12_month」などの接尾辞または接頭辞を、新しく生成されたリーセンシ、フリクエンシ、マネタリー・フィールドに適用します。付ける拡張子の種類に応じて、「接尾辞」 または 「接頭辞」 を選択してください。例えば、複数の期間を調べる場合に役立ちます。

次の値以下のレコードを破棄 必要な場合、RFM の合計を計算する場合に使用されないトランザクションの詳細の最小値を指定することができます。値の単位は、選択された 「値」 フィールドに関連します。

最近のトランザクションのみを含める 大規模なデータベースを分析する場合、最新のレコードのみが使用されるよう指定することができます。次のように特定の日付または最新の期間内のいずれかに記録された日付を選択できます。

- 次以降のトランザクション データ: トランザクションの日付を指定します。この日付以降のレコードが分析に含まれます。
- 次の期間のトランザクション データ レコードが分析に含まれる後の 「リーセンシ基準日」 の日付からさかのぼった期間の数および種類 (日、週、月または年数) を指定します。

2 番目のリーセント トランザクション日付を保存 各顧客の 2 番目に最近のトランザクションの日付を知りたい場合は、このオプションを選択します。また、「**3** 番目のリーセント トランザクション日付を保存」 チェック・ボックスも選択できます。例えば、かなり前に多くの取引を行っているが最近は 1 回の取引しかない顧客を識別するために役に立ちます。

ソート・ノード

ソート・ノードを使用すれば、レコードを 1 つまたは複数のフィールド値に基づいて昇順または降順にソートすることができます。ソート・ノードは、例えば最も一般的なデータ値のレコードを表示および選択するために頻繁に使用されます。一般的に、まずレコード集計ノードを使用してデータを集計し、次にソート・ノードを使用して集計データをレコード度数の降順にソートします。この結果をテーブルに表示してデータを探索し、上位 10 位までの顧客のレコードを選択するような決定を下します。

ソート・ノードの「設定」タブには次のフィールドがあります。

ソート項目: ソート キーとして使用するすべてのフィールドがテーブル内に表示されます。数値のキー・フィールドが、ソートに最も適しています。

- このリストにフィールドを追加するには、右側にあるフィールド・ピッカー・ボタンを使用します。
- 並び順を選択するには、テーブルの 「順序」列にある 「昇順」 または 「降順」の矢印をクリックします。
- フィールドを削除するには、赤い削除ボタンをクリックします。
- 式をソートするには、赤い上矢印および下矢印ボタンを使用します。

デフォルトのソート順。新しいフィールドが上に追加された場合にデフォルトで使用するソート順序として、「昇順」または「降順」を選択します。

注: モデルス トリームの下流に重複レコード ノードがある場合、ソート ノードは適用されません。重複レコード ノードについて詳しくは、 98 ページの『重複レコード・ノード』を参照してください。

ソートの最適化設定

すでにいくつかのキー・フィールドでソートされているデータを処理している場合は、残りのデータをシステムが効率よくソートできるように、すでにソートされたフィールドを指定できます。例えば、*Age* (降順) と *Drug* (昇順) でソートするときに、データがすでに *Age* (降順) でソート済みであることがわかっているとします。

データは事前ソート済み **:** データがすでに 1 つ以上のフィールドでソート済みかどうかを指定します。

既存のソート順を指定**:** すでにソートされているフィールドを指定します。「フィールドの選択」ダイアログ・ボックスを使用して、フィールドをリストに追加します。「順序」 列で、各フィールドが昇順でソートされているか、または降順でソートされているかを指定します。複数のフィールドを選択する場合は、正しいソート順で一覧になっていることを確認してください。リストの右にある矢印を使用して、フィールドを正しい順序で配置します。正しい既存のソート順の指定に誤りがあると、ストリームを実行するときに、ソートが指定どおりではないレコード番号を示した、エラーが表示されます。

注 : 並列処理を有効にすると、ソートの速度が上がる場合があります。

レコード結合ノード

レコード結合ノードでは、複数の入力レコードから、入力フィールドの全部または一部を含む 1 つの出力レコードが作成されます。この機能は、内部顧客データと購入人口データのような、異なるソースからのデータを結合する場合に役立ちます。以下の方法で、データを結合できます。

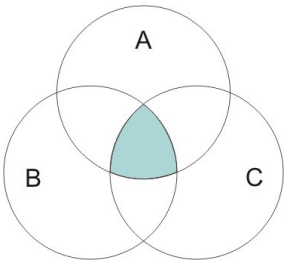
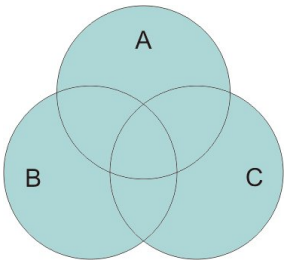
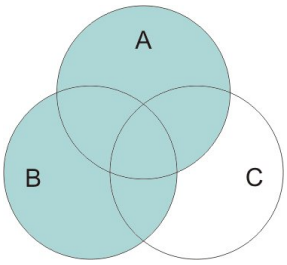
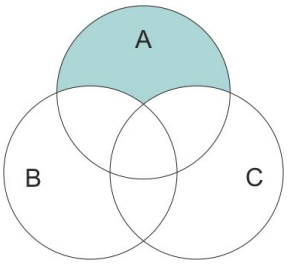
- 順序による結合では、最も小さいデータ・ソース中のデータがなくなるまで、入力順序にしたがってすべてのソースから対応するレコードを結合します。ソート・ノードを使用してデータをソートしている場合に、このオプションが重要になります。
- キー・フィールドを使った結合では、*Customer ID* (顧客 ID) などのキー・フィールドを使用して、あるデータ・ソース中のレコードと他のデータ・ソース中のレコードとの結合方法を指定します。内部結合、完全外部結合、部分外部結合、および逆結合など、さまざまな種類の結合を利用できます。 詳しくは、トピック『結合の種類』を参照してください。
- 条件による結合では、結合を行うために満たす必要のある条件を指定できます。ノードで直接条件を指定するか、Clem 式ビルダーを使用して条件を作成できます。
- 「ランク付けされた条件」による結合は左側の外部結合であり、結合を行うために満たす必要のある条件と、低いものから高いものの順にソートするランク付け式を指定します。多くの場合は地理空間データの結合に使用し、ノードで直接条件を指定するか、Clem 式ビルダーを使用して条件を作成できます。

結合の種類

データ結合にキー・フィールドを使用する場合、まずどのレコードを除外して、どのレコードを対象にするかを検討することをお勧めします。後述するように、さまざまな種類の結合手段があります。

基本的な結合の種類としては、内部結合と外部結合があります。これらの方法は、*Customer ID* などのキー・フィールドの共通する値に基づいて、関連データ・セットからテーブルを併合するために頻繁に使用されます。内部結合によって、制限のない併合が行われ、完全なレコードのみが含まれるデータ・セットが出力されます。外部結合の場合も結合データからの完全なレコードが含まれますが、それ以外に 1 つまたは複数の入力テーブルから固有のデータを入れることもできます。

利用できる各種の結合手段の詳細は、後述します。

	内部結合では、キー・フィールドの値がすべての入力テーブルで共通のレコードだけが含まれます。つまり、出力データ・セットには、一致しないレコードは含まれません。
	完全外部結合では、入力テーブルからの一致するレコードと一致しないレコードの両方のレコード (すべてのレコード) が含まれます。左外部結合および右外部結合は、部分外部結合と呼ばれています。
	部分外部結合では、キー・フィールドを使ったすべての一致するレコード、および特定のテーブルからの一致しないレコードが含まれます。(または、別な方法では、いくつかのテーブルからのすべてのレコードと、そのほかのテーブルからの一致したレコードのみ。)外部結合に入れるテーブル (ここの A や B など) は、「レコード結合」タブの「データの選択」ボタンを使用して選択することができます。2 つのテーブルだけを結合する場合、部分結合は左外部結合または右外部結合と呼ばれることもあります。IBM SPSS Modeler では、3 つ以上のテーブルを結合することもできるため、ここでは部分外部結合と呼んでいます。
	逆結合では、最初の入力テーブルの一致しないレコードだけが含まれます (ここではテーブル A)。この結合では、内部結合とは反対に、出力データ・セットに完全なレコードは含まれません。

例えば、あるデータ・セット内の農場についての情報があり、農業関連の保険金請求が別のデータ・セットにある場合に、「レコード結合」オプションを使用して、最初のソースのレコードを 2 番目のソースに照合できます。

農場サンプル内の顧客が保険金請求をファイリングしているかどうかを判断するには、内部結合オプションを使用して、すべての ID が 2 つのサンプルで一致する箇所を示すリストを返します。

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalue
1	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0...
2	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0...
3	id620	name620	north	880.000	74.000	6.000	426988.000	rapeseed	arable_de	118885.0...

図 2. 内部結合のサンプル出力

完全外部結合オプションを使用すると、入力テーブルから一致するレコードと一致しないレコードの両方が返されます。システム欠損値 ($null$) が、不完全な値に対して使用されます。

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalu
1	id601	$null$	$null$	$null$	$null$	$null$	$null$	$null$	decomm...	74703.1(
2	id602	name602	north	1780.000	42.000	9.000	734118.000	maize	$null$	$nul
3	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0
4	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0
5	id606	$null$	$null$	$null$	$null$	$null$	$null$	$null$	arable_de	122135.0

図 *3.* 完全外部結合のサンプル出力

部分外部結合では、指定されたテーブルから一致しないレコードと同様に、キー・フィールドを使用して一致したすべてのレコードが含まれます。テーブルには、最初のデータ・セットから一致したレコードと同様に、ID フィールドから一致したすべてのレコードが表示されます。

	id	claimtype	claimvalue	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	$null$	$null$	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id604	decomm...	281082.0...	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes
3	id605	decomm...	122006.0...	name605	north	1700.000	46.000	8.000	621148.000	wheat
4	id607	$null$	$null$	name607	southeast	1820.000	29.000	6.000	211605.000	maize
5	id608	$null$	$null$	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
6	id609	$null$	$null$	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
7	id615	$null$	$null$	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
8	id618	$null$	$null$	name618	southeast	1180.000	98.000	3.000	368646.000	maize

図 *4.* 部分外部結合のサンプル出力

逆結合オプションを使用する場合は、最初の入力テーブルで一致しないレコードのみが返されます。

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id607	name607	southeast	1820.000	29.000	6.000	211605.000	maize
3	id608	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
4	id609	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
5	id615	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
6	id618	name618	southeast	1180.000	98.000	3.000	368646.000	maize
7	id619	name619	north	840.000	64.000	8.000	457552.000	potatoes

図 *5.* 逆結合のサンプル出力

結合方法とキーの指定

レコード結合ノードの「レコード結合」タブには次のフィールドがあります。

レコード結合方法 レコードの結合に使用する方法を選択します。「キー」または「条件」を選択すると、ダイアログボックスの下半分が有効になります。

- 順序 各入力からの *n* 番目のレコードを結合して *n* 番目の出力レコードを生成するように順序でレコードを結合します。一致する入力レコードがなくなると、それ以上の出力レコードは生成されません。つまり、作成されるレコード数は一番小さいデータ・セットのレコード数と等しくなります。
- キー トランザクション *ID* などのキー フィールドを使用して、キー フィールドの値が同じであるレコードを結合します。これは、データベースの「等結合」と同じ処理です。キーの値が複数あるような場合は、可能なすべての組み合わせが返されます。例えば、同じキーフィールドの値 *A* を持つ複数のレコードが、それぞれ別のフィールドでは値 *B*、*C*、および *D* を持つ場合、結合後のフィールドでは *A* と *B*、*A* と *C* 、および *A* と *D* のように、個別のレコードを組み合わせた結果が生成されます。

注 : キーによるレコード結合では、ヌル値は同一とみなされず、結合されません。

- 条件 結合の条件を指定するには、このオプションを使用します。詳しくは、 93 ページの『結合の条件の指定』を参照してください。
- ランク付けされた条件 このオプションを使用して、1 次データセットとすべての 2 次データセットの各行ペアを結合するかどうかを指定します。ランク付け式を使用して、複数の一致を低いものから高いものの順にソートします。詳しくは、 93 ページの『結合のためのランク付けされた条件の指定』を参照してください。

キーの候補 すべての入力データ ソースの完全に一致するフィールド名のフィールドだけがリストされます。このリストからフィールドを選択して矢印ボタンをクリックすると、レコードを結合するためのキー・フィールドにそのフィールドが追加されます。複数のキーフィールドを使用できます。一致しない入力フィールドの名前を変更するには、フィルター ノードまたはソース・ノードの「フィルター」タブを使用します。

結合キー すべての入力データ・ソースから、キー・フィールドの値に基づいたレコードの結合に使用するすべてのフィールドが表示されます。リストからキーを削除するには、該当するキーを選択して矢印ボタンをクリックし、そのキーを「キーの候補」リストに戻します。複数のキー・フィールドが選択されている場合は、次のオプションが有効になります。

重複するキー・フィールドをまとめる 上で複数のキー・フィールドが選択されている場合に、その名前を持つ出力フィールドを 1 つだけにします。以前のバージョンの IBM SPSS Modeler からストリームをインポートした場合を除いて、このオプションはデフォルトで有効になります。このオプションの選択を解除した場合、「レコード結合ノード」ダイアログ・ボックスの「フィルター」タブを使用して、重複するキー・フィールドの名前を変更するか、または除外する必要があります。

一致レコードだけを含める **(**内部結合**)** 完全なレコードだけを結合する場合に選択します。

一致レコードと不一致レコードを含める **(**完全外部結合**)**「完全外部結合」を実行する場合に選択します。キー・フィールドの値がすべての入力テーブルに存在しない場合、不完全なレコードはそのまま保持されます。未定義値 **(**$null$**)** は、キー・フィールドに追加され、出力レコードに含まれます。

一致レコードおよび選択した不一致レコードを含める **(**部分外部結合**)** サブダイアログ・ボックスで選択したテーブルに対して、「部分外部結合」を実施する場合に選択します。不完全なレコードを保持するテーブルを指定するには、「データの選択」 をクリックします。

最初のデータ・セット中の他と一致しないレコードを含める **(**逆結合**)** 最初のデータ・セットの不一致レコードだけを下流に渡す、逆結合を実施する場合に選択します。「入力」タブの矢印を使用して、入力データ・セットの順序を指定することができます。この結合では、出力データ・セットに完全なレコードは含まれません。 詳しくは、 89 ページの『結合の種類』を参照してください。

部分結合のデータの選択

部分外部結合の場合、不完全なレコードを保持するテーブルを選択する必要があります。例えば、顧客テーブルからのすべてのレコードを保持しながら、住宅ローン・テーブル中の一致するレコードだけを保持することができます。

「外部結合」列**:** 「外部結合」列では、全体をそのまま含めるデータ・セットを選択します。部分結合の場合、ここで選択したデータ・セットに対して、重複レコードや不完全レコードが保持されます。 詳しくは、トピック 89 ページの『結合の種類』を参照してください。

結合の条件の指定

結合の方法を 「条件」 に設定して、結合を行うために満たす必要のある条件を指定できます。

「条件」フィールドに条件を直接入力、またはフィールドの右側の計算機の記号をクリックして Clem 式ビルダーを使用し、条件を作成することができます。

重複したフィールド名にタグを追加して結合の競合を回避 結合する複数のデータセットに同じフィールド名が存在する場合は、このチェック ボックスを選択して、フィールドの列見出しの先頭に個別の接頭辞タグを追加します。例えば、*Name* というフィールドが 2 つ存在する場合の結合結果には、*1_Name* と *2_Name* が含まれることになります。データ ソースのタグを名前変更した場合は、番号付き接頭辞タグの代わりに新しい名前が使用されます。このチェック ボックスを選択しないと、重複する名前がデータに存在する場合に、チェック ボックスの右側に警告が表示されます。

結合のためのランク付けされた条件の指定

ランク付けされた条件の結合は、条件による左側外部結合と考えることができます。結合の左側は、各レコードがイベントになっている 1 次データセットです。例えば、犯罪データのパターンを検出するために使用するモデルでは、1 次データセットの各レコードが犯罪とその関連情報 (場所や種類など) になります。この例の場合、右側には、関連した地理空間データセットを含めることになります。

結合では、結合条件とランク付け式の両方を使用します。結合条件では、*within* や *close_to* などの地理空間関数を使用できます。結合中には、右側のデータセットにあるすべてのフィールドが左側のデータセットに追加されますが、複数一致の場合はリスト フィールドが生成されます。以下に例を示します。

- 左側: 犯罪データ
- 右側: 地域のデータセットと道路のデータセット
- 結合条件: 犯罪データ *within* 地域かつ *close_to* 道路。加えて、*close_to* と見なす場合の定義。

この例で、3 本の道路から所定の *close_to* 距離以内で犯罪が発生し、かつ返される一致件数が 3 以上に設定されている場合は、その 3 本の道路すべてがリスト項目として返されます。

結合の方法を「ランク付けされた条件」に設定することで、結合を行うために満たす必要のある 1 つ以上の条件を指定できます。

1 次データセット 結合用の 1 次データセットを選択します。他のすべてのデータセットからのフィールドが、選択したデータセットに追加されます。これは外部結合の左側と考えることができます。

1 次データセットを選択すると、レコード結合ノードに接続する他のすべての入力データセットが自動的に「結合」テーブルにリストされます。

重複したフィールド名にタグを追加して結合の競合を回避 結合する複数のデータセットに同じフィールド名が存在する場合は、このチェック ボックスを選択して、フィールドの列見出しの先頭に個別の接頭辞タグを追加します。例えば、*Name* というフィールドが 2 つ存在する場合の結合結果には、*1_Name* と *2_Name* が含まれることになります。データ ソースのタグを名前変更した場合は、番号付き接頭辞タグの代わりに新しい名前が使用されます。このチェック ボックスを選択しないと、重複する名前がデータに存在する場合に、チェック ボックスの右側に警告が表示されます。

結合

データセット

レコード結合ノードへの入力として接続する 2 次データセットの名前が表示されます。複数の 2 次データセットが存在するとき、デフォルトでは、レコード結合ノードに接続された順序でリストされます。

結合条件

1 次データセットを含むテーブルの各データセットを結合するための固有の条件を入力します。セルに条件を直接入力、またはセルの右側の計算機の記号をクリックして Clem 式ビルダーを使用し、条件を作成することができます。 例えば、地理空間述部を使用して、あるデータセットから取得した犯罪データを別のデータセットの地域データに格納する結合条件を作成することができます。デフォルトの結合条件は、以下のリストに示すように地理空間の尺度によって決まります。

- ポイント、行ストリング、複数点、複数行ストリング - デフォルト条件は *close_to*。
- 多角形、多角形群 - デフォルト条件は *within*。

これらの尺度について詳しくは、 148 ページの『地理空間のサブ尺度』を参照してください。

異なる種類の複数の地理空間フィールドがデータセットに含まれる場合に使用されるデフォルト条件は、以下の降順で、データで検出された最初の尺度に応じて決定されます。

- ポイント
- 行ストリング
- 多角形

注: 地理空間データ フィールドが 2 次データベースに存在する場合は、デフォルトのみが使用可能です。

ランク付け式

データセットの結合をランク付けするための式を指定します。この式は、複数の一致をランク付けの基準による順序にソートするために使用します。セルに条件を直接入力、またはセルの右側の計算機の記号をクリックして Clem 式ビルダーを使用し、条件を作成することができます。

Clem 式ビルダーには距離と領域のデフォルトのランク付け式が用意されており、いずれのランク付けも低から高の順です (例えば、距離の一致度が最も高いものが最小の値になります)。距離によるランク付けの例として、1 次データセットに犯罪とその関連位置の情報が入っており、他のデータセットにはそれぞれ位置情報付きの目標物が入っているとすると、犯罪と目標物の間の距離をランク付けの基準として使用できます。デフォルトのランク付け式は、以下のリストに示すように地理空間の尺度によって決まります。

- ポイント、行ストリング、複数点、複数行ストリング - デフォルトの式は *distance* です。
- 多角形、多角形群 - デフォルトの式は *area* です。

注: 地理空間データ フィールドが 2 次データベースに存在する場合は、デフォルトのみが使用可能です。

一致の数

条件およびランク付け式に基づいて取得する一致の件数を指定します。以下のリストに示すように、デフォルトの一致件数は 2 次データセットの地理空間尺度によって決まります。ただし、セルをダブルクリックして任意の値を入力することができます。このときの最大値は 100 です。

- ポイント、行ストリング、複数点、複数行ストリング - デフォルト値は 3。
- 多角形、多角形群 - デフォルト値は 1。

- 地理空間フィールドを含まないデータセット - デフォルト値は 1。

例えば、*close_to* の「結合条件」と *distance* の「ランク付け式」に基づいて結合をセットアップした場合は、2 次データセットから 1 次データセットの各レコードへの一致のうち、上位の (最も近い順に) 3 件の一致が結果リスト フィールドの値として返されます。

レコード結合ノードからのフィールドのフィルタリング

レコード結合ノードでは、複数のデータ・ソースを結合した結果として発生する重複フィールドをフィルタリングしたり、名前を変更することができます。フィルタリング・オプションを選択するには、ダイアログ・ボックスの 「フィルター」 タブをクリックします。

ここのオプションは、フィルター・ノードのオプションとほとんど変わりありません。ただし、ここに記載されていなくても、「フィルター」メニューで利用できるオプションもあります。 詳しくは、トピック 159 ページの『フィールドのフィルタリングまたは名前の変更』を参照してください。

フィールド: 現在接続しているデータ・ソースからの入力フィールドを表示します。

タグ: データ・ソース・リンクに関連付けられたタグ名 (または数字) が表示されます。このレコード結合ノードへのアクティブなリンクを変更するには、「入力」 タブをクリックします。

ソース・ノード: データを結合するソース・ノードを表示します。

接続済みノード: レコード結合ノードに接続しているノードのノード名を表示します。複雑なデータ・マイニング作業では、しばしば複数の結合または追加操作が行われ、同じソース・ノードが含まれることがあります。接続されたノード名を表示することによって、この問題を解消することができます。

フィルター: 入力フィールドと出力フィールドの間の現在の接続を表示します。アクティブな接続は、正常な矢印で表示されます。赤い X が付けられている接続は、そのフィールドがフィルタリングされていることを表します。

フィールド: 結合後または追加後に出力フィールドをリストします。重複するフィールド名は赤で表示されます。重複するフィールドを無効にするには、上の「フィルター」フィールドをクリックします。

現在のフィールドを表示: キー・フィールドとして使用されているフィールドの情報を表示します。

未使用のフィールド設定を表示: 現在使用されていないフィールドの情報を表示します。

入力順序とタグの設定

「レコード結合ノード」ダイアログ・ボックスまたは「レコード追加ノード」ダイアログ・ボックスの「入力」タブを使用して、データ・ソースの入力順序を指定したり、各ソースのタグ名を変更することができます。

入力データ・セットのタグと順序: 完全なレコードだけを結合または追加する場合に選択します。

- **タグ:**各入力データ・ソースの現在のタグ名を表示します。タグ名またはタグは、結合または追加操作のデータ・リンクを一意に識別するための手段を提供します。例えば、複数の水道管を流れている水が、ある場所で 1 本の水道管に合流するような状況を想像してください。IBM SPSS Modeler 内のデータも同じように流れていきます。また、合流点では、さまざまなデータ・ソース間でしばしば複雑なやり取りが行われます。タグは、レコード結合ノードやレコード追加ノードで入力 (先ほどの例の「水道管」にあたる) を管理し、ノードが保存されたり切断された場合でも、リンクをそのまま保持して簡単に識別できるようにするための手段を提供します。

レコード結合ノードまたはレコード追加ノードにデータ・ソースを追加した場合、ノードの接続順序を示す数字を使用して、デフォルトのタグが自動的に作成されます。この順序は、入力または出力データ・セット中のフィールドの順序とは関係ありません。デフォルトのタグを変更するには、「タグ」 列に新しい名前を入力します。

- ソース・ノード**:** データを結合するソース・ノードを表示します。
- 接続済みノード**:** レコード結合ノードまたはレコード追加ノードに接続しているノードのノード名を表示します。複雑なデータ・マイニング作業では、しばしば複数の結合操作が行われ、同じソース・ノードが含まれることがあります。接続されたノード名を表示することによって、この問題を解消することができます。
- フィールド**:** 各データ・ソース中のフィールド数を表示します。

現在のタグを表示**:** レコード結合ノードまたはレコード追加ノードで現在使用されているタグを表示する場合に選択します。現在のタグは、データが流れているノードへのリンクを識別します。水道管にたとえると、現在のタグは現在水が流れ込んでいる水道管を示します。

未使用のタグ設定の表示**:** 以前レコード結合ノードまたはレコード追加ノードへの接続に使用されたことがあり、現在データ・ソースに接続されていないタグ、またはリンクを表示する場合に選択します。先ほどの例にたとえると、これは給水システム内にまだ現存しているけれども、水が流れていない水道管を表しています。これらの「水道管」を新しいソースに接続することも、削除することもできます。ノードから未使用のタグを削除するには、「消去」 をクリックします。この操作を行うと、未使用のタグがすべて削除されます。

レコード結合の最適化設定

一定の状況でデータをより効率的にレコード結合できるように、システムには 2 つのオプションが用意されています。これらのオプションによって、1 つの入力データ・セットがほかのデータ・セットより著しく大きい場合や、データが、レコード結合に使用するキー・フィールドのすべて、または一部によってすでにソート済みの場合に、レコード結合を最適化できるようになります。

注**:** このタブでの最適化は、IBM SPSS Modeler ネイティブ ノード実行のみ、つまりレコード結合ノードが SQL にプッシュバックしない場合にのみ適用されます。最適化の設定は SQL 生成に影響しません。

入力データ・セットが比較的大きい **:** 入力データ・セットの 1 つがほかのデータ・セットよりもかなり大きいことを示すために選択します。システムにより小さいほうのデータ・セットがメモリー内にキャッシュされ、その後、大きいデータ・セットをキャッシングもソートもしないで処理することで、レコード結合が実行されます。通常、このような結合は、星状のスキーマや類似した設計を使用して設計されたデータとともに使用します。星状のスキーマには、例えばトランザクション形式のデータ内のような、共有データの大きな中央テーブルがあります。このオプションを選択する場合は、「選択」 をクリックして大きなデータ・セットを指定します。ただ *1* つ の大きなデータ・セットのみを選択できることに注意してください。次の表に、この方法を使用して最適化できる結合を要約します。

表 *15.* 結合の最適化の要約：

結合の種類	大規模入力データ・セット用に最適化できるか**?**
内部	はい
部分	大規模データ・セット内に不完全レコードがない場合に、「はい」。
完全	いいえ
逆結合	大規模データ・セットが最初の入力の場合に、「はい」。

すべての入力はキー・フィールド基準ですでにソートされている: 入力データが、レコード結合に使用するキー・フィールドのうちの 1 つ以上ですでにソートされていることを示すために選択します。入力データ・セットが すべて ソート済みであることを確認します。

既存のソート順を指定: すでにソートされているフィールドを指定します。「フィールドの選択」ダイアログ・ボックスを使用して、フィールドをリストに追加します。レコード結合に使用されている (「レコード結合」タブで指定された) キー・フィールドの中からのみ、選択できます。「順序」 列で、各フィールドが昇順でソートされているか、または降順でソートされているかを指定します。複数のフィールドを選択する場合は、正しいソート順で一覧になっていることを確認してください。リストの右にある矢印を使用して、フィールドを正しい順序で配置します。正しい既存のソート順の指定に誤りがあると、ストリームを実行するときに、ソートが指定どおりではないレコード番号を示した、エラーが表示されます。

データベースが使用する照合方式の大文字と小文字の区別によっては、入力がデータベースによってソートされている場合最適化が正しく機能しないことがあります。例えば、一方の入力が大文字と小文字を識別し、もう一方が大文字と小文字を区別しない 2 つの入力がある場合、ソートの結果が異なる可能性があります。最適化を結合した場合、レコードがソートされた順で処理される場合があります。その結果、異なる照会方法で入力がソートされた場合、結合ノードはエラーを報告し、ソートが一貫しないレコード番号を表示します。すべての入力が 1 つのソースである場合、または相互に包括的な照合によってソートされている場合、レコードは正常に結合できます。

注 : レコード結合の速度は、並行処理を有効にすると、有利になる可能性があります。

レコード追加ノード

レコード追加ノードを使用すると、一連のレコードを連結することができます。異なるソースからのレコードを結合するレコード結合ノードとは異なり、レコード追加ノードでは、1 つのソースからレコードがなくなるまですべてのレコードが読み込まれて下流に渡されます。次に、最初の入力または主入力と同じデータ構造 (レコード数やフィールド数など) を使用して、次のソースからレコードが読み込まれます。主ソースのフィールド数が他の入力ソースのフィールド数より多い場合は、不完全な値に対してはシステム ヌル文字列 (`$null$`) が使用されます。

レコード追加ノードは、構造が似ていてもデータが異なるデータ・セットを組み合わせる場合に役立ちます。例えば、3 月の売り上げデータ・ファイルと 4 月の売り上げデータ・ファイルのように、期間の異なるファイルにトランザクション・データが保存されている場合を想定します。これらのファイルの構造は同じである (同じフィールドが同じ順序で並んでいる) と仮定すれば、レコード追加ノードを使用して両方のファイルを 1 つの大きなファイルに結合することができます。このファイルを、後ほど分析に利用します。

注 : ファイルを追加するには、フィールドの尺度が同じでなければなりません。例えば、名義型フィールドを尺度が連続型のフィールドの追加することができません。

追加オプションの設定

フィールド一致基準: 追加するフィールドを一致させるための方法を選択します。

- 位置: メイン・データ・ソース中のフィールドの位置を基準にしてデータ・セットを追加する場合に選択します。この方法を使用する場合、正しく追加を行うためには、あらかじめデータをソートしておく必要があります。
- 「名前」。入力データ・セット中のフィールド名を基準にしてデータ・セットを追加する場合に選択します。また、フィールド名の一致に際して大文字と小文字を区別する場合は、「大文字と小文字を区別」 を選択します。

出力フィールド: レコード追加ノードに接続されているソース・ノードを一覧表示します。リストの最初のノードが主入力ソースになります。表示されているフィールドをソートするには、列見出しをクリックしてください。ソートを行っても、データ・セット中のフィールドの並びは変わりません。

フィールド入力元: メイン・データ・セット内のフィールドを基準にして出力フィールドを生成する場合は、「メイン データセットだけ」 を選択します。メイン データセットは、「入力」タブで指定されている主入力のことです。すべての入力データ・セットに渡って一致フィールドがあるかどうかに関係なく、すべてのデータ・セット中のすべてのフィールドに対して、出力フィールドを選択する場合は、「すべてのデータセット」 を選択します。

フィールドに入力データセットを入れてレコードにタグ付け: 各レコードのソース・データ・セットを示す値を持つフィールドを出力ファイルに追加する場合に選択します。テキスト・フィールドに名前を指定します。デフォルトのフィールド名は、入力 です。

重複レコード・ノード

データ・セット内の重複レコードは、データ・マイニングを開始する前に削除する必要があります。例えば、マーケティング・データベースで、異なる住所または会社情報を持つ個人が複数回出現する場合があります。重複レコード ノードを使用すると、データ内の重複するレコードを検出または削除したり、重複するレコードのグループから単一の複合レコードを作成したりすることができます。

重複レコード ノードを使用するには、まず 2 つのレコードが重複していると見なす条件を規定する一連のキー フィールドを定義する必要があります。

一部のフィールドをキー フィールドとして選択しない場合は、2 つの「重複する」レコードにおいて、依然として残りのフィールドの値が異なる可能性があるため、2 つのレコードを完全には同一と見なせない場合があります。この場合は、重複するレコードのグループそれぞれの中で適用するソート順序を定義することもできます。このソート順序により、グループの中で最初のレコードと見なすレコードを詳細に制御することができます。それ以外の場合は、重複するレコードがすべて同等に扱われ、いずれかのレコードが任意に選択される場合があります。レコードの着信順序は考慮されないため、上流のソート ノードの使用は効果がありません (下記の『重複レコード ノード内のレコードの並べ替え』を参照してください)。

モード: 複合レコードを作成するか、最初のレコードを含めるか除外する (破棄する) かを指定します。

- 各グループの複合レコードを作成。これにより、非数値フィールドを集計することができます。このオプションを選択すると、「複合」タブが使用可能になり、複合レコードの作成方法を指定することができます。詳しくは、 100 ページの『重複レコードの複合の設定』を参照してください。
- 各グループの最初のレコードだけを含める。 重複するレコードのグループそれぞれの中で最初のレコードを選択し、残りを破棄します。最初 のレコードは、レコードの着信順ではなく、以下で定義するソート順で判別されます。
- 各グループの最初のレコードだけを破棄。 重複するレコードのグループそれぞれの中で最初のレコードを破棄し、残りを選択します。最初 のレコードは、レコードの着信順ではなく、以下で定義するソート順で判別されます。このオプションは、データ内の重複を検出 し、後からストリームで調査する場合に役立ちます。

グループ化のキー・フィールド。レコードが同一であるかどうかを判断するために使われるフィールドを表示します。以下を行うことができます。

- このリストにフィールドを追加するには、右側にあるフィールド・ピッカー・ボタンを使用します。
- フィールドを削除するには、赤い X (削除) ボタンをクリックします。

グループ内でレコードをソート。重複するレコードの各グループの中でレコードをどのように並べ替えるか、および昇順と降順のいずれで並べ替えるかを決定するために使用するフィールドをリストします。次の操作が可能です。

- このリストにフィールドを追加するには、右側にあるフィールド・ピッカー・ボタンを使用します。
- フィールドを削除するには、赤い X (削除) ボタンをクリックします。
- フィールドを移動するには、上下 2 つのボタンがあり、複数のフィールドごとに並べ替える場合に使用します。

各グループの最初のレコードを含めるか除外することを選択した場合、どのレコードを先頭と見なすかが重要なときは、ソート順を指定する必要があります。

複合レコードの作成を選択した場合、「複合」タブで指定するオプションによっては、ソート順も指定する必要があります。詳しくは、 100 ページの『重複レコードの複合の設定』を参照してください。

デフォルトのソート順。デフォルトで、ソート キー値の「昇順」または「降順」のどちらでレコードを並べ替えるかを指定します。

重複レコード ノード内のレコードの並べ替え

重複するレコードのグループの中でのレコードの順序が重要な場合は、重複レコード ノードで「グループ内でレコードをソート」オプションを使用して順序を指定する必要があります。上流のソート ノードに依存しないでください。レコードの着信順序は考慮されず、ノード内で指定された順序のみが考慮されることに注意してください。

ソート フィールドを指定しない (またはソート フィールドの指定が不十分な) 場合は、それぞれの重複レコードのグループに含まれるレコードは並べ替えられず (または不完全な並べ替えが行われ)、予測できない結果になる場合があります。

例えば、複数のマシンに関する非常に大規模なログ レコードのセットがあるとします。ログに含まれるデータは以下のようになっています。

表 *16.* マシンのログデータ

タイムスタンプ	マシン	温度
17:00:22	マシン A	31
13:11:30	マシン B	26
16:49:59	マシン A	30
18:06:30	マシン X	32
16:17:33	マシン A	29
19:59:04	マシン C	35
19:20:55	マシン Y	34
15:36:14	マシン X	28
12:30:41	マシン Y	25
14:45:49	マシン C	27
19:42:00	マシン B	34
20:51:09	マシン Y	36
19:07:23	マシン X	33

レコード数を削減して各マシンの最新レコードを残すには、キー フィールドとして「**マシン**」を使用し、ソート フィールド (降順) として「**タイムスタンプ**」を使用します。ソートの選択内容は、指定された「マシン」の多くの行のうち、どれが返されるかを指定するため、入力順序は結果に影響しません。最終的なデータ出力は以下のようになります。

表 *17.* ソート後のマシンのログ データ

タイムスタンプ	マシン	温度
17:00:22	マシン A	31
19:42:00	マシン B	34
19:59:04	マシン C	35
19:07:23	マシン X	33
20:51:09	マシン Y	36

重複レコード最適化設定

処理しているデータのレコード数が少ない場合、またはすでにソートされている場合、IBM SPSS Modeler がデータをより効率的に処理できるようにする方法を最適化できます。

注 : 「入力データのキー フィールド値のバリエーションが少ない」を選択するか、ノードに SQL 生成を使用する場合、異なるキー値内の行を返すことができます。異なるキー内で返される行を制御するには、「設定」タブで 「グループ内でレコードをソート」を使用して、ソート順を指定する必要があります。最適化オプションは、「設定」タブでソート順を指定している限り、重複レコード・ノードによって出力された結果には影響を与えません。

入力データのキー フィールド値のバリエーションが少ない**:** キー・フィールドのレコードが少ないまたは一意の値が少ない場合に指定します。このオプションを選択すると、パフォーマンスが向上します。

入力データ・セットが「設定」タブでフィールドをグループ化およびソートして並べ替えられる**:**「設定」タブの 「グループ内でレコードをソート」 に表示されたすべてのフィールドでデータがすでにソートされている場合、データの降順または昇順が同じ場合にのみこのオプションを選択します。このオプションを選択すると、パフォーマンスが向上します。

SQL 生成を無効にする **:** ノードの SQL 生成を無効にする場合に選択します。

重複レコードの複合の設定

処理するデータに (例えば同一人物に対して) 複数のレコードがある場合は、処理用に単一の複合 (集計) レコードを作成することで、データの操作方法を最適化することができます。

注**:** このタブは、「設定」タブで「各グループの複合レコードを作成」を選択した場合にのみ使用可能です。

「コンポジット」タブのオプションの設定

フィールド**:** この列には、データ・モデルのキー・フィールド以外のすべてのフィールドが、自然なソート順序で表示されます。ノードが接続されていない場合は、フィールドは表示されません。フィールド名のアルファベット順に行をソートするには、列見出しをクリックします。Shift キーまたは Ctrl キーを押しながらクリックすることで、複数の行を選択することができます。また、フィールドを右クリックするとメニューが表示され、すべての行を選択したり、フィールドの名前または値の昇順または降順に行をソートした

り、指標またはストレージのタイプによってフィールドを選択したり、選択したすべての行に同じ「次を基準とする値を入れる」項目を自動的に追加するための値を選択したりすることができます。

次を基準とする値を入れる: 「フィールド」の複合レコードに使用する値のデータ型を選択します。使用できるオプションは、フィールド・タイプにより異なります。

- 数値範囲のフィールドの場合は、以下の中から選択できます。
  - グループ内の最初のレコード
  - グループ内の最後のレコード
  - 合計
  - 平均値
  - 最小値
  - 最大値
  - カスタム
- 時刻または日付のフィールドの場合は、以下の中から選択できます。
  - グループ内の最初のレコード
  - グループ内の最後のレコード
  - 最も早い
  - 最新
  - カスタム
- 文字列またはデータ型不明のフィールドの場合は、以下の中から選択できます。
  - グループ内の最初のレコード
  - グループ内の最後のレコード
  - 最初の英数字
  - 最後の英数字
  - カスタム

いずれの場合も、「ユーザー設定」オプションを使用して、複合レコードの設定にどの値を使用するかを詳細に制御することができます。詳しくは、『重複レコードの複合 - 「ユーザー設定」タブ』を参照してください。

フィールドにレコード度数を含める。このオプションを選択すると、各出力レコードにフィールド (デフォルトでは Record_Count) が追加されます。このフィールドは、各集計レコードを作成するために集計された入力レコード数を示します。このフィールドの名前を自身で指定するには、その名前を編集フィールドに入力してください。

重複レコードの複合 - 「ユーザー設定」タブ

「カスタム入力」ダイアログ・ボックスを使用すると、新しい複合レコードの作成にどの値を使用するかを詳細に制御できます。「コンポジット」タブで単一のフィールド行しかカスタマイズしない場合は、このオプションを使用する前に必ずデータをインスタンス化してください。

注: このダイアログ・ボックスは、「コンポジット」タブの「次を基準とする値を入れる」列で「ユーザー設定」の値を選択した場合にのみ使用可能です。

フィールドのデータ型に応じて、以下のいずれかのオプションから選択することができます。

- 出現頻度で選択。データ・レコードでの出現頻度に基づいて値を選択します。

注: データ型が「連続型」、「データ型不明」、または「日付/時刻」であるフィールドの場合は使用できません。

– 使用。「最も高い頻度」または「最も低い頻度」のいずれかから選択してください。
– 同順位。複数のレコードの出現頻度が同じである場合に、必要なレコードの選択方法を指定します。4 つのオプションが用意されており、「最初を使用」、「最後を使用」、「最低を使用」、または「最高を使用」のいずれかから選択できます。

- 値を含める **(T/F)**。フィールドを、グループ内のすべてのレコードが指定の値を持つかどうかを示すフラグに変換するには、これを選択します。次に、選択したフィールドの項目のリストから「値」を選択します。

  注: 「コンポジット」タブで複数のフィールド行を選択する場合は使用できません。
- リスト内の最初の一致。複合レコードに割り当てる値の優先順位を決定するには、これを選択します。次に、選択したフィールドの項目のリストからいずれかの「項目」を選択します。

  注: 「コンポジット」タブで複数のフィールド行を選択する場合は使用できません。
- 値の連結。グループ内のすべての値を 1 つの文字列に連結することで保持するには、これを選択します。それぞれの値の間に使用する区切り文字を指定する必要があります。

  注: データ型が「連続型」、「データ型不明」、または「日付/時刻」であるフィールド行を 1 つ以上選択する場合は、これ以外のオプションを使用できません。
- 区切り文字を使用。連結した文字列の区切り文字の値として「スペース」または「カンマ」の使用を選択できます。また、「その他」フィールドに、独自の区切り文字の値を入力することもできます。

  注: 「値の連結」オプションを選択する場合にのみ使用可能です。

ストリーミング時系列ノード

ストリーミング時系列ノードを使用し、1 つのステップで時系列モデルを作成してスコアリングします。対象フィールドごとに個別の時系列モデルが構築されますが、生成されたモデルのパレットにはモデル・ナゲットは追加されず、モデル情報も参照できません。

時系列データをモデル作成する方法では、欠損値を空の行で示し、各測定間を均一な区分とすることが求められます。データがこの要件を満たしていない場合、必要に応じて、値を変換する必要があります。

時系列ノードとの使用で注意すべきその他の点は、次のとおりです。

- フィールドは数値型。
- 日付フィールドは入力として使用できない。
- データ区分は無視されること。

ストリーミング時系列ノードは、時系列から指数平滑法、1 変量の自己回帰積分移動平均法 (ARIMA)、および多変量 ARIMA (または伝達関数) モデルを推測し、その時系列データに基づいて予測を作成します。さらに、エキスパート・モデラーも使用できます。これは、1 つ以上の対象フィールドに対して、最も適合する ARIMA または指数平滑化モデルを自動的に識別して評価します。

時系列モデル作成について詳しくは、「SPSS Modeler Modeling Nodes guide」の『時系列モデル』セクションを参照してください。

ストリーミング時系列ノードのストリーミング展開環境での使用は、IBM SPSS Collaboration and Deployment Services Scoring Service を使用して、IBM SPSS Modeler Solution Publisher を通じてサポートされています。

ストリーミング時系列ノード - フィールド オプション

定義済みの役割を使用: このオプションを選択すると、上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) が使用されます。

カスタム・フィールド割り当てを使用: 対象、予測、およびその他の役割を手動で割り当てるには、このオプションを選択します。

注: データを区分した場合、データ区分が考慮されるのは、「定義済みの役割を使用」を選択した場合です。「ユーザー設定フィールドの割り当てを使用」を選択した場合は、考慮されません。

フィールド: 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

対象: 1 つ以上のフィールドを予測の対象として選択します。

入力の候補 1 つ以上のフィールドを予測の入力として選択します。

イベントおよび干渉 特定の入力フィールドをイベント・フィールドまたは干渉フィールドとして指定する場合に、この領域を使用します。この指定により、イベント (販売促進などの予測可能な繰り返し発生する状況) または干渉 (停電や従業員のストライキなど、一時的な出来事) に影響を受ける時系列データを含むものとして、フィールドが識別されます。

ストリーミング時系列ノード - データ指定オプション

「データ指定」タブで、データをモデルに含めるためのすべてのオプションを設定します。「日付/時刻フィールド」と「時間区分」の両方を指定すると、「実行」 ボタンをクリックして、すべてデフォルト・オプションを含むモデルを構築できます。ただし、通常は、それぞれの目的でビルドをカスタマイズする必要があります。

このタブに含まれる数種類のペインを使用して、モデルに固有のカスタマイズを設定します。

ストリーミング時系列ノード - 観測

このペインの設定を使用して、観測を定義するフィールドを指定します。

日付/時刻フィールドによって指定される観測

観測を日付、時刻、またはタイムスタンプのフィールドによって定義することを指定できます。観測を定義するフィールドに加えて、観測を記述する適切な時間区分を選択します。指定した時間区分によっては、観測間の区分 (増分) や週当たりの日数などの他の設定も指定できます。時間区分には、以下の考慮事項があります。

- 観測の時間区分が不規則のときは (例えば、販売注文の処理時)、「不規則」の値を使用します。「不規則」を選択したときは、「データ指定」タブの「時間区分」の設定で、分析に使用する時間区分を指定する必要があります。

- 観測が日付と時刻を表しており、時間区分が時、分または秒の場合は、「**1** 日あたりの時間数」、「**1** 日あたりの分数」、または「**1** 日あたりの秒数」を使用します。観測が日付を参照しない時刻 (期間) を表しており、時間区分が時、分、または秒の場合は、「時間数 **(**非周期的**)**」、「分数 **(**非周期的**)**」、または「秒数 **(**非周期的**)**」を使用します。
- 選択した時間区分に基づき、手続きで欠損観測値を検出できます。欠損観測値の検出が必要であるのは、すべての観測の時間区分が等しく、欠損観測値がないものと手続きで想定されているためです。例えば、時間区分が日であり、日付 2015-10-27 の後に 2015-10-29 がある場合は、2015-10-28 が欠損観測値です。欠損観測値に対して値が代入されます。「データ指定」タブの「欠損値の処理」エリアを使用して、欠損値を処理するための設定を指定してください。
- 指定した時間区分により、手続きは、同じ時間区分内の集計が必要な複数の観測値を検出し、区分境界 (月の先頭など) の観測値を調整できます。これにより、観測値の区分が等しくなります。例えば、時間区分が月の場合は、同じ月の複数の日付が集計されます。このタイプの集計は、グループ化 と呼ばれます。デフォルトでは、観測値はグループ化のときに合計されます。別のグループ化の方法 (観測値の平均など) を指定するには、「データ指定」タブの「集計と分布」の設定を使用します。
- 時間区分によっては、追加設定により、通常の等間隔の区分の区切りを定義できます。例えば、時間区分が日であるが、平日のみが有効の場合は、1 週間の日数が 5 日で、週が月曜日に始まることを指定できます。

期間または循環する期間として定義される観測

観測は、任意の数の循環レベルまでの期間または期間の繰り返しサイクルを表す 1 つ以上の整数フィールドによって定義できます。この構造では、標準の時間区分に適合しない観測の系列を記述できます。例えば、月数が 10 カ月のみの会計年度を、年を表す循環フィールドと月を表す期間フィールドで記述できます (1 つの循環の長さは 10)。

循環する期間を指定するフィールドにより、周期的レベルの階層が定義されます。ここで、最低レベルは「期間」フィールドで定義されます。次の最高のレベルは、レベルが 1 の循環フィールド、レベルが 2 の循環フィールド... という順で指定されます。各レベルのフィールド値は (最高レベルを除く)、次の最高のレベルに関して周期的でなければなりません。最高レベルの値は、周期的にできません。例えば、10 カ月の会計年度の場合、月は年内で周期的であり、年は周期的ではありません。

- 特定レベルの循環の長さが、次の最低レベルの周期性になります。会計年度の例では、循環レベルは 1 つのみあり、循環の長さは 10 でした。これは、次の最低レベルが月を表し、指定した会計年度の月数が 10 カ月であるからです。
- 1 から始まらないすべての周期的フィールドに対して、開始値を指定します。この設定は、欠損値を検出するために必要です。例えば、周期的フィールドが 2 から始まるが、開始値が 1 として指定されている場合、手続きでは、そのフィールドの各循環の最初の期間に欠損値があるものと想定されます。

ストリーミング時系列ノード - 分析の時間区分

分析で使用する時間区分は、観測の時間区分とは異なる場合があります。例えば、観測の時間区分が日の場合に、分析の時間区分に月を選択することがあります。その場合は、データが日次データから月次データに集計されてからモデルが構築されます。データの分布を長い時間区分から短い時間区分にすることも選択できます。例えば、観測が四半期の場合、データを四半期から月次データに分布できます。

このペインの設定を使用して、分析の時間区分を指定します。データを集計または分布する方法は、「データ指定」タブの「集計と分布」の設定から指定します。

分析を実行する時間区分で使用できる選択肢は、観測の定義方法とそれらの観測の時間区分によって決まります。特に、循環する期間によって観測が定義されている場合は、集計のみがサポートされます。その場合、分析の時間区分は、観測の時間区分と等しいか、それより長い必要があります。

ストリーミング時系列ノード - 集計オプションと分布オプション

このペインの設定を使用して、観測の時間区分に関する入力データの集計または分布を行うための設定を指定します。

集計関数**(G)**

分析に使用する時間区分が観測の時間区分よりも長い場合は、入力データが集計されます。例えば、観測の時間区分が日で、分析の時間区分が月のときは集計が実行されます。使用可能な集計関数は、平均、合計、モード、最小、または最大です。

分布関数**(D)**

分析に使用する時間区分が観測の時間区分よりも短い場合は、入力データが分布します。例えば、観測の時間区分が四半期で、分析の時間区分が月のときは分布が実行されます。使用できる分布関数は、平均または合計です。

グループ化関数**(O)**

観測が日付/時刻によって定義されており、複数の観測が同じ時間区分で実行される場合は、グループ化が適用されます。例えば、観測の時間区分が月の場合は、同じ月の複数の日付がグループ化され、観測が実行される月に関連付けられます。使用できるグループ化関数は、平均、合計、モード、最小、または最大です。観測が日付/時刻によって定義されており、観測の時間区分が不規則として指定されている場合は、グループ化が必ず実行されます。

注**:** グループ化は、集計の一つの形式ですが、欠損値を処理する前に実行されます。一方、公式の集計は、欠損値を処理した後に実行されます。観測の時間間隔が不規則と指定されている場合、集計はグループ化関数でのみ実行されます。

日をまたぐ観測を前日に集計**(C)**

時間が日の境界をまたぐ観測を前日の値に集計するかどうかを指定します。例えば、1 日 8 時間の毎時観測を 20:00 に開始する場合、00:00 から 04:00 の間の観測を前日の集計結果に含めるかどうかをこの設定で指定します。この設定が適用されるのは、観測の時間区分が 1 日あたりの時間数、1 日あたりの分数、または 1 日あたりの秒数で、分析の時間区分が日の場合に限られます。

指定されたフィールドのカスタム設定

集計関数、分布関数、およびグループ化関数をフィールドごとに指定できます。この設定により、集計関数、分布関数、およびグループ化関数のデフォルト設定は上書きされます。

ストリーミング時系列ノード - 欠損値オプション

このペインの設定を使用して、入力データの欠損値を代入値で置換する方法を指定します。使用できる置換方法を以下に示します。

線形補間

線形補間を使用して欠損値を置換します。補間には、欠損値の前の最後の有効な値と、欠損値の後の最初の有効な値が使用されます。系列の最初または最後の観測に欠損値がある場合は、その系列の先頭または末尾にある最も近い 2 つの欠損値以外の値が使用されます。

系列平均

欠損値を系列全体の平均値に置換します。

周囲平均値
: 欠損値を、有効な周囲の値の平均値に置換します。隣接ポイントのスパンは、その平均値の計算に使用された欠損値の前後の有効値の数です。

周囲中央値
: 欠損値を、有効な周囲の値の中央値に置換します。隣接ポイントのスパンは、その中央値の計算に使用された欠損値の前後の有効値の数です。

線形トレンド
: このオプションは、系列の欠損観測値以外のすべての値を使用して、単純な線形回帰モデルを適合させます。次にこのモデルは、欠損値を代入するために使用されます。

その他の設定:

最小データ品質スコア **(%) (Lowest data quality score (%))**
: 各時系列に対応する時刻変数および入力データのデータ品質指標を計算します。データ品質スコアがこのしきい値よりも低い場合、対応する時系列は破棄されます。

ストリーミング時系列ノード - 推定期間

「推定期間」ペインで、モデルの推定に使用するレコードの範囲を指定することができます。デフォルトでは、推定期間は、最も早い観測の時刻から始まり、すべての系列における最も遅い観測の時刻に終わります。

開始時刻と終了時刻で指定**(B)**
: 推定期間の開始と終了の両方を指定することも、開始のみまたは終了のみを指定することもできます。推定期間の開始または終了を省略した場合は、デフォルト値が使用されます。

  - 観測が「日付/時刻」フィールドによって定義されている場合は、「日付/時刻」フィールドに使用されているのと同じ形式で開始と終了の値を入力します。
  - 循環する期間によって観測が定義されている場合は、それぞれの「循環する期間」フィールドの値を指定します。各フィールドは、別々の列に表示されます。

最も遅い時間区分または最も早い時間区分で指定 **(By latest or earliest time intervals)**
: データの最も早い時間区分で開始または最も遅い時間区分で終了する時間区分の指定回数として推定期間を定義します。必要に応じてオフセットを指定することができます。このコンテキストでは、時間区分は、分析の時間区分を参照します。例えば、観測は月 1 回であるが、分析の時間区分は四半期であると想定します。「最新」を指定し、「時間区分の数 **(Number of time intervals)**」の値として 24 を指定することは、最新の 24 個の四半期を意味します。

  必要な場合は、指定した数の時間区分を除外することもできます。例えば、最新の 24 個の時間区分を指定し、除外する数値として 1 を指定した場合、推定期間は、最新の区分の前の 24 個の区分から構成されます。

ストリーミング時系列ノード - 作成オプション

「作成オプション」タブで、モデルを構築するすべてのオプションを設定します。もちろん、「実行」ボタンをクリックすると、すべてデフォルト オプションのモデルが構築されますが、通常は、それぞれの目的で作成をカスタマイズする必要があります。

このタブには、2 つの異なるペインが含まれています。これらのペインで、ご使用のモデルに固有のカスタマイズを設定します。

ストリーミング時系列ノード - 一般的な作成オプション

このペインで使用可能なオプションは、「方法」リストから以下の 3 つの設定のうち、どれを選択したかによって異なります。

- エキスパート モデラー。これで、独立した各時系列に最も適合するモデルが自動的に検出されます。
- 指数平滑法。このオプションは、カスタム指数平滑法モデルを指定する場合に使用します。
- **ARIMA**。このオプションは、カスタム ARIMA モデルを指定する場合に使用します。

エキスパート モデラー

「モデル タイプ」で、作成するモデルのタイプを以下から選択します。

- すべてのモデル。エキスパート・モデラーは、ARIMA と指数平滑法モデルの両方を考慮します。
- 指数平滑法モデルのみ。エキスパート・モデラーは、指数平滑法モデルのみを考慮します。
- **ARIMA** モデルのみ。エキスパート・モデラーは ARIMA モデルのみを考慮します。

エキスパート・モデラーが季節モデルを考慮する。このオプションは、アクティブなデータ・セットに周期性が定義されている場合にのみ有効です。このオプションがオンの場合、エキスパート・モデラーは季節性および非季節性の両方のモデルを検討します。このオプションを選択しないと、エキスパート・モデラーは非季節モデルのみを考慮します。

エキスパート モデラーが指数平滑化モデルを考慮。 このオプションを選択した場合、エキスパート モデラーは合計 13 個の指数平滑法モデル (このうちの 7 個は元の時系列ノードに存在していたもので、6 個はバージョン 18.1 で追加されたもの) を検索します。このオプションを選択しないと、エキスパート・モデラーは元の 7 個の指数平滑化モデルのみを検索します。

「外れ値」で、以下のオプションから選択します。

「自動的に外れ値を検出」。デフォルトでは、外れ値の自動検出は実行されません。外れ値の自動検出を実行するには、このオプションをオンにしてから、希望する外れ値のタイプを選択します。

入力フィールドは、測定の尺度がフラグ型、名義型、または順序型である必要があり、このリストに含まれる前に数値 (フラグ型フィールドの場合は True/Falseでなく 1/0) になっていなければなりません。

エキスパート・モデラーは単純な回帰のみを検討し、「フィールド」タブのイベント・フィールドまたは干渉フィールドとして識別される入力フィールドの任意の伝達関数を考慮しません。

指数平滑化

モデル・タイプ**:** 指数平滑化モデルは、季節性または非季節性のどちらかに分類されます。[1] 季節性モデルは、「データ指定」タブの「時間区分」ペインを使用して定義される周期性が季節性である場合にのみ使用できます。季節的な周期性には、循環する期間、年数、四半期数、月数、曜日数、1 日あたりの時間数、1 日あたりの分数、1 日あたりの秒数があります。 選択可能なモデル タイプは、以下のとおりです。

- 「単純」。このモデルは、トレンドまたは季節性のない時系列に適しています。関連する平滑化パラメーターは水準のみです。単純指数平滑化は、0 次の自己回帰、1 次の差分、1 次の移動平均、および定数なしの ARIMA に最もよく似ています。
- **Holt** の線型トレンド**:** このモデルは、線型トレンドがあり季節性のない時系列に適しています。関連する平滑化パラメーターは水準パラメーターとトレンド・パラメーターであり、このモデル内では、互いの値に制約を受けません。Holt のモデルは Brown のモデルよりも一般的ですが、大きな系列の推定値

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

の計算には余計に時間がかかる場合があります。Hplt の指数平滑化は、0 次の自己回帰、2 次の差分、移動平均が 2 次の ARIMA に最もよく似ています。

- 減衰トレンド**:** このモデルは、線型トレンドがあり季節性のない時系列に適しています。関連する平滑化パラメーターは、水準、トレンド、および減衰トレンドです。減衰指数平滑化は、1 次の自己回帰、1 次の差分、および 2 次の移動平均の ARIMA に最もよく似ています。
- 「乗法トレンド」。このモデルは、系列の水準に依存するトレンドがあり、季節性のない系列に適しています。関連する平滑化パラメーターは、水準パラメーターとトレンド・パラメーターです。「乗法トレンド」の指数平滑法は、いかなる ARIMA モデルにも類似しません。
- **Brown** の線型トレンド**:** このモデルは、線型トレンドがあり季節性のない時系列に適しています。関連する平滑化パラメーターは水準パラメーターとトレンド・パラメーターで、モデル内では等しいと見なされます。したがって、Brown モデルは Holt モデルの特別なケースです。Brown の指数平滑化は、ARIMA に最もよく似ています。0 次の自己回帰、2 次の差異、および 2 次の移動平均があり、移動平均の 2 次目の係数が一次の二乗の係数の 1/2 です。
- 単純季節**:** このモデルは、トレンドがなく常に一定の季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準パラメーターと季節パラメーターです。季節指数平滑化は、0 次の自己回帰、1 次の差分、1 次の季節差分、および 1、*p*、および移動平均が *p*+1 の ARIMA に最もよく似ています。この *p* は、季節区間 (季節的な間隔) の周期数です。月次データの場合、*p* = 12 です。
- **Winters** の加法**:** このモデルは、線型トレンドと常に一定の季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、および季節です。Winters の加法指数平滑化は、0 次の自己回帰、1 次の差分、1 次の季節差分、および移動平均が *p*+1 の ARIMA に最もよく似ています。この *p* は、季節区分 (季節的な間隔) の周期数です。月次データの場合、*p* = 12 です。
- 「減衰トレンドと加法的季節性」。このモデルは、減衰する線型トレンドおよび常に一定の季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、減衰トレンド、および季節です。「減衰トレンドと加法的季節性」の指数平滑法は、いかなる ARIMA モデルにも類似しません。
- 「乗法トレンドと加法的季節性」。このモデルは、系列の水準に依存するトレンドおよび常に一定の季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、および季節です。「乗法トレンドと加法的季節性」の指数平滑法は、いかなる ARIMA モデルにも類似しません。
- 「乗法的季節性」。このモデルは、トレンドがなく系列の水準に依存する季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準パラメーターと季節パラメーターです。「乗法的季節性」の指数平滑法は、いかなる ARIMA モデルにも類似しません。
- **Winters** の乗法**:** このモデルは、線型トレンドと系列の水準に依存する季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、および季節です。Winters の相乗指数平滑法は、いかなる ARIMA モデルにも類似しません。
- 「減衰トレンドと乗法的季節性」。このモデルは、減衰する線型トレンドおよび系列の水準に依存する季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、減衰トレンド、および季節です。「減衰トレンドと乗法的季節性」の指数平滑法は、いかなる ARIMA モデルにも類似しません。
- 「乗法トレンドと乗法的季節性」。このモデルは、系列の水準に依存するトレンドおよび季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、および季節です。「乗法トレンドと乗法的季節性」の指数平滑法は、いかなる ARIMA モデルにも類似しません。

対象の変換**:** 各従属変数に、モデル化される前に実行される変換を指定できます。

- なし**:** 変換は実行されません。
- 平方根: 平方根変換が実行されます。

- 自然対数: 自然対数変換が実行されます。

ARIMA

カスタム ARIMA モデルの構造を指定します。

「**ARIMA** の順序」。モデルのさまざまな ARIMA 成分の値を、グリッドの対応するセルに入力します。すべての値は負でない整数にする必要があります。自己回帰と移動平均の成分については、値は最大次数を表します。すべての正の低い次数はモデルに含まれます。例えば、2 を指定すると、モデルには次数 2 と 1 が含まれます。「季節性」列のセルは、周期性がアクティブ データ セットに定義されている場合にのみ有効です。

- 自己回帰 **(p)**。モデル内の自己回帰の次数の数値です。自己回帰の次数は、現在の値を予測するために使用される系列からの前 (過去) の値を指定します。例えば、自己回帰の次数 2 は、現在の値を予測するために系列の値を過去の 2 期間使用するように指定します。
- 差分 **(d)**。モデルを推定する前に系列に適用される差分の次数を指定します。トレンドが存在する場合は差分を取る必要があり (トレンドの存在する系列は通常非定常性であり、ARIMA モデルは定常性を前提としている)、その効果を取り除くために行います。差分の次数は、系列のトレンドの次数に対応しています (1 次差分は線型トレンドを表し、2 次差分は 2 次トレンドを表す、など)。
- 移動平均 **(q)**。モデル内の移動平均の次数の数値。移動平均の次数は、過去の値の系列平均の偏差が、現在の値を予測するためにどのように使用されるかを指定します。例えば、移動平均の次数 1 および 2 は、系列の現在の値を予測する際に最近の 2 期間のそれぞれから取得した系列の平均値の偏差を考慮することを指定します。

季節性。季節型の自己回帰、移動平均、および差分成分は、対応する非季節の成分と同様の役割を果たします。ただし、季節次数については、現在の系列値は、1 つ以上の季節期間で区切られた過去の系列値に影響されます。例えば、毎月のデータ (季節期間 12) については、季節次数 1 は、現在の系列値は現在の期間より 12 期間以前の系列値により影響されることを意味しています。毎月のデータについて、季節次数 1 は、非季節次数 12 を指定するのと同じことになります。

「自動的に外れ値を検出」。外れ値の自動検出を実行するには、このオプションをオンにしてから、使用可能な外れ値のタイプを 1 つ以上選択します。

検出する外れ値のタイプ。検出する外れ値の型を選択します。サポートされるタイプは、次のとおりです。

- 相加的 (デフォルト)
- レベル・シフト (デフォルト)
- 技術革新的
- 過渡
- 季節性相加
- 局所トレンド
- 相加的パッチ

転送関数の順序と変換。ARIMA モデル内の任意またはすべての入力フィールドに対する変換の指定および伝達関数の定義を行うには、「設定」をクリックします。別のダイアログ ボックスが表示され、そこで転送と変換の詳細を入力します。

モデル内に定数項を含める。系列値の全体平均が 0 だという確信がない限り、通常は定数を含めます。差分を適用する場合は、定数を除外することをお勧めします。

伝達関数および変換関数**:** 伝達関数の次数および「変換」ダイアログ ボックスを使用して、ARIMA モデル内の任意またはすべての入力フィールドに対する変換の指定および伝達関数の定義を行います。

対象の変換。このペインでは、各目標変数に、モデル化される前に実行される変換を指定できます。

- なし**:** 変換は実行されません。
- 平方根: 平方根変換が実行されます。
- 自然対数: 自然対数変換が実行されます。

転送の関数と変換の入力の候補。伝達関数を使用して、入力フィールドの過去の値が対象系列の将来の値を予測するために使用される方法を指定します。ペインの左側のリストには、すべての入力フィールドが表示されます。このペインのその他の情報は、選択した入力フィールドに固有のものです。

伝達関数の次数。伝達関数のさまざまな成分の値を、「構造」グリッドの対応するセルに入力します。すべての値は負でない整数にする必要があります。分子と分母の成分については、値は最大次数を表します。すべての正の低い次数はモデルに含まれます。さらに、次数 0 は常に分子成分に含まれます。例えば、分子に 2 を指定すると、モデルには次数 2、1 および 0 が含まれます。分母に 3 を指定するとモデルには次数 3、2、および 1 が含まれます。「季節性」列のセルは、周期性がアクティブ データ セットに定義されている場合にのみ有効です。

分子。伝達関数の分子の次数で、従属系列の現在の値を予測するために使用される選択した独立 (予測値) 系列からの前 (過去) の値を指定します。例えば、分子次数 1 は、独立系列の現在の値だけでなく、過去の 1 期間における独立系列の値が各従属系列の現在の値を予測するために使用されることを指定します。

分母。伝達関数の分母の次数で、従属系列の現在の値を予測するために、選択された独立 (予測値) 系列の前 (過去) の値に対して系列の平均からどのくらいの偏差が使用されるかを指定します。例えば、分母次数 1 は、各従属系列の現在の値を予測する際に、過去の 1 期間における独立系列の平均値の偏差が考慮されることを指定します。

差分。モデルを推定する前に、選択された独立 (予測) 系列に適用される差分の次数を指定します。トレンドが存在する場合は差分を取る必要があり、トレンドの効果を取り除くために差分を使用します。

季節性。季節分子、分母、および差分成分は、対応する非季節の成分と同様の役割を果たします。ただし、季節次数については、現在の系列値は 1 つ以上の季節期間で区切られた過去の系列値に影響されます。例えば、毎月のデータ (季節期間 12) については、季節次数 1 は、現在の系列値は現在の期間より 12 期間以前の系列値により影響されることを意味しています。毎月のデータについて、季節次数 1 は、非季節次数 12 を指定するのと同じことになります。

遅延。遅延を設定すると、入力フィールドへの影響が指定した間隔数で遅延されます。例えば遅延が 5 に設定された場合、時間 t での入力フィールドの値は、5 期間が経過するまで $(t + 5)$ 予測に影響しません。

変換**:** 独立変数のセットに対する伝達関数の仕様にも、そのような変数に実行されるオプションの変換が含まれます。

- なし**:** 変換は実行されません。
- 平方根: 平方根変換が実行されます。
- 自然対数: 自然対数変換が実行されます。

ストリーミング時系列ノード - モデル オプション

信頼限界幅 **(%)**。信頼区間は、モデルの予測と残差自己相関に対して計算されます。100 未満の正の値を指定できます。デフォルトでは、95% の信頼区間が使用されます。

「レコードの将来への拡張」のオプションで、推定期間の終わりを超えて予測する時間区分の数を設定します。この場合の時間区分は、「データ指定」タブで指定した分析の時間区分です。予測が要求されると、対象でもない任意の入力系列に対して、自己回帰モデルが自動的に構築されます。次にこれらのモデルは、予測期間のこれらの入力系列の値を生成するために使用されます。この設定の最大値制限はありません。

予測で使用する将来の値

- 将来の入力値を計算**:** このオプションを選択した場合、予測、ノイズ予測、分散推定、および将来の時間値の予測値が自動的に計算されます。予測が要求されると、対象でもない任意の入力系列に対して、自己回帰モデルが自動的に構築されます。次にこれらのモデルは、予測期間のこれらの入力系列の値を生成するために使用されます。
- 値をデータに追加するフィールドの選択**:** 予測する各レコードに対し (ホールド・アウトは除く)、予測フィールド (役割を**入力**に設定) を使用する場合、各予測の予測期間に対し、推定値を指定できます。手動で値を指定することも、リストから選択することもできます。
  - フィールド**:** フィールド選択ボタンをクリックし、予測として使用するフィールドを選択します。ここで選択したフィールドは、モデル作成で使用されることも、使用されないこともあります。フィールドを予測フィールドとして実際に使用するには、下流のモデル作成ノードで選択する必要があります。このダイアログ・ボックスは将来の値を指定する便利な場所であり、下流のモデル作成ノードによって共有できるので、各ノードで将来の値を個別に指定しなくても済みます。また、利用できるフィールドの一覧が「作成オプション」タブでの選択に制約される場合があります。

    将来の値がストリーム内で以後使用されないフィールドに指定された場合 (削除されたか、「作成オプション」タブで選択が更新されたことが原因)、そのフィールドは赤で表示されることに注意してください。
  - 値**:** 各フィールドに対し、関数のリストから選択するか、または「指定」をクリックして手動で入力または事前に定義された値から選択することができます。予測フィールドが、管理するまたは事前に検知できる項目と関連する場合、値を手動で入力する必要があります。例えば、部屋の予約数に基づいてホテルの翌月の収益を予測する場合、当月実際に取得した予約数を指定することができます。それに対し、予測フィールドが株価など管理外のものと関連する場合、最も最近使用した値や最近使用したポイントの平均などの関数を使用することができます。

    利用できる関数は、フィールドの尺度によって異なります。

表 *18.* 測定の尺度に使用可能な関数

尺度	関数
連続型フィールドまたは名義型フィールド	ブランク 最近使用したポイントの平均 最も最近使用した値 指定
フラグ型フィールド	ブランク 最も最近使用した値 真 偽 指定

「最近使用したポイントの平均」は、最近の 3 つのデータ・ポイントの平均から将来の値を計算します。

「最も最近使用した値」は、将来の値を最新のデータ・ポイントの値に設定します。

「真/偽」は、フラグ型フィールドの将来の値を、指定した真または偽に設定します。

「指定」を選択すると、手動で将来の値を指定するため、または事前定義されたリストから値を選択するためのダイアログ・ボックスが開きます。

スコアリングで使用可能にする

モデル・ナゲットのダイアログ・ボックスに表示されるスコアリング・オプションのデフォルト値を設定できます。

- 確信度の上限および下限の計算**:** このオプションを選択すると、各対象フィールドの信頼区間の上限と下限にそれぞれ対応する新規フィールド (デフォルトの接頭辞 $TSLCI- および $TSUCI- が付きます) が作成されます。
- ノイズの残差の計算**:** このオプションを選択すると、各対象フィールドのモデル残差に対応する新規フィールド (デフォルトの接頭辞 $TSResidual- が付きます) が、これらの値の合計とともに作成されます。

モデルの設定

出力に表示するモデルの最大数**:** 出力に含めるモデルの最大数を指定します。構築されたモデルの数がこのしきい値を超えると、モデルは出力に表示されませんが、スコアリングには引き続き使用できるということに注意してください。デフォルト値は 10 です。多数のモデルを表示すると、パフォーマンスが低下したり、アプリケーションが不安定になったりする可能性があります。

SMOTE ノード

SMOTE (Synthetic Minority Over-sampling Technique) ノードは不均衡データ・セットを扱うためのオーバーサンプリング・アルゴリズムを提供します。これにより、データの均衡化のための高度な手法が提供されます。SMOTE プロセス ノードは Python で実装されており、`imbalanced-learn`© Python ライブラリーを必要とします。imbalanced-learn ライブラリーについて詳しくは、http://contrib.scikit-learn.org/imbalanced-learn/about.html[1] を参照してください。

「ノード パレット」の「Python」タブには、SMOTE ノードおよびその他の Python ノードがあります。

[1]Lemaître, Nogueira, Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research*, vol. 18, no. 17, 2017, pp. 1-5. (http://jmlr.org/papers/v18/16-365.html)

SMOTE ノードの設定

SMOTE ノードの「設定」タブで、次の設定を定義します。

対象の設定

対象フィールド**:** 対象フィールドを選択します。すべての尺度タイプ (フラグ型、名義型、順序型、および離散型) がサポートされます。「データ区分」セクションで「データ区分データを使用」オプションを選択した場合、学習データのみがオーバーサンプリングされます。

オーバーサンプリング率

「自動」を選択してオーバーサンプリング率を自動的に選択するか、「比率 **(**マジョリティーに対するマイノリティーの比率**)** の設定」を選択してカスタムの比率の値を選択します。これは、マジョリティー クラスのサンプル数に対するマイノリティー クラスのサンプル数の比率です。値は **0** より大きく **1** 以下でなければなりません。

ランダム シード

ランダム・シードの設定**:** 乱数発生ルーチンによって使用されるシードを生成するには、このオプションを選択し、「生成」をクリックします。

方法

アルゴリズムの種類**:** 使用する SMOTE アルゴリズムの種類を選択します。

サンプルのルール

K Neighbours: 合成サンプルを作成するために使用する最近傍の数を指定します。

M Neighbours: マイノリティー サンプルが危険な状況にあるかをどうか判断するために使用する最近傍の数を指定します。これは、SMOTE アルゴリズムの種類「**Borderline1**」または「**Borderline1**」が選択された場合にのみ使用されます。

データ区分

データ区分データを使用。 学習データのオーバーサンプリングのみを行う場合は、このオプションを選択します。

SMOTE ノードは、`imbalanced-learn`© Python ライブラリーを必要とします。次の表に、SPSS Modeler の SMOTE ノードのダイアログの設定と Python アルゴリズムとの間の関係を示します。

表 *19.* ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**Python API** パラメータ名
オーバーサンプリング率 (数値の入力コントロール)	`sample_ratio_value`	`ratio`
ランダム シード	`random_seed`	`random_state`
K_Neighbours	`k_neighbours`	`k`
M_Neighbours	`m_neighbours`	`m`
アルゴリズムの種類	`algorithm_kind`	`kind`

拡張変換ノード

拡張変換ノードを使用すると、IBM SPSS Modeler ストリームからデータを取得し、R スクリプトまたは Python for Spark スクリプトを使用して、取得したデータに変換を適用することができます。変更されたデータは、さらなる処理、モデル構築、およびモデル・スコアリングのために、ストリームに返されます。拡張変換ノードを使用すると、R または Python for Spark で記述したアルゴリズムを使用して、データを変換することができます。また、ユーザーは、拡張変換ノードを使用して、特定の問題に対応したデータ変換手法を開発することができます。

このノードを R と共に使用するには、IBM SPSS Modeler - Essentials for R をインストールする必要があります。インストール手順と、互換性に関する情報については、「*IBM SPSS Modeler - Essentials for R:* インストール手順」を参照してください。また、ご使用のコンピューターに、互換性のあるバージョンの R がインストールされている必要もあります。

拡張変換ノードの「シンタックス」タブ

シンタックスのタイプ (**R** または **Python for Spark**) を選択します。詳しくは、以下のセクションを参照してください。シンタックスの準備ができたら、「実行」をクリックして、拡張変換ノードを実行できます。

R シンタックス

「**R** シンタックス」。データ分析用のカスタムの R スクリプト・シンタックスを、このフィールドに入力するか、貼り付けることができます。

「フラグ型フィールドの変換」。フラグ型フィールドの処理方法を指定します。「文字列を因子に、整数および実数を倍精度に」および「論理値 **(**真、偽**)**」の 2 つのオプションがあります。「論理値 **(**真、偽**)**」を選択した場合、フラグ型フィールドの元の値は失われます。例えば、フィールドに「Male」および「Female」という値がある場合、これらの値は「真」および「偽」に変更されます。

「欠損値を **R** の欠損値 **(NA)** に変換」。選択すると、欠損値はすべて、R の `NA` 値に変換されます。R では、欠損値の識別に値 `NA` が使用されます。使用する R 関数によっては、データに `NA` が含まれていた場合の関数の動作を制御するために使用される引数が含まれている場合があります。例えば、関数によって `NA` を含むレコードを自動的に除外することを選択できる場合があります。このオプションが選択されない場合、すべての欠損値はそのまま R に渡されます。これらの欠損値は R スクリプトの実行時にエラーの原因となる可能性があります。

「時間帯を考慮した特殊な制御で日時フィールドを **R** のクラスに変換」。選択すると、日付形式または日付/時刻形式の変数が R の日付/時刻形式に変換されます。次のいずれかのオプションを選択する必要があります。

- 「**R POSIXct**」。日付形式または日付/時刻形式の変数が R の `POSIXct`オブジェクトに変換されます。
- 「**R POSIXlt (**リスト**)**」。日付形式または日付/時刻形式の変数が R の `POSIXlt`オブジェクトに変換されます。

注**:** POSIX 形式は、拡張オプションです。これらのオプションは、ご使用の R スクリプトで、これらの形式を必要とする方法で日付/時刻フィールドを処理するように指定している場合にのみ使用してください。POSIX 形式は、時刻形式の変数には適用されません。

Python シンタックス

「**Python** シンタックス」。データ分析用のカスタムの Python スクリプト・シンタックスを、このフィールドに入力するか、貼り付けることができます。Python for Spark について詳しくは、Python for Spark およびPython for Spark を使用したスクリプトを参照してください。

拡張変換ノード **-** 「コンソール出力」タブ

「コンソール出力」タブには、「シンタックス」タブの R スクリプトまたは Python for Spark スクリプトが実行されたときに受信するすべての出力が含まれます (例えば、R スクリプトを使用する場合、「シンタックス」タブの「**R** シンタックス」フィールドにある R スクリプトが実行されたときに R コンソールから受信する出力が表示されます)。この出力には、R スクリプトまたは Python スクリプトの実行時に

生成される R または Python のエラー・メッセージや警告が含まれる場合があります。出力は、主にスクリプトをデバッグするために使用できます。「コンソール出力」タブには、「**R** シンタックス」フィールドまたは「**Python** シンタックス」フィールドのスクリプトも表示されます。

拡張変換スクリプトが実行されるたびに、R コンソールまたは Python for Spark から受信した出力で「コンソール出力」タブの内容が上書きされます。出力を編集することはできません。

Space-Time-Box ノード

Space-Time-Box (STB) は、Geohash の空間的な場所を拡張したものです。具体的には、STB は英数字の文字列で、空間および時間を規則的に分割した領域です。

例えば、STB「**dr5ru7|2013-01-01 00:00:00|2013-01-01 00:15:00**」は、次の 3 つの部分から構成されます。

- geohash **dr5ru7**
- 開始タイム・スタンプ **2013-01-01 00:00:00**
- 終了タイム・スタンプ **2013-01-01 00:15:00**

例として、空間と時間の情報を使用することにより、2 つのエンティティーが実際同じ時間に同じ場所に存在しているため、これらのエンティティーが同じものであるという信頼性を向上させることができます。また、2 つのエンティティーが空間と時間において近接しているため、これらのエンティティーには関連性があるということを示すことにより、関係性識別の精度を向上させることもできます。

要件に応じて、「個々のレコード」または「ハングアウト」モードを選択できます。いずれのモードでも、以下に示す基本的な同じ情報が必要になります。

「緯度」フィールド**:** 緯度を (WGS84 座標系で) 識別するフィールドを選択します。

「経度」フィールド**:** 経度を (WGS84 座標系で) 識別するフィールドを選択します。

「タイム・スタンプ」フィールド**:** 時間または日付を識別するフィールドを選択します。

個々のレコード オプション

レコードにフィールドを追加して、特定の時間のレコードの場所を識別する場合に、このモードを使用します。

フィールド作成**:** 新規フィールドを作成するための空間と時間の密度を 1 つ以上選択します。詳しくは、117 ページの『Space-Time-Box 密度の定義』を参照してください。

フィールド名拡張子**:** 新しいフィールド名に追加する拡張子を入力します。「接尾辞」または「接頭辞」のいずれかを選択して、この拡張子を追加できます。

ハングアウト オプション

ハングアウトは、エンティティーが継続的または繰り返し検出される場所または時間と考えることができます。例えば、定期的に輸送される車両を識別したり、通常の状態からの偏差を識別したりするのに使用できます。

ハングアウト検出機能は、エンティティーとフラグ条件の挙動をモニターします。これにより、領域の中でエンティティーが「ハングアウト」していることが識別されます。ハングアウト検出機能は、フラグが立て

られた各ハングアウトを自動的に 1 つ以上の STB に割り当て、メモリー内のエンティティーとイベント追跡を使用して、最良の効率でハングアウトを検出します。

STB 密度: 新規フィールドを作成するための空間と時間の密度を選択します。例えば、**STB_GH4_10MINS** の値を、約 20km 四方の大きさの 4 文字の geohash ボックスと 10 分の時間枠に対応させます。詳しくは、 117 ページの『Space-Time-Box 密度の定義』を参照してください。

「エンティティー **ID**」フィールド**:** ハングアウト ID として使用するエンティティーを選択します。この ID フィールドでイベントを識別します。

イベントの最小数**:** データ内では、1 つのイベントは 1 つの行で表されます。エンティティーがハングアウトしていると見なされる、イベントの発生数の最小値を選択します。ハングアウトは、下記の「滞留時間の下限」フィールドに基づいて限定する必要もあります。

滞留時間の下限**:** エンティティーが同じ場所に留まる必要がある最小期間を指定します。これは、例えば、信号で停止している車両をハングアウトの対象から除外する場合に有効です。ハングアウトは、前記の「イベントの最小数」フィールドに基づいて限定する必要もあります。

ハングアウトを限定する要素の詳細は以下のとおりです。

$e_1, \ldots, e_n$ が、期間 (t_1, t_n) に所定のエンティティー ID から受信したイベントをすべて時刻順に並べたものを示すとします。これらのイベントは、以下の場合にハングアウトに該当します。

- $n >=$ イベントの最小数
- $t_n - t_1 >=$ 最小滞留時間
- すべてのイベント $e_1, \ldots, e_n$ が同じ STB で発生する

STB 境界をまたぐハングアウトを許可 **(Allow hangouts to span STB boundaries):** このオプションを選択すると、ハングアウトの定義が緩和されます。例えば、複数の Space-Time-Box でハングアウトしているエンティティーを含めることができます。例えば、STB を全時間として定義している場合にこのオプションを選択すると、午前 0 時の 30 分前から午前 0 時の 30 分後の 1 時間であっても、1 時間ハングアウトしているエンティティーは有効であるとして認識されます。このオプションを選択していない場合は、ハングアウト時間の 100% が単一の Space-Time-Box 内で無ければなりません。

限定時間ボックス内でのイベントの最小比率**(%):** 「**STB** 境界をまたぐハングアウトを許可 **(Allow hangouts to span STB boundaries)**」を選択した場合にのみ有効です。このオプションは、1 つの STB 内でレポートされたハングアウトが実際に別の STB にどの程度オーバーラップしているかを制御する場合に使用します。ハングアウトを識別するために、単一の STB 内で発生する必要があるイベントの最小割合を選択します。この値を 25% に設定して、イベントの割合が 26% である場合は、ハングアウトであると認定されます。

例えば、1 つの 4 バイト geohash スペース ボックスおよび 10 分の時間ボックス (STB_NAME = STB_GH4_10MINS) にイベントが少なくとも 2 件必要であり (イベントの最小数 = 2)、連続する滞在時間が少なくとも 2 分でなければならないという条件でハングアウト検出機能を構成したとします。ハングアウトが検出された場合は、例えば、同じ 4 バイト geohash スペース ボックスにエンティティーが滞在し、午後 4:57 から午後 5:07 までの 10 分間のうち、午後 4:58、午後 5:01、午後 5:03 に 3 つの限定イベントが発生しています。限定時間ボックスの割合はハングアウトの評価となる STB を指定し、以下のようになります。

- **100%**。ハングアウトは、午後 5:00 から 5:10 の時間ボックスの中で報告され、午後 4:50 から 5:00 の時間ボックスでは報告されません (午後 5:01 と午後 5:03 のイベントは、ハングアウト限定のために必要なすべての条件を満たし、これらのイベントの 100% が 5:00 から 5:10 の時間ボックスで発生しています)。
- **50%**。両方の時間ボックスのハングアウトが報告されます (午後 5:01 と午後 5:03 のイベントがハングアウト限定のために必要なすべての条件を満たし、これらのイベントのうち 50% 以上が 4:50 から 5:00 までの時間ボックスで発生し、これらのイベントの 50% 以上が 5:00 から 5:10 の時間ボックスで発生しています)。
- **0%**。両方の時間ボックスのハングアウトが報告されます。

0% と指定すると、限定期間に接触するすべての時間ボックスを表す STB がハングアウトの報告の対象になります。限定期間は、STB の時間ボックスの対応する期間以下でなければなりません。つまり、20 分の限定期間と共に 10 分の STB を構成するようなことはできません。

ハングアウトは、限定条件が満たされるとただちに報告されます。報告は STB ごとに 1 回までです。ハングアウトに該当するイベントが 3 件あり、合計 10 件のイベントがすべて同じ STB の限定期間の中で発生したとします。その場合は、3 番目の該当イベントが発生したときにハングアウトが報告されます。その後の 7 件のイベントでは、いずれもハングアウトの報告がトリガーされません。

注**:**

- ハングアウト検出機能のメモリー内のイベント データはプロセス間で共有されません。そのため、特定のエンティティーが特定のハングアウト検出機能ノードと類似します。つまり、エンティティーの受信モーション データは、常に一貫して、そのエンティティーを追跡するハングアウト検出機能ノードに渡されます。通常、このノードは、実行を通じて同じノードです。
- ハングアウト検出機能のメモリー内のイベント データは一時的なものです。ハングアウト検出機能を終了して再起動すると、処理中であったハングアウトはすべて失われます。そのため、プロセスを停止して再起動すると、システムで実際のハングアウトが報告されない原因になる可能性があります。有効である可能性がある対策として、一部の履歴モーション データを再生する (例えば、48 時間前にさかのぼって、再起動したすべてのノードに適用可能なモーション レコードを再生する) 方法があります。
- ハングアウト検出機能には、時系列でデータを供給する必要があります。

Space-Time-Box 密度の定義

各 Space-Time-Box (STB) に含める物理領域と経過時間を指定することにより、STB のサイズ (密度) を選択します。

領域密度 **(Geo density)**。各 STB に含める領域のサイズを選択します。

時間区分。各 STB に含める時間数を選択します。

フィールド名。STB を接頭辞として、前の 2 つのフィールドの選択に基づいて自動的に入力されます。

ストリーミング TCM ノード

ストリーミング TCM ノードを使用して、1 つのステップで時間的因果モデルを作成およびスコアリングできます。

時間的因果モデリングについて詳しくは、「SPSS Modeler モデル作成ノード」ガイドの『時系列モデル』セクションにある『時間的因果モデリング』のトピックを参照してください。

ストリーミング TCM ノード - 時系列オプション

「フィールド」タブで「時系列」の設定を使用すると、モデル システムに含める系列を指定できます。

データに適用するデータ構造のオプションを選択します。多次元データの場合は、「次元の選択」をクリックして、次元フィールドを指定します。次元フィールドを指定した順序により、後続のすべてのダイアログおよび出力に次元フィールドが表示される順序が定義されます。次元フィールドを並べ替えるには、「次元の選択」サブダイアログの上下の矢印ボタンを使用します。

列ベースのデータの場合、系列 という語は、フィールド という語と同じ意味を持ちます。多次元データの場合、時系列を含むフィールドは、メトリック フィールドと呼ばれます。多次元データの時系列は、メトリック フィールドと各次元フィールドの値によって定義されます。列ベースのデータと多次元データの両方に以下の考慮事項があります。

- 入力の候補として、または対象と入力の両方として指定された系列は、各対象のモデルに含めるものとして考慮されます。各対象のモデルには、対象自身の遅れた値が必ず含まれます。
- 強制入力として指定された系列は、必ず各対象のモデルに含められます。
- 少なくとも一つの系列を対象、または対象と入力の両方として指定する必要があります。
- 「定義済みの役割を使用」を選択した場合、入力の役割を持つフィールドは、入力の候補として設定されます。定義済みの役割が強制入力にマップされることはありません。

多次元データ

多次元データの場合は、メトリック フィールドと関連役割をグリッドに指定します (グリッドの各行には、単一のメトリックと役割が指定されます)。デフォルトでは、モデル システムには、グリッドの各行の次元フィールドのすべての組み合わせの系列が含まれます。例えば、*region* および *brand* の次元がある場合、デフォルトでは、メトリック *sales* を対象として指定することは、*region* および *brand* の組み合わせごとに別々の販売対象系列があることを意味します。

グリッドの行ごとに、次元の省略記号ボタンをクリックすることにより、任意の次元フィールドの値のセットをカスタマイズできます。このアクションにより、「次元の値の選択」サブダイアログが開きます。グリッドの行を追加、削除、またはコピーすることもできます。

「系列の数」列には、関連メトリックに現在指定されている次元の値のセットの数が表示されます。表示される値は、実際の系列の数 (セット当たり 1 つの系列) よりも大きい場合があります。この条件が発生するのは、指定した次元の値の組み合わせのいくつかが、関連メトリックに含まれる系列に対応していないときです。

ストリーミング TCM ノード - 次元の値の選択

多次元データの場合、特定の役割を持つ特定のメトリック フィールドに適用する次元の値を指定して、分析をカスタマイズできます。例えば、*sales* がメトリック フィールドで、*channel* が値「retail」および「web」を持つ次元の場合は、「web」の販売を入力にし、「retail」の販売を対象にすることを指定できます。分析で使用されるすべてのメトリック フィールドに適用される次元のサブセットを指定することもできます。例えば、*region* が地理的領域を示す次元フィールドの場合は、分析を特定の領域に限定できます。

すべての値
: 現在の次元フィールドのすべての値を含めることを指定します。このオプションはデフォルトです。

含める値または除外する値の選択**(L)**
: このオプションは、現在の次元フィールドの値のセットを指定する場合に使用します。「モード」

に「含める」が選択されている場合は、「選択した値」リストに指定されている値のみが含まれます。「モード」に対して「除外」が選択されている場合は、「選択した値」リストに指定されている値以外のすべての値が含まれます。

選択元の値のセットをフィルタリングすることが可能です。フィルター条件に一致した値は、「一致」タブに表示されます。フィルター条件に一致しない値は、「選択されていない値」リストの「不一致」タブに表示されます。「すべて」タブには、フィルター条件にかかわらず、選択されていない値がすべてリストされます。

- フィルターを指定するときは、ワイルドカード文字を示すアスタリスク (*) を使用できます。
- 現在のフィルターをクリアするには、「表示された値のフィルター」ダイアログで検索語に空の値を指定します。

ストリーミング TCM ノード - 観測オプション

「フィールド」タブにある「観測」の設定を使用すると、観測を定義するためのフィールドを指定できます。

日付/時刻によって定義された観測

観測を日付、時刻、またはタイムスタンプのフィールドで定義することを指定できます。観測を定義するフィールドに加えて、観測を記述する適切な時間区分を選択します。指定した時間区分によっては、観測間の区分 (増分) や週当たりの日数などの他の設定も指定できます。時間区分には、以下の考慮事項があります。

- 観測の時間区分が不規則のときは (例えば、販売注文の処理時)、「不規則」の値を使用します。「不規則」を選択したときは、「データ指定」タブの「時間区分」の設定で、分析に使用する時間区分を指定する必要があります。
- 観測が日付と時刻を表しており、時間区分が時、分または秒の場合は、「**1** 日あたりの時間数」、「**1** 日あたりの分数」、または「**1** 日あたりの秒数」を使用します。観測が日付を参照しない時刻 (期間) を表しており、時間区分が時、分、または秒の場合は、「時間数 **(**非周期的**)**」、「分数 **(**非周期的**)**」、または「秒数 **(**非周期的**)**」を使用します。
- 選択した時間区分に基づき、手続きで欠損観測値を検出できます。欠損観測値の検出が必要であるのは、すべての観測の時間区分が等しく、欠損観測値がないものと手続きで想定されているためです。例えば、時間区分が日であり、日付 2014-10-27 の後に 2014-10-29 がある場合は、2014-10-28 が欠損観測値です。欠損観測値がある場合は、値が代入されます。欠損値を処理するための設定は、「データ指定」タブから指定できます。
- 指定した時間区分により、手続きは、同じ時間区分内の集計が必要な複数の観測値を検出し、区分境界 (月の先頭など) の観測値を調整できます。これにより、観測値の区分が等しくなります。例えば、時間区分が月の場合は、同じ月の複数の日付が集計されます。このタイプの集計は、グループ化 と呼ばれます。デフォルトでは、観測値はグループ化のときに合計されます。別のグループ化の方法 (観測値の平均など) を指定するには、「データ指定」タブの「集計と分布」の設定を使用します。
- 時間区分によっては、追加設定により、通常の等間隔の区分の区切りを定義できます。例えば、時間区分が日であるが、平日のみが有効の場合は、1 週間の日数が 5 日で、週が月曜日に始まることを指定できます。

期間または循環する期間によって定義される観測

観測は、任意の数の循環レベルまでの期間または期間の繰り返しサイクルを表す 1 つ以上の整数フィールドによって定義できます。この構造では、標準時間区分の一つを適合させない観測の系列

を記述できます。例えば、月数が 10 カ月のみの会計年度を、年を表す循環フィールドと月を表す期間フィールドで記述できます (1 つの循環の長さは 10)。

循環する期間を指定するフィールドにより、周期的レベルの階層が定義されます。ここで、最低レベルは「期間」フィールドで定義されます。次の最高のレベルは、レベルが 1 の循環フィールド、レベルが 2 の循環フィールド... という順で指定されます。各レベルのフィールド値は (最高レベルを除く)、次の最高のレベルに関して周期的でなければなりません。最高レベルの値は、周期的にできません。例えば、10 カ月の会計年度の場合、月は年内で周期的であり、年は周期的ではありません。

- 特定レベルの循環の長さが、次の最低レベルの周期性になります。会計年度の例では、循環レベルは 1 つのみあり、循環の長さは 10 でした。これは、次の最低レベルが月を表し、指定した会計年度の月数が 10 カ月であるからです。
- 1 から始まらないすべての周期的フィールドに対して、開始値を指定します。この設定は、欠損値を検出するために必要です。例えば、周期的フィールドが 2 から始まるが、開始値が 1 として指定されている場合、手続きでは、そのフィールドの各循環の最初の期間に欠損値があるものと想定されます。

ストリーミング TCM ノード - 時間区分オプション

分析に使用される時間区分は、観測の時間区分と異なる場合があります。例えば、観測の時間区分が日の場合に、分析の時間区分に月を選択することがあります。その場合は、データが日次データから月次データに集計されてからモデルが構築されます。データの分布を長い時間区分から短い時間区分にすることも選択できます。例えば、観測が四半期の場合、データを四半期から月次データに分布できます。

分析を実行する時間区分で使用できる選択肢は、観測の定義方法とそれらの観測の時間区分によって決まります。特に、循環する期間によって観測が定義されている場合は、集計のみがサポートされます。その場合、分析の時間区分は、観測の時間区分と等しいか、それより長い必要があります。

分析の時間間隔は、「データ仕様 (Data Specifications)」タブの「時間区分」設定から指定されます。データを集計または分布する方法は、「データ指定」タブの「集計と分布」の設定から指定します。

ストリーミング TCM ノード - 集計と分布のオプション

集計関数**(G)**

分析に使用する時間区分が観測の時間区分よりも長い場合は、入力データが集計されます。例えば、観測の時間区分が日で、分析の時間区分が月のときは集計が実行されます。使用可能な集計関数は、平均、合計、モード、最小、または最大です。

分布関数**(D)**

分析に使用する時間区分が観測の時間区分よりも短い場合は、入力データが分布します。例えば、観測の時間区分が四半期で、分析の時間区分が月のときは分布が実行されます。使用できる分布関数は、平均または合計です。

グループ化関数**(O)**

観測が日付/時刻によって定義されており、複数の観測が同じ時間区分で実行される場合は、グループ化が適用されます。例えば、観測の時間区分が月の場合は、同じ月の複数の日付がグループ化され、観測が実行される月に関連付けられます。使用できるグループ化関数は、平均、合計、モード、最小、または最大です。観測が日付/時刻によって定義されており、観測の時間区分が不規則として指定されている場合は、グループ化が必ず実行されます。

注**:** グループ化は、集計の一つの形式ですが、欠損値を処理する前に実行されます。一方、公式の集計は、欠損値を処理した後に実行されます。観測の時間間隔が不規則と指定されている場合、集計はグループ化関数でのみ実行されます。

日をまたぐ観測を前日に集計**(C)**

時間が日の境界をまたぐ観測を前日の値に集計するかどうかを指定します。例えば、1 日 8 時間の毎時観測を 20:00 に開始する場合、00:00 から 04:00 の間の観測を前日の集計結果に含めるかどうかをこの設定で指定します。この設定が適用されるのは、観測の時間区分が 1 日あたりの時間数、1 日あたりの分数、または 1 日あたりの秒数で、分析の時間区分が日の場合に限られます。

指定されたフィールドのカスタム設定

集計関数、分布関数、およびグループ化関数をフィールドごとに指定できます。この設定により、集計関数、分布関数、およびグループ化関数のデフォルト設定は上書きされます。

ストリーミング TCM ノード - 欠損値オプション

入力データの欠損値は、代入値で置換されます。使用できる置換方法を以下に示します。

線形補間

線形補間を使用して欠損値を置換します。補間には、欠損値の前の最後の有効な値と、欠損値の後の最初の有効な値が使用されます。系列の最初または最後の観測に欠損値がある場合は、その系列の先頭または末尾にある最も近い 2 つの欠損値以外の値が使用されます。

系列平均

欠損値を系列全体の平均値に置換します。

周囲平均値

欠損値を、有効な周囲の値の平均値に置換します。隣接ポイントのスパンは、その平均値の計算に使用された欠損値の前後の有効値の数です。

周囲中央値

欠損値を、有効な周囲の値の中央値に置換します。隣接ポイントのスパンは、その中央値の計算に使用された欠損値の前後の有効値の数です。

線形トレンド

このオプションは、系列の欠損観測値以外のすべての値を使用して、単純な線形回帰モデルを適合させます。次にこのモデルは、欠損値を代入するために使用されます。

その他の設定:

欠損値の最大パーセンテージ**(%)(X)**

すべての系列に許可される欠損値の最大パーセンテージを指定します。指定した最大数よりも欠損値が多い系列は、分析から除外されます。

ストリーミング TCM ノード - 一般的なデータ オプション

次元フィールドあたりの異なる数値の最大数**(M)**

この設定は、多次元データに適用されるものであり、いずれか一つの次元フィールドに許可される異なる数値の最大数を指定します。デフォルトでは、この制限は 10000 に設定されていますが、必要に応じて大きな数値に増やすことができます。

ストリーミング TCM ノード - 一般的な作成オプション

信頼区間の幅 **(%)(C)**

この設定では、予測およびモデル・パラメーターの両方の信頼区間を制御します。100 未満の正の値を指定できます。デフォルトでは、95% の信頼区間が使用されます。

対象ごとの入力フィールドの最大数**(M)**

この設定では、各対象のモデルで許可される入力の最大数を指定します。1 から 20 までの範囲の整数を指定できます。各対象のモデルには、それ自身の遅れた値が必ず含まれます。そのため、この値を 1 に設定すると、その入力のみが対象自身になることが指定されます。

モデル許容度**(O)**

この設定では、各対象の最適な入力セットを判定するために使用される反復プロセスを制御します。ゼロよりも大きい任意の数を指定できます。デフォルトは 0.001 です。モデル許容度は、予測値の選択の停止基準です。最終モデルに組み込まれる予測値の数に影響する可能性があります。しかし、対象自体がうまく予測することができる場合は、その他の予測値が最終モデルに組み込まれない可能性があります。多少の試行錯誤が必要な場合があります (例えば、この値を大きい値に設定した場合、それよりも小さい値を設定してみることにより、その他の予測値を組み込むことが可能かどうかを判断できます)。

外れ値のしきい値**(%)(T)**

モデルから計算された外れ値の確率がこのしきい値を超えると、観測に外れ値を示すフラグが立てられます。50 から 100 までの範囲の値を指定できます。

各入力のラグ数 **(Number of Lags for Each Input)**

この設定は、各対象のモデル内の各入力のラグ項目数を指定します。デフォルトでは、ラグ項目数は分析に使用される時間間隔から自動的に決定されます。例えば、時間間隔が (月単位の増分である) 月である場合のラグ数は 12 になります。オプションで、明示的にラグ数を指定することもできます。指定する値は、1 から 20 の範囲内の整数でなければなりません。

既存のモデルを使用して推定を続行

時間的因果モデルが既に生成されている場合、新規モデルを作成するのではなく、そのモデルに指定された基準設定を再利用するには、このオプションを選択します。この方法で、以前と同じモデル設定でも最新データを使用してそのモデルに基づいて新しい予測を再推定および作成できるので、時間の節約になります。

ストリーミング TCM ノード - 推定期間オプション

デフォルトでは、推定期間は、最も早い観測の時刻から始まり、すべての系列における最も遅い観測の時刻に終わります。

開始時刻と終了時刻で指定**(B)**

推定期間の開始と終了の両方を指定することも、開始のみまたは終了のみを指定することもできます。推定期間の開始または終了を省略した場合は、デフォルト値が使用されます。

- 観測が「日付/時刻」フィールドによって定義されている場合は、「日付/時刻」フィールドに使用されているのと同じ形式で開始と終了の値を入力します。
- 循環する期間によって観測が定義されている場合は、それぞれの「循環する期間」フィールドの値を指定します。各フィールドは、別々の列に表示されます。

最も遅い時間区分または最も早い時間区分で指定 **(By latest or earliest time intervals)**

データの最も早い時間区分で開始または最も遅い時間区分で終了する時間区分の指定回数として推定期間を定義します。必要に応じてオフセットを指定することができます。このコンテキストでは、時間区分は、分析の時間区分を参照します。例えば、観測は月 1 回であるが、分析の時間区

分は四半期であると想定します。「最新」を指定し、「時間区分の数 **(Number of time intervals)**」の値として 24 を指定することは、最新の 24 個の四半期を意味します。

必要な場合は、指定した数の時間区分を除外することもできます。例えば、最新の 24 個の時間区分を指定し、除外する数値として 1 を指定した場合、推定期間は、最新の区分の前の 24 個の区分から構成されます。

ストリーミング TCM ノード - モデル・オプション

モデル名
: モデルのカスタム名を指定することも、自動生成された名前 (*TCM*) を受け入れることもできます。

予測
: 「レコードの将来への拡張」のオプションで、推定期間の終わりを超えて予測する時間区分の数を設定します。この場合の時間区分は、「データ指定」タブで指定した分析の時間区分です。予測が要求されると、対象でもない任意の入力系列に対して、自己回帰モデルが自動的に構築されます。次にこれらのモデルは、予測期間のこれらの入力系列の値を生成するために使用されます。この設定の最大値制限はありません。

CPLEX の最適化ノード

CPLEX の最適化ノードにより、OPL (Optimization Programming Language) モデル・ファイルを介した複雑な数学 (CPLEX) ベースの最適化の機能が提供されます。この機能は IBM Analytical Decision Management 製品で使用可能ですが、IBM Analytical Decision Management の必要なしに、SPSS Modeler でも CPLEX ノードを使用できるようになりました。

CPLEX の最適化および OPL について詳しくは、IBM ILOG CPLEX Optimization Studio の資料を参照してください。

CPLEX の最適化ノードは、複数のデータ ソース (つまり、多次元入力データ) をサポートします。CPLEX の最適化ノードにはいくつかのノードを接続でき、各事前ノードを使用して OPL モデル計算にデータを提供できます。個別のフィールド マッピングを持つ個別のタプル セットとして設定します。

CPLEX の最適化ノードによって生成されたデータを出力するときに、結果の単一のインデックスまたは多次元インデックスとして、データ ソースからの元のデータを共に出力できます。

注**: IBM SPSS Modeler Server** で CPLEX 最適化ノードを含むストリームを実行している場合、デフォルトでは、組み込みの Community Edition CPLEX ライブラリーが使用されます。この場合、変数は 1000 個まで、制約も 1000 件までに制限されます。完全版の IBM ILOG CPLEX をインストールして、このような制限のない完全版の CPLEX エンジンを代わりに使用したい場合は、ご使用のプラットフォームに応じて以下の手順を実行してください。

- Windows の場合は、`options.cfg` を編集して OPL ライブラリーのパスを追加します。以下に例を示します。

  ```
  cplex_opl_lib_path="<CPLEX_path>¥opl¥bin¥<Platform_dir>"
  ```

  ここで、`<CPLEX_path>` は CPLEX のインストール・ディレクトリー (`C:¥Program Files¥IBM¥ILOG¥CPLEX_Studio127` など) であり、`<Platform_dir>` はプラットフォーム固有のディレクトリー (`x64_win64` など) です。
- Linux の場合は、`modelersrv.sh` を編集して OPL ライブラリーのパスを追加します。以下に例を示します。

  ```
  CPLEX_OPL_LIB_PATH=<CPLEX_path>/opl/bin/<Platform_dir>
  ```

ここで、<CPLEX_path> は CPLEX のインストール・ディレクトリー (/root/Libs_127_FullEdition/Linux_x86_64 など) であり、<Platform_dir> はプラットフォーム固有のディレクトリー (x86-64_linux など) です。

注: **SPSS Modeler Solution Publisher** で CPLEX 最適化ノードを含むストリームを実行している場合、デフォルトでは、組み込みの Community Edition CPLEX ライブラリーが使用されます。この場合、変数は 1000 個まで、制約も 1000 件までに制限されます。完全版の IBM ILOG CPLEX をインストールして、このような制限のない完全版の CPLEX エンジンを代わりに使用したい場合は、ご使用のプラットフォームに応じて以下の手順を実行してください。

- Windows の場合は、modelerrun.exe のコマンド・ライン引数として OPL ライブラリーのパスを追加します。以下に例を示します。

  ```
  -o cplex_opl_lib_path="<CPLEX_path>¥opl¥bin¥<Platform_dir>"
  ```

  ここで、<CPLEX_path> は CPLEX のインストール・ディレクトリー (C:¥Program Files¥IBM¥ILOG¥CPLEX_Studio127 など) であり、<Platform_dir> はプラットフォーム固有のディレクトリー (x64_win64 など) です。
- Linux の場合は、modelerrun を編集して OPL ライブラリーのパスを追加します。以下に例を示します。

  ```
  CPLEX_OPL_LIB_PATH=<CPLEX_path>/opl/bin/<Platform_dir>
  ```

  ここで、<CPLEX_path> は CPLEX のインストール・ディレクトリー (/root/Libs_127_FullEdition/Linux_x86_64 など) であり、<Platform_dir> はプラットフォーム固有のディレクトリー (x86-64_linux など) です。

CPLEX の最適化ノードのオプションの設定

CPLEX の最適化ノードの「オプション」タブには次のフィールドがあります。

OPL モデル ファイル: OPL (Optimization Programming Language) モデル・ファイルを選択します。

OPL モデル: OPL モデルを選択すると、その内容がここに表示されます。

入力データ

「入力データ」タブで、「データ ソース」ドロップダウンには、現在の CPLEX の最適化ノードに接続されたすべてのデータ ソース (事前ノード) がリストされています。ドロップダウンからデータ ソースを選択すると、下の「入力マッピング」セクションが更新されます。「すべてのフィールドの適用 **(Apply All Fields)**」をクリックして、選択されたデータ ソースのすべてのフィールド マッピングを自動的に生成します。「入力マッピング」テーブルにデータが自動的に設定されます。

入力データに対応する、OPL モデルのタプル セット名を入力します。次に、必要に応じて、タプル定義の順序に従って、すべてのタプル フィールドがデータ入力フィールドにマップされていることを確認します。

データ ソースの入力マッピングを設定した後で、ドロップダウンから別のデータ ソースを選択し、プロセスを繰り返すことができます。以前のデータ ソース マッピングは自動的に保存されます。終了したら、「適用」または「**OK**」をクリックします。

その他のデータ

最適化について他に指定する必要があるデータがある場合は、「その他のデータ」タブの「**OPL** データ」セクションを使用します。

出力

出力が決定変数である場合、決定変数は事前データ ソース (入力データ) をインデックスとして受け入れる必要があり、それらのインデックスを「入力データ」タブの「入力マッピング」セクションで事前定義する必要があります。その他のタイプの決定変数は現在サポートされていません。決定変数は単一のインデックスまたは複数のインデックスを受け入れることができます。SPSS Modeler は、CPLEX 結果と共に、元の入力データのすべてまたは一部を出力します。これは、他の SPSS Modeler ノードと整合性を持ちます。参照先の対応するインデックスは、下で説明されている「出力タプル」フィールドで指定される必要があります。

「出力」タブで出力モード(「未加工出力」または「決定変数」) を選択し、必要に応じて他のオプションを指定します。「未加工出力」オプションの場合、名前に関係なく、目的関数が直接出力されます。

OPL での目的関数値の変数名**:** このフィールドは、「決定変数」出力モードを選択した場合に有効になります。OPL モデルの目的関数の値変数の名前を入力してください。

出力に使用する目的関数値のフィールド名**:** 出力に使用するフィールド名を入力します。デフォルトは `_OBJECTIVE` です。

出力タプル**:** 入力データからの事前定義されたタプルの名前を入力します。これは、決定変数のインデックスとしての役割を果たし、「変数の出力」で出力されることが予期されます。「出力タプル」は、OPL 内の決定変数定義と整合性を持つ必要があります。複数のインデックスがある場合、タプル名は、コンマ (**,**) で結合される必要があります。

変数の出力**:** 出力に含める 1 つ以上の変数を追加します。

第 4 章 フィールド設定ノード

フィールド設定の概要

初期データを探索した後に、解析の準備として、データの選択、クリーニング、または構成が必要になるでしょう。「フィールド設定」パレットには、このデータ変換と準備作業に役立つさまざまなノードが用意されています。

例えば、フィールド作成ノードを使用して、現在データには含まれていない属性を新しく作成できます。または、例えばデータ分割ノードを使用して、対象の分析用に自動的にフィールド値を再コード化することができます。データ型ノードは、使用頻度の高いノードです。このノードを使用すると、測定の尺度、値、モデリングの役割をデータ・セット内の各フィールドに割り当てることができます。この操作は、欠損値の処理と下流のモデル作成において役に立ちます。

「フィールド設定」パレットには次のようなノードがあります。

自動データ準備 (ADP) ノードでは、データ分析、固定値の識別、問題のあるまたは役に立たない可能性のあるフィールドのスクリーニング、必要に応じた新しい属性の取得、詳細なスクリーニングおよびサンプリング手法を使用したパフォーマンスの向上などを行うことができます。完全に自動化された方法でノードを使用し、ノードで固定値を選択および適用できます。または必要に応じて変更の作成および承認、拒否または修正の前に変更をプレビューできます。

データ型ノードで、フィールドのメタデータとプロパティーを指定します。例えば、各フィールドに、測定の尺度 (連続型、名義型、順序型、またはフラグ) を指定し、欠損値とシステムヌルの処理のためのオプションを設定し、モデル作成の目的に対するフィールドの役割を設定し、フィールドと値のラベルを指定し、フィールドの値を指定します。

フィルター ノードは、フィールドのフィルター操作（破棄）、フィールド名の変更、またはソース ノードから別のノードへのフィールドのマップを行います。

フィールド作成ノードで、1 つまたは複数の既存フィールドから、データ値を変更するか、新しいフィールドを作成します。これで、 タイプ式、フラグ、名義、ステート、カウント、および条件式の各フィールドが作成されます。

アンサンブル・ノードでは、2 つまたはそれ以上のモデル・ナゲットを組み合わせて 1 つのモデルよりもより正確な予測を取得します。

置換ノードで、フィールド値の置換やストレージの変更を行います。`@BLANK(@FIELD)` のような、CLEM 条件に基づいて値を置換することができます。また、すべての空白値やヌル値を特定の値に置換することもできます。置換ノードは、データ型ノードと一緒に使用される場合が多く、欠損値の置き換えが行われます。

匿名化ノードは、フィールド名や値の下流の表示方法を変換し、元のデータを隠します。これは、他のユーザーが顧客名やその他の詳細情報などの機密データを使用してモデルを構築できるようにする場合に有用です。

データ分類ノードにより、あるカテゴリー値のセットが別のセットに変換されます。データ分類は、カテゴリーを再編成したり、分析用のデータをグループ化しなおす場合に役立ちます。

データ分割ノードで、既存の 1 つまたは複数の連続型 (数値範囲) フィールドの値に基づいて、自動的に新しい名義型 (セット型) フィールドを作成します。例えば、連続型収入フィールドを、平均からの偏差による収入グループを含む、新しいカテゴリー・フィールドに変換することができます。新規フィールドのビンを作成すると、分割点に基づいてフィールド作成ノードを生成することができます。

リーセンシ、フリクエンシ、マネタリー (RFM) の分析ノードを使用すると、最後に購入したのがどのくらい最近か (リーセンシ)、どのくらい頻繁に購入するか (フリクエンシ)、トランザクション全体でいくら消費したか (マネタリー) を検証することによって、最も良い顧客となると考えられるのはどの顧客かを量的に決定することができます。

データ区分ノードで、モデル構築の学習、テスト、および検証の各ステージ用に、データを独立したサブセットに分割するデータ区分フィールドが生成されます。

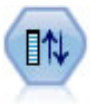
フラグ設定ノードで、1 つ以上の名義型フィールドに定義されたカテゴリー値に基づいた、複数のフラグ型フィールドが派生します。

再構成ノードで、名義型またはグラグ型フィールドを、これから別のフィールドの値で埋めることができるフィールドのグループへ変換します。例えば、*credit*、*cash*、および *debit* の値の *payment type* という名前のフィールドがある場合、3 つの新しいフィールド (*credit*、*cash*、*debit*) が作成されます。その各々には、実際の支払の値を含めることができます。

行列入替ノードで、レコードがフィールドになり、フィールドがレコードになるように、行内と列内のデータを交換します。

間隔を指定し、新しい時間フィールドを作成して推定や予測を行う場合は、時間区分ノードを使用してください。秒単位から年単位まで、すべての時間区分がサポートされます。

時系列ノードにより、以前レコードのフィールドのデータを含む、新規フィールドが作成されます。時系列ノードは、多くの場合、時系列データなどの継続的なデータに使用されます。時系列ノードを使用する前に、ソート・ノードを使用して、データをソートしておくこともできます。



フィールド順序ノードで、下流のフィールド表示に使用する順序を定義します。この順序は、テーブル、リスト、およびフィールド・ピッカーなど、さまざまな場所のフィールドの表示に適用されます。この操作は、さまざまなデータセットにおいて、特定のフィールドをより参照しやすくする場合に役立ちます。

SPSS Modeler では、式ビルダーの空間処理関数、時空間予測 (STP) ノード、マップ視覚化ノードなどの項目は、投影座標系を使用します。地理座標系を使用するインポート データの座標系を変更するには、投影ノードを使用してください。

これらのノードの大半は、データ検査ノードが作成する検査レポートから直接生成することができます。詳しくは、トピック 342 ページの『データ準備用のその他のノードの生成』を参照してください。

データの準備の自動化

分析用にデータを準備することは、どのようなプロジェクトにおいても最も重要な段階の 1 つですが、従来は最も時間がかかる段階の 1 つでもありました。自動データ準備 (ADP) は、データ分析および修正の特定、問題となる、または有用でないと考えられるフィールドの除外、必要に応じた新しい属性の取得、高度なスクリーニング手法を用いたパフォーマンスの改善を行い、タスクを処理します。完全に自動化された方法でアルゴリズムを使用し、そのアルゴリズムによって修正を選択したり適用することができます。または、インタラクティブな方法でそのアルゴリズムを使用して適用前の変更内容をプレビューし、必要に応じてその変更内容を承認するか拒否するかを選択することもできます。

ADP を使用すると、実行する統計の概念の事前情報を必要とせず、モデルを迅速かつ用意に作成できるよう、データを準備することができます。モデルはより迅速に構築およびスコアリングするようになります。また、ADP を使用すると、モデル更新やチャンピオン/チャレンジャーなど、自動モデル作成プロセスの強固さをより向上させます。

注**:** ADP で分析用のフィールドを準備する場合、古いフィールドの既存の値およびプロパティーを置き換えるのではなく、調整または変換を含む新しいフィールドを作成します。古いフィールドは高度な分析には使用されません。

例: 世帯主の保険請求を調査するためのリソースが制限されている保険会社が、不正請求の疑いのある請求を区別するためのモデルを作成したいと考えています。モデルを作成する前に、自動データ準備を使用して、モデル作成のためのデータを準備します。変換が適用される前に提案される変換を確認できる必要があるため、自動データ準備をインタラクティブ・モードで使用します。

自動車産業グループは、さまざまな個人用自動車の売り上げを記録します。採算ベースを上回るモデルおよび下回るモデルを特定できるように、自動車の売り上げと自動車の特性との関係を確立したいと考えます。自動データ準備を使用して分析用のデータを準備し、準備「前」および準備「後」のデータを使用してモデルを作成し、結果がどのように異なるかを確認します。

目的は何ですか**?:** 自動データ準備では、他のアルゴリズムがモデルを作成する速度に影響し、それらのモデルの予測精度を改善するデータ準備ステップが推奨されます。これには、特徴の変換、構成、および選択が含まれます。対象を変換することもできます。データ準備プロセスで重点を置く必要があるモデル作成の優先度を指定できます。

- 速度と精度のバランス**:** このオプションでは、モデル作成アルゴリズムによるデータ処理の速度と、予測精度の両方を同等に優先するように、データを準備します。
- 速度の最適化**:** このオプションでは、モデル作成アルゴリズムによるデータ処理の速度を優先するよう、データを準備します。大きいデータ・セットを処理する場合、または迅速な回答を求めている場合は、このオプションを選択します。
- 精度の最適化**:** このオプションでは、モデル作成アルゴリズムによって生成される予測の精度を優先するよう、データを準備します。
- カスタム分析**:** 「設定」タブでアルゴリズムを手動で修正する場合、このオプションを選択します。継続して「設定」タブのオプションに変更を行うも、その他の目的と互換性がない場合、この設定が自動的に選択されます。

ノードの学習

ADP ノードはプロセス・ノードとして実装され、データ型ノードと同じように機能します。ADP ノードの学習は、データ型ノードのインスタンス化に対応しています。分析が実行されると、上流データが変更されない限り、高度な分析を行わずに指定された変換がデータに適用されます。データ型ノードやフィルター・ノード同様、ADP ノードの接続が解除されても、データ・モデルや変換は記憶され、再接続された場合に再度学習する必要がなくなります。必要に応じて、一般データのサブセットのデータ・モデルについて学習し、実データに使用するためにコピーまたは展開することができます。

ツールバーの使用

ツールバーを使用すると、データ分析の表示を実行および更新し、元のデータと組み合わせて使用できるノードを生成できます。

- ノードの生成 フィルター・ノードまたはフィールド生成ノードを生成できます。このメニューは、分析が「分析」タブに表示されている場合にのみ使用できます。

  フィルター・ノードは、変換された入力 フィールドを削除します。データ セットに元の入力フィールドを残すように ADP ノードを設定すると、元の入力フィールドのセットが復元されるので、入力フィールドに関連してスコア フィールドを解釈できます。例えば、これはさまざまな入力に対してスコア フィールドのグラフを生成したい場合に役立ちます。

フィールド生成ノードは、元のデータ・セットと目標の単位を復元します。ADP ノードに範囲型目標を再調整する分析が含まれている場合 (「入力および目標の順位」パネルで Box-Cox 再調整を選択) のみ、フィールド生成ノードを生成できます。目標が範囲型でない場合、または Box-Cox 再調整が選択されていない場合、フィールド生成ノードは生成できません。詳しくは、トピック 143 ページの『フィールド生成ノードの生成』を参照してください。

- 表示「分析」タブに表示される項目を制御するオプションが用意されています。ここには、グラフ編集コントロールや、メイン・パネルおよびリンク ビュー双方の表示選択が用意されています。
- プレビュー 入力データに適用される変換のサンプルが表示されます。
- データの分析 現在の設定を使用して分析を開始し、「分析」タブに結果を表示します。
- 分析のクリア 既存の分析を削除します (現在の分析が存在する場合にのみ使用できます)。

ノードの状態

矢印、または分析が行われたかどうかを示すアイコンをクリックすることによって、IBM SPSS Modeler 領域に ADP ノードの状態が示されます。

データの準備の自動化ノードで実行される計算について詳しくは、「*IBM SPSS Modeler* アルゴリズム・ガイド」の『*Automated Data Preparation Algorithms*』セクションを参照してください。このガイドは、PDF 形式で用意されており、製品ダウンロードの一部として、インストール・ディスクの `¥Documentation` ディレクトリーから入手するか、Web で入手できます。

「フィールド」タブ

モデルを作成する前に、対象フィールドや入力フィールドを指定する必要があります。いくつかの例外を除いて、すべてのモデル・ノードは、上流のデータ型ノードからのフィールド情報を使用します。データ型ノードを使用して入力フィールドおよび対象フィールドを選択する場合、このタブで何も変更する必要はありません。

データ型ノードの設定を使用**:** このオプションを選択すると、上流のデータ型ノードからのフィールド情報がこのノードで使用されます。これがデフォルトです。

ユーザー設定を使用**:** このオプションを選択すると、上流のデータ型ノードからのフィールド情報ではなく、ここで指定したフィールド情報がこのノードで使用されます。このオプションを選択した後に、必要に応じて以下のフィールドを指定します。

目標**:** 1 つまたは複数の対象フィールドが必要なモデルの場合に、対象フィールドを選択します。これは、データ型ノードのフィールドの役割を 「対象」 に設定するのと似ています。

入力**:** 1 つ以上の入力フィールドを選択します。これは、データ型ノードのフィールドの役割を 「入力」 に設定するのと似ています。

「設定」タブ

「設定」タブは、アルゴリズムがデータをどのように処理するかを調整するために変更できる、複数グループの設定で構成されています。他の目的と互換性のない変更をデフォルト設定に対して行うと、「目的」タブが自動的に更新され、「分析のカスタマイズ」オプションを選択します。

フィールド設定

度数フィールドを使用: このオプションにより、フィールドを出現頻度の重みとして選択することができます。集計データを使用している場合など、学習データのレコードがそれぞれ複数の単位を示す場合に使用します。フィールド値は、レコードごとに示した単位数です。

重みフィールドを使用: フィールドをケースの重みとして選択できます。ケースの重みを使用して、出力フィールドのレベル間の分散における相違を処理します。

モデル作成から除外されたフィールドの処理方法**:** 除外されたフィールドの処理方法を指定します。データを除外するか、「役割」 を 「なし」 に設定するかを選択できます。

注 : この操作は、目標フィールドが変換される場合、目標フィールドにも適用されます。例えば、新しく派生したバージョンの対象が 「対象」 フィールドとして使用されている場合、元の対象は除外されるか、「なし」 に設定されます。

受信フィールドが既存の分析と一致しない場合。学習済み ADP ノードを実行するとき 1 つまたは複数の必須入力フィールドが受信データ・セットにない場合の処理方法を指定します。

- 実行を停止して既存の分析を保存する。実行プロセスを停止し、現在の分析情報を保存して、エラーを表示します。
- 既存の分析をクリアして新しいデータを分析する。既存の分析をクリアし、受信データを分析して、推奨されている変換をそのデータに適用します。

日時の準備

多くのモデル作成アルゴリズムは、日付や時刻の詳細を直接処理することはできません。これらの設定を使用して、既存データの日付および時刻から、モデル入力として使用できる新しい期間データを取得できます。日付および時刻を含むフィールドは、日付または時間のストレージ・タイプで事前定義する必要があります。元の日付および時間フィールドは、自動データ準備に従うモデル入力としては推奨されません。

モデリング用の日時を準備**:** このオプションの選択を解除すると、他のすべての「日時の準備」コントロールが無効になりますが、選択状態は維持されます。

基準日からの経過時間を計算**:** 日付を含む各変数の基準日以降の年/月/日の数を生成します。

- 基準日**:** 入力データの日付情報に関して、計算する期間の開始日を指定します。「今日の日付」を選択すると、ADP を実行するときに常に現在のシステムの日付が使用されます。特定の日付を使用するには、「固定日付」 を選択して、該当する日付を入力します。ノードが初めて作成された場合、「固定日付」 フィールドには、現在の日付が自動的に入力されます。
- 期間 **(**日数**)** の単位**:** 期間の単位を ADP により自動的に決定するか、または「固定単位」(「年」、「月」、または「日」) から選択するかを指定します。

基準時刻からの経過時間を計算**:** 時刻を含む各変数の基準時刻以降の時間/分/秒の数を生成します。

- 基準時刻**:** 入力データの時刻情報に関して、計算する期間の開始時刻を指定します。「現在の時刻」を選択すると、ADP が実行されている場合、現在のシステムの時刻が常に使用されます。特定の時刻を使用するには、「固定時刻」 を選択して、該当する時刻を入力します。ノードが初めて作成された場合、「固定時刻」 フィールドには、現在の時刻が自動的に入力されます。
- 期間 **(**時間数**)** の単位**:** 期間の単位を ADP により自動的に決定するか、または「固定単位」(「時間」、「分」、または「秒」) から選択するかを指定します。

サイクル時間要素を抽出: これらの設定を使用して、1 つの日付または時刻フィールドを 1 つ以上のフィールドに分割します。例えば、3 つの日付のチェック ボックスをすべてオンにすると、入力日付フィールド「1954-05-23」は 1954、5、および 23 の 3 つのフィールドに分割されます。それぞれ 「フィールド名」 パネルで定義された接尾辞を使用すると、元の日付フィールドは無視されます。

- 日付から取得: 日付入力について、年、月、日付またはそれらの組み合わせを抽出するかどうかを指定します。
- 時刻から取得: 時刻入力について、時間、分、秒またはそれらの組み合わせを抽出するかどうかを指定します。

フィールドの除外

品質の悪いデータは、予測の精度に影響を与える場合があります。そのため、入力フィールドに適切な品質レベルを指定することができます。定数または 100% 欠損値であるすべてのフィールドは、自動的に除外されます。

低品質の入力フィールドを除外: このオプションを選択解除すると、選択内容を維持した状態で「フィールドの除外」のその他のすべてのコントロールが無効になります。

欠損値の多いフィールドの除外: 欠損値の割合が指定されている割合を超えているフィールドが、以降の分析から除外されます。 0 以上 100 以下の値を指定しますが (0 はオプションの選択解除を示す)、すべての欠損値を含むフィールドは自動的に除外されます。デフォルトは 50 です。

一意のカテゴリーが多すぎる名義型フィールドを除外: カテゴリー数が指定された数を超えている名義型フィールドが、以降の分析から除外されます。正の整数を指定してください。デフォルトは 100 です。ID、住所、名前などのモデル作成からレコード特有の情報を含むフィールドを自動的に削除する場合に役立ちます。

単一カテゴリーの値が多いカテゴリー・フィールドの除外: 1 つのカテゴリーに指定されている割合を超えるレコードが含まれている順序型フィールドおよび名義型フィールドが、以降の分析から除外されます。0 以上 100 以下の値を指定しますが (0 はオプションの選択解除を示す)、定数フィールドは自動的に除外されます。デフォルトは 95 です。

入力フィールドおよび目標フィールドの準備

処理しているデータで完全な状態のものがないため、分析を実行する前にいくつかの設定を調整する必要があります。例えば、外れ値の削除、欠損値の処理方法の指定、タイプの調整などです。

注 : このパネルで値を変更した場合、「目的」 タブが自動的に更新され、「カスタム分析」 オプションが選択されます。

モデル作成の入力フィールドの目標フィールドを準備する: パネルのすべてのフィールドをオンまたはオフに切り替えます。

タイプを調整してデータ品質を改善する: 入力フィールドおよび目標フィールドについて、それぞれいくつかのデータ変換を指定できます。目標フィールドの値を変更する必要がないためです。例えば、収入の予測 (単位: ドル) は、ログで測定される予測よりもより意味があります。また、目標フィールドに欠損値がある場合、欠損値を入力する予測ゲインがないため、入力フィールドに欠損値を入力すると、一部のアルゴリズムは、欠損する情報を処理することができます。

外れ値の分割値など、これらの変換の追加設定は、目標フィールドと入力フィールドに共通しています。

入力フィールドと目標フィールドのいずれか、または両方に次の設定を選択できます。

- 数値型フィールドのタイプを調整する: 測定の尺度が順序型の数値型フィールドを連続型に変換できるかどうか、またはその逆を指定します。最小しきい値および最大しきい値を指定して、変換を制御できます。
- 名義型フィールドの並べ替え: 数値型 (セット型) フィールドを最小カテゴリーから最大カテゴリーの順に並べ替えます。
- 連続型フィールドの外れ値を置換: 外れ値を置き換えるかどうかを指定します。次の 「外れ値の置換方法」 オプションと組み合わせて使用します。
- 連続型フィールド: 欠損値を平均値に置換: 連続型 (範囲型) フィールドの欠損値を置き換えます。
- 名義フィールド: 欠損値を最頻値に置換: 名義型 (セット型) フィールドの欠損値を置き換えます。
- 順序型フィールド: 欠損値を中央値に置換: 順序型 (順序セット型) フィールドの欠損値を置き換えます。

>順序フィールドの値の最大数: 順序型 (順序セット型) フィールドを連続型 (範囲型) として再定義する際の閾値を指定します。デフォルトは 10 です。そのため、順序型フィールドに 10 を超えるカテゴリーがある場合、連続型 (範囲型) に再定義されます。

連続型フィールドの値の最小数: 尺度または連続型 (範囲型) フィールドを順序型 (順序セット型) として再定義する際の閾値を指定します。デフォルトは 5 です。そのため、連続型フィールドに 5 を超えるカテゴリーがある場合、順序型 (順序セット型) に再定義されます。

外れ値の分割値: 標準偏差で測定される、外れ値の分割基準を指定します。デフォルトは 3 です。

外れ値の置換方法: 外れ値を、分割値でトリム化 (coerce) して置き換えるか、外れ値を削除して欠損値として設定するかを選択します。欠損値に設定した外れ値は、上記で選択された欠損値処理の設定にしたがって処理されます。

すべての連続型入力フィールドを共通尺度に設定: 連続型入力フィールドを正規化するには、このチェック・ボックスをオンにして、正規化方法を選択します。デフォルトは 「**z** スコア変換」 で、デフォルトが 0 の 「最終平均値」、デフォルトが 1 の 「最終標準偏差」 を指定できます。また、「最小値/最大値の変換」 を使用して最小値と最大値を指定できます (デフォルトはそれぞれ 0 および 100 です)。

このフィールドは、「変数の構築と選択」パネルで「変数構築の実行」を選択する場合に特に便利です。

Box-Cox 変換で連続型目標を再調整する: 連続型 (スケールまたは範囲型) 目標フィールドを正規化するには、このチェック・ボックスをオンにします。Box-Cox 変換は、「最終平均値」 のデフォルト値が 0、「最終標準偏差」 のデフォルト値が 1 です。

注 : 目標を正規化する場合、目標の次元が変換されます。この場合、フィールド生成ノードを生成して、逆変換を適用し、変換された単位を、高度な処理で認識可能な形式に戻す必要があります。 詳しくは、トピック 143 ページの『フィールド生成ノードの生成』を参照してください。

構築および変数選択

データの予測精度を向上させるために、入力フィールドを変換したり、既存フィールドに基づいて新しいフィールドを構築できます。

注 : このパネルで値を変更した場合、「目的」 タブが自動的に更新され、「カスタム分析」 オプションが選択されます。

入力フィールドを変換、構築および選択して予測制度を改善する: パネルのすべてのフィールドをオンまたはオフに切り替えます。

まばらなカテゴリーを結合して目標との関連性を最大化: 目標と関連して処理する変数の数を減らして、より節約的なモデルを作成します。必要に応じて、デフォルト値の 0.05 から確率値を変更します。

すべてのカテゴリーが 1 つのカテゴリーに結合される場合、予測値としての値がないため、元のバージョンのフィールドおよび派生した化されたフィールドは除外されます。

目標がない場合、度数に基づいてまばらなカテゴリーを結合する: 目標フィールドがないデータを処理する場合、順序型 (順序セット型) フィールドおよび名義型 (セット型) フィールドのいずれか、または両方のまばらなカテゴリーを結合できます。結合するカテゴリーを特定するデータのケースまたはレコードの最小パーセントを指定します。デフォルトは 10 です。

カテゴリーは、次の規則に従って結合されます。

- 2 値フィールドの結合は実行しない。
- 結合時にカテゴリーが 2 つしかない場合、結合は停止する。
- 元のカテゴリーがない場合、または結合時にカテゴリーが作成されない場合、指定したケースの最小パーセントより少ない場合、結合は停止する。

予測精度を保持しながら連続型フィールドを分割: データにカテゴリー型目標が含まれている場合、強い関連を持つ連続型入力フィールドを分割して、処理のパフォーマンスを向上させることができます。必要に応じて、等質サブグループの確率値をデフォルト値の 0.05 から変更します。

カテゴリー化操作によって特定フィールドに単一ビンが生成される場合、予測値としての値がないため、元のバージョンのフィールドおよびカテゴリー化されたフィールドは除外されます。

注 : IBM SPSS Modeler のその他の部分で使用される ADP のカテゴリー化は最適カテゴリー化とは異なります。最適カテゴリー化では、エントロピー情報を使用して、連続型変数をカテゴリー変数に変換します。最適カテゴリー化では、データを並べ替え、メモリ内にすべて保存する必要があります。ADP では、等質サブグループを使用して、連続型変数を分割します。ADP カテゴリー化では、データを並べ替え、メモリ内にすべて保存する必要はありません。等質サブグループの方法を使用して連続型変数をカテゴリー化すると、カテゴリー化したあとのカテゴリー数は、常に目標内のカテゴリー数と等しいか少なくなります。

変数選択の実行: 相関係数が低いフィールドを削除します。必要に応じて、デフォルト値の 0.05 から確率値を変更します。

このオプションは、目標が連続型の連続型入力フィールドと、カテゴリー型入力フィールドに適用されます。

変数構築の実行: 複数の既存フィールドの組み合わせから新しいフィールドを取得します (既存フィールドはモデル作成から削除されます)。

このオプションは、目標が連続型の場合または目標がない場合にのみ、連続型入力フィールドに適用されます。

フィールド名

新しいフィールドや変換されたフィールドを用意に特定できるようにするために、ADP は新しい基本名、接頭辞または接尾辞を作成し、適用します。それらの名前を修正して、ニーズおよびデータにより関連付けることができます。別のラベルを指定する場合は、下流のデータ型ノードで指定する必要があります。

変換され構築されたフィールド: 変換された目標フィールドおよび入力フィールドの適用する名前の拡張子を指定します。

ADP ノードでは、何も入力されない文字列フィールドを設定すると、未使用フィールドの処理方法によってはエラーが発生する場合があります。「設定」タブの「フィールド設定」パネルで、「モデル作成から除外されたフィールドの処理方法」 が 「未使用フィールドを除外」 に設定された場合、入力フィールドおよび目標フィールドの名前の拡張子を設定しないこともできます。元のフィールドが除外され、変換されたフィールドが上書き保存されます。この場合、新しく変換されたフィールドは元のフィールドと同じ名前がつけられます。

ただし、「未使用フィールドの方向を「なし」に設定」を選択した場合、対象の名前と入力の名前の拡張子を空 (ヌル) にすると、重複したフィールド名が作成されるため、エラーが発生します。

さらに、「選択および構築」設定を使用して、構築されるフィールドに適用する接頭辞名を指定します。数値の接尾辞をこの接頭辞のルート名に追加して、新しい名前を作成します。番号の形式は、次のように、取得された新しいフィールドの数によって異なります。

- 1 から 9 の構築済みフィールドの名前は、feature1 から feature9 です。
- 10 から 99 の構築済みフィールドの名前は、feature01 から feature99 です。
- 100 から 999 の構築済みフィールドの名前は、feature001 から feature999 です。

これにより、構築されたフィールドは、フィールド数に関係なく、合理的な順序で並べ替えられます。

日時から計算した期間: 日付および時刻の両方から計算された期間に適用する名前の拡張子を指定します。

日時から抽出したサイクル要素: 日付および時刻の両方から抽出したサイクル要素に適用する名前の拡張子を指定します。

「分析」タブ

1. 「目的」タブ、「フィールド」タブ、「設定」タブで行った変更など、ADP 設定に問題がない場合、「データの分析」 をクリックしてください。アルゴリズムにより設定がデータ入力に適用され、「分析」タブに結果が表示されます。

「分析」タブには、データの処理の概要を示すテーブル形式の出力およびグラフィック出力が含まれ、スコアリング用のデータをどのように修正または改善するかについての推奨事項が表示されます。これらの推奨事項を確認し、承認したり拒否したりすることができます。

「分析」タブは 2 つのパネルで構成されています。左側はメイン・ビュー、右側はリンク ビューまたは補助ビューです。メイン・ビューには、次の 3 種類があります。

- 処理したフィールドの要約 (デフォルト)。 詳しくは、トピック 137 ページの『処理したフィールドの要約』を参照してください。
- フィールド : 詳しくは、トピック 137 ページの『フィールド』を参照してください。
- アクションの概要。 詳しくは、トピック 139 ページの『アクションの概要』を参照してください。

リンク/補助ビューには、以下の 4 つがあります。

- 予測の精度 (デフォルト)。 詳しくは、トピック 139 ページの『予測精度』を参照してください。
- フィールド・テーブル。 詳しくは、トピック 139 ページの『「フィールド」テーブル』を参照してください。
- フィールド詳細。 詳しくは、トピック 140 ページの『フィールドの詳細』を参照してください。
- アクションの詳細。 詳しくは、トピック 141 ページの『アクションの詳細』を参照してください。

ビュー間のリンク

メイン・ビューで、表内の下線付きテキストは、リンク ビューの表示を制御します。テキストをクリックすると、特定のフィールド、一連のフィールドまたは処理中のステップに関する詳細を取得できます。最後に選択したリンクは濃い色で表示されます。これにより、2 つのビュー・パネルのコンテンツ間の接続を特定できます。

ビューのリセット

元の分析に関する推奨事項を再度表示し、「分析」ビューに行った変更を取り消す場合、メイン・ビュー・パネルの一番下にある 「リセット」 をクリックしてください。

処理したフィールドの要約

「処理したフィールドの要約」表には、フィールドの状態や構築フィールド数への変更など、処理に対する全体の影響の射影したスナップショットが表示されます。

モデルは実際に構築されていないため、データ準備の前後に予測精度船体の変更に対する測定またはグラフはありません。その代わり、推奨された各予測の予測精度についてのグラフを表示できます。

表には、次の情報が表示されます。

- 目標フィールド数。
- 元 (入力) 予測フィールド数。
- 分析およびモデル作成に使用が推奨される予測値。ここには、推奨フィールド数の合計、元の (変換されていない) 推奨フィールド数、変換された推奨フィールド数 (中間バージョンのフィールド、日付/時刻フィールドから算出したフィールド、構築されたフィールドは除外)、日付/時刻フィールドから算出した推奨フィールド数、構築された推奨予測フィールド数が表示されます。
- 元の形式でも、派生フィールドとしても、あるいは構築された予測値に対する入力としても、いかなる形式でも使用が推奨されない入力予測値の数。

「フィールド」 情報に下線がある場合、クリックするとリンク ビューに詳細が表示されます。「目標」、「入力フィールド」、および 「未使用の入力フィールド」 の詳細は、「フィールド・テーブル」リンク ビューに表示されます。詳しくは、 139 ページの『「フィールド」テーブル』を参照してください。「分析の使用が推奨されるフィールド」は、「予測精度」リンク ビューに表示されます。詳しくは、トピック 139 ページの『予測精度』を参照してください。

フィールド

「フィールド」メイン・ビューには、処理済みフィールドと、ADP が下流モデルにそれらのフィールドの使用を推奨するかどうかを表示します。任意のフィールドについての推奨事項を上書きできます。例えば、構築済みフィールドを除外する、または ADP が除外を推奨するフィールドを追加するなどです。フィールドが変換された場合、推奨された変換を受け入れるか、元のバージョンを使用するかを決定できます。

「フィールド」ビューは、2 つのテーブルで構成されています。1 つは目標フィールドについてのテーブル、もう 1 つは処理されたまたは作成された予測フィールドについてのテーブルです。

「目標」テーブル

「目標」 テーブルには、目標がデータに定義されているかどうかだけが表示されます。

テーブルには、次の 2 つの列があります。

- 「名前」。フィールドが変換された場合でも、元の名前が常に使用されます。

- 測定の尺度: 測定の尺度を示すアイコンが表示されます。マウス・ポインタをアイコンの上に停止させると、データについて説明するラベル (連続型、順序型、名義型など) が表示されます。

  目標が変換されると、「尺度」 列には、最終的な変換バージョンが反映されます。注 : 目標の変換をオフにすることはできません。

予測値テーブル

「予測値」 テーブルが常に表示されます。テーブルの各行は、フィールドを示します。デフォルトでは、行は予測精度の高い順に並んでいます。

通常のフィールドの場合、元の名前は常に行の名前として使用されます。元のバージョンおよび派生バージョンの日付/時刻フィールドがテーブルの各行に表示されます。また、テーブルには構築済み予測フィールドも表示されます。

テーブルに表示される変換されたバージョンのフィールドは、常に最終バージョンを示します。

デフォルトでは、推奨されたフィールドのみが、「予測値」テーブルに表示されます。残りのフィールドを表示するには、テーブルの上にある 「テーブルに非推奨フィールドを含む」 ボックスを選択します。これらのフィールドは、テーブルの一番下に表示されます。

テーブルには、次の列が表示されます。

- 使用バージョン: フィールドを下流で使用するかどうか、および推奨される変換を使用するかどうかを制御するドロップダウン・リストが表示されます。デフォルトでは、ドロップダウン・リストには推奨事項が反映されます。

  変換された通常の予測値の場合、ドロップダウン・リストには 「変換済み」、「変換前」、「使用しない」 の 3 つの選択肢があります。

  変換されていない通常の予測値の場合、「変換前」 および 「使用しない」 の選択肢があります。

  派生した日付/時刻フィールドおよび構築済み予測フィールドの場合、「変換済み」 および 「使用しない」 の選択肢があります。

  元の日付フィールドの場合、ドロップダウン・リストは無効となり、「使用しない」 に設定されます。

  注 : 変換前バージョンと変換済みバージョンの両方の予測フィールドの場合、「変換前」 と 「変換済み」 でバージョンを変更すると、自動的にそれらのフィールドの 「尺度」 および 「予測精度」 の設定が更新されます。
- 「名前」。各フィールドの名前はリンクになっています。名前をクリックすると、フィールドに関する詳細情報がリンク ビューに表示されます。 詳しくは、トピック 140 ページの『フィールドの詳細』を参照してください。
- 測定の尺度: データ型を示すアイコンが表示されます。マウス・ポインタをアイコンの上に停止させると、データについて説明するラベル (連続型、順序型、名義型など) が表示されます。
- 予測精度: ADP が推奨するフィールドに対してのみ予測精度が表示されます。この列は、目標が定義されている場合に表示されます。予測精度は 0 ～ 1 で、値が大きいほど、予測精度が「良い」ことを示します。一般的に、予測精度は ADP 分析の予測を比較するのに役立ちますが、予測精度の値を分析間で比較することはできません。

アクションの概要

自動データ準備で実行された各アクションについて、入力予測フィールドは変換および/または除外されます。ステップを通過したフィールドは、次のステップで使用されます。最後のステップまで通過したフィールドがモデル作成に推奨されます。変換された入力予測フィールドおよび構築されたフィールドは除外されます。

アクションの概要は、ADP で実行された処理のアクションが表示された、単純な表です。「アクション」に下線がある場合、クリックすると実行された操作の詳細がリンク ビューに表示されます。 詳しくは、トピック 141 ページの『アクションの詳細』を参照してください。

注 : 元のバージョンおよび最終変換されたバージョンのフィールドのみが表示され、分析中に使用された中間バージョンのフィールドは表示されません。

予測精度

デフォルトでは、分析が初めて実行された場合に、または「ファイル処理の要約」ビューで 「分析での使用が推奨された予測値」 を選択した場合に表示され、図用には推奨予測フィールドの予測精度が表示されます。フィールドは、予測精度によって並べ替えられ、値が最も大きいフィールドが最上位に表示されます。

変換されたバージョンの通常の予測フィールドの場合、フィールドは「設定」タブの「フィールド名」パネルでの *_transformed* など接尾辞の選択内容を反映します。

各フィールド名の後に、尺度を示すアイコンが表示されます。

推奨される各予測の予測精度は、目標が連続型かカテゴリー型かに応じて、線型回帰モデルまたは naïve Bayes モデルのいずれかから計算されます。

「フィールド」テーブル

「処理したフィールドの要約」メイン・ビューで 「目標」、「予測値」、「未使用の予測値」 をクリックすると表示され、「フィールド表」ビューには関連するフィールドを示す単純なテーブルが表示されます。

テーブルには、次の 2 つの列があります。

- 「名前」。予測の名前。

  目標フィールドの場合、目標が変換されている場合でも、フィールドの元の名前またはラベルが使用されます。

  変換されたバージョンの通常の予測フィールドの場合、フィールド名は「設定」タブの「フィールド名」パネルでの *_transformed* など接尾辞の選択内容を反映します。

  日付と時刻から算出したフィールドの場合、最終的に変換されたバージョンの名前が使用されます。例えば、*bdate_years* です。

  構築済み予測フィールドの場合、*Predictor1* など、構築済み予測フィールドの名前が使用されます。
- 測定の尺度**:** データ型を示すアイコンが表示されます。

  目標フィールドの場合、「尺度」 は常に変換されたバージョンが反映されます (目標フィールドが変換されている場合)。例えば、順序型 (順序セット型) から連続型 (範囲型、スケール) への変更、またはその逆も同様です。

フィールドの詳細

「フィールド」メイン・ビューで 「名前」 をクリックすると表示され、「フィールド詳細」ビューには選択したフィールドの分布、欠損値、予測精度グラフ (該当する場合) が表示されます。さらに、該当する場合は、フィールドの処理履歴や変換対象フィールドの名前も表示されます。

各図表セットについて、2 つのバージョンが並んで表示され、変換が適用されたフィールドと適用されていないフィールドを比較します。変換されたバージョンのフィールドがない場合、元のバージョンの図表のみが表示されます。派生した日付/時刻フィールドおよび構築済み予測フィールドの場合、新しい予測フィールドの図表のみ表示されます。

注 : カテゴリー数が多すぎるためにフィールドが除外された場合、処理の履歴のみが表示されます。

分布図

連続型フィールドの分布は、正規曲線が重なり、平均値を表す垂直参照線を使用したヒストグラムで表示されます。カテゴリー・フィールドは棒グラフで表示されます。

ヒストグラムには、標準偏差や歪度を示すラベルがつけられています。ただし、値の数が 2 以下の場合、または元のフィールドの分散が 10 ～ 20 より小さい場合、歪度は表示されません。

図表の上にマウスポインタを停止させると、ヒストグラムの平均値、またはカテゴリーのレコード数合計の度数またはパーセンテージを棒グラフで表示します。

欠損値のグラフ

円グラフは、変換が適用された場合、変換が適用されていない場合の欠損値の割合を比較します。グラフのラベルはパーセンテージを示します。

ADP が欠損値の処理を実行した場合、変換後の円グラフには置換値、つまり欠損値の変わりに使用される値もラベルで表示します。

グラフにマウスポインタを停止させると、全体のレコード数の欠損値数と全体の割合が表示されます。

予測精度グラフ

推奨フィールドについて、棒グラフに変換前後の予測精度が表示されます。目標フィールドが変換されると、予測精度は変換後の目標フィールドについて計算されます。

注 : 目標が定義されていない場合、またはメイン・ビュー・パネルで目標をクリックした場合、予測精度のグラフは表示されません。

グラフの上のマウス・ポインタを停止させると、予測精度の値が表示されます。

処理履歴表

表には、変換されたバージョンのフィールドがどのように取得されたかを示されます。ADP によって行われた処理が、実行順に表示されます。ただし、特定のステップにおいては、特定のフィールドに対して複数の処理が実行されている場合があります。

注 : この表は、変換されていないフィールドには表示されません。

表内の情報は、2 つまたは 3 列に分けて表示されます。

- アクション**:** アクションの名前。例えば「連続型予測値」などです。詳しくは、トピック『アクションの詳細』を参照してください。
- 詳細**:** 実行された処理のリストです。例えば、標準単位への変換などです。
- 関数**:** 構築された予測値の場合にのみ表示され、入力フィールドの線型結合が表示されます (.06*age + 1.21*height など)。

アクションの詳細

「アクションの概要」メイン・ビューで下線の付いた 「アクション」 を選択した場合に表示されます。「アクションの詳細」リンク ビューには、実行された各アクションのアクション固有の情報およびおよび共通情報が表示されます。アクション固有の詳細情報が最初に表示されます。

各アクションについて、説明が、リンク ビューの一番上にタイトルとして表示されます。アクション固有の詳細がタイトルの下に表示され、派生予測フィールド数、フィールドの再計算、目標の変換、結合または並べ替えられたカテゴリー、構築または除外された予測フィールドの詳細が含まれる場合があります。

各アクションが処理されるごとに、予測フィールドが除外されたり結合されたりするなどの処理中に使用される予測フィールド数が変わる場合があります。

注 : アクションが無効になった場合、または指定された目標がなかった場合、「アクションの概要」メイン・ビューでアクションがクリックされた場合、アクションの詳細の代わりにエラー・メッセージが表示されます。

アクション数は 9 つですが、すべての分析で、すべての処理が行われるわけではありません。

テキスト・フィールド・テーブル

テーブルには、次の数が表示されます。
- 削除される空白値
- 分析から除外された予測フィールド

日付および時刻の予測フィールド・テーブル

テーブルには、次の数が表示されます。
- 日付および時刻予測フィールドから算出した期間
- 日付および時刻の要素
- 派生した日付および時刻の予測フィールドの合計

期間 (日数) が計算された場合、基準日または基準時刻が脚注として表示されます。

予測フィールドスクリーニング・テーブル

テーブルには、処理から除外された次の予測フィールドの数が表示されます。
- 定数
- 欠損値の多い予測フィールド
- 単一カテゴリーのケース数が多い予測フィールド
- カテゴリー数の多い名義型フィールド (セット)
- 除外された予測フィールドの合計

尺度テーブルのチェック

テーブルには再計算されたフィールド数を、次の項目に分けて表示します。

- 連続型として計算された順序型フィールド (順序セット型)
- 順序型フィールドとして計算された連続型フィールド
- 再計算の合計

連続型または順序型である入力フィールド (対象または予測フィールド) がない場合、脚注として表示されます。

外れ値テーブル

テーブルには、外れ値の処理方法の数が表示されます。

- 外れ値が検出されて削除された連続型フィールドの数、または外れ値が検出されて欠損値に設定された連続型フィールドの数のいずれか (どちらの数になるかは、「設定」タブの「入力と対象の準備」パネルの設定によって異なります)。
- 外れ値を処理した後定数項となったために除外される連続型フィールドの数。

1 つの脚注には外れ値の分割値、連続型である入力フィールド (目標または予測フィールド) がない場合、別の脚注が表示されます。

欠損値テーブル

テーブルには欠損値を置換したフィールド数を、次の項目に分けて表示します。

- 目標。目標が指定されていない場合はこの行は表示されません。
- 予測値。名義型 (セット)、順序型 (順序セット)、および連続型に分類して表示されます。
- 置換された欠損値の合計数。

ターゲット・テーブル

テーブルには、目標が変換されたかどうかについて、次のように表示されます。

- 正規性への Box-Cox 変換。指定の基準 (平均および標準偏差) およびラムダを示す列に分割されます。
- 安定性を向上させるために並べ替えられた目標カテゴリー。

カテゴリー型予測フィールド・テーブル

テーブルには、次のようなカテゴリー型予測フィールドの数が表示されます。

- カテゴリーが安定性が低いものから高いもの順に並べ替えられている。
- 目標との関連性を最大化するためにカテゴリーが結合されている。
- まばらなカテゴリーを処理するためにカテゴリーが結合されている。
- 目標との関連性の低さにより除外されている。
- 結合後定数項となったため除外されている。

カテゴリー予測フィールドがない場合、脚注が表示されます。

連続型予測フィールド・テーブル

連続型フィールド・テーブルには、2 つのテーブルがあります。一方には変換に関する次の数値のいずれかが表示されます。

- 標準の単位に変換された予測フィールド値。また、変換された予測フィールドの数、指定された平均値、標準偏差が表示されます。
- 共通範囲にマッピングされた予測フィールド値。また、指定された最小値や最大値のほか、min-max 変換を使用して変換された予測フィールド数も表示されます。
- 分割された予測フィールド値と分割された予測フィールド数。

もう一方のテーブルには、予測領域構築の詳細が、次のような予測フィールド数で表示されます。

- 構築済み。
- 目標との関連性の低さにより除外されている。
- 分割後定数項となったため除外されている。
- 構築後定数項となったため除外されている。

入力となっている連続型予測フィールドがない場合、脚注が表示されます。

フィールド生成ノードの生成

フィールド生成ノードを生成すると、目標の逆変換がスコア フィールドに適用されます。デフォルトで、ノードは自動モデラー (自動分類または自動数値) またはアンサンブル・ノードで作成されたスコア フィールドの名前を投入します。尺度型 (範囲型) 目標が変換されると、スコア フィールドは変換された単位で表示されます (例えば、$ ではなく log($))。結果を解釈して使用するために、予測値を元の尺度に変換する必要があります。

注 : フィールド作成ノードを生成できるのは、ADP ノードに範囲型目標を再調整する分析が含まれている場合 (「入力と対象の準備」パネルで Box-Cox 再調整を選択した場合) だけです。目標が範囲型でない場合、または Box-Cox 再調整が選択されていない場合、フィールド生成ノードは生成できません。

フィールド生成ノードは複数モードで作成され、式に @FIELD を使用して、必要に応じて変換された目標を追加できます。例えば、次の詳細情報を使用します。

- 対象フィールド名 :応答
- 変換された対象フィールド名 :response_transformed
- スコア フィールド名 :$XR-response_transformed

新しいフィールド $XR-response_transformed_inverse を作成する、フィールド作成ノードを生成します。

注 : 自動モデラーまたはアンサンブル・ノードを使用していない場合、フィールド生成ノードを編集して、モデルの適切なスコア フィールドを変換する必要があります。

連続型目標の正規化

デフォルトでは、「入力と対象の準備」パネルで「**Box-Cox** 変換で連続型対象をスケール変更」チェック・ボックスを選択すると、対象が変換されます。その際、モデルを構築するための新しいフィールドを作成します。例えば、元の目標が 「*response*」 の場合、新しい目標は 「*response_transformed*」 となります。ADP ノードの下流モデルは、この目標を自動的に選択します。

元の目標によっては、この操作で問題が発生する場合があります。例えば、目標が 「年齢」 だった場合、新しい目標の値は 「年」 ではなく、変換されたバージョンの 「年」 となります。認識可能な単位ではないため、スコアを確認できず解釈もできないということになります。こうした場合、変換された単位を以前の単位に戻す、逆変換を適用できます。これを行うには、次のようにします。

1. 「データの分析」 をクリックして、ADP 分析を実行した後、「生成」 メニューから 「フィールド生成ノード」 を選択します。
2. モデル領域のナゲットの後にフィールド生成ノードを投入します。

フィールド生成ノードがスコア フィールドを元の次元に復元し、予測値 が元の 「年」 の値となるようにします。

デフォルトでは、フィールド生成ノードは、アンサンブル化されたモデルの自動モデラーで生成されたスコア フィールドを変換します。個別モデルを作成している場合、フィールド生成ノードを編集して、実際のスコア フィールドから取得する必要があります。モデルを評価する場合、変換済み目標をフィールド生成ノードの 「派生元」 フィールドに追加する必要があります。これにより、同じ逆変換が目標に適用され、下流の評価ノードまたは分析ノードが、これらのノードをメタデータではなくフィールド名を使用する限り、変換済みデータを適切に使用します。

変換前の名前も復元する場合、存在する場合はフィルター・ノードを使用して、元の目標フィールドを削除し、目標フィールドおよびスコア フィールドの名前を変更できます。

データ型ノード

フィールドのプロパティーは、ソース・ノードまたは、個別のデータ型ノードで指定できます。どちらのノードでも機能は同じです。次のプロパティが使用できます。

- フィールド IBM SPSS Modeler でデータの値とフィールド ラベルを指定するには、そのフィールド名をダブルクリックします。例えば、IBM SPSS Statistics からインポートされるフィールド メタデータを表示したり、データ型ノードで変更したりできます。同様に、フィールドの新しいラベルとそれらの値を作成できます。データ型ノードで指定したラベルは、「ストリームのプロパティー」ダイアログ・ボックスの選択内容に応じて、IBM SPSS Modeler 全体にわたって表示されます。
- 尺度 測定の尺度で、特定フィールドのデータの特性を記述するために使用します。フィールドの詳細がすべてわかっている場合には、完全にインスタンス化済みとされます。詳しくは、 145 ページの『尺度』を参照してください。

  注: フィールドの尺度はストレージ タイプとは異なります。ストレージ タイプは、データが文字列、整数、実数、日付、時間、タイムスタンプ、リストのいずれで格納されているかを示します。
- 値 この列を使用して、データセットからデータ値を読み込むオプションを指定したり、「指定」オプションを使用して別のダイアログ ボックスで尺度および値を指定したりすることができます。値を読み込まないでフィールドを渡すこともできます。詳しくは、 150 ページの『データ値』を参照してください。

  注: 対応する「フィールド」エントリーがリストを含む場合、この列のセルは修正できません。
- 欠損値 フィールドの欠損値の処理方法を指定するために使用します。詳しくは、 155 ページの『欠損値の定義』を参照してください。

  注: 対応する「フィールド」エントリーがリストを含む場合、この列のセルは修正できません。
- 検査 この列には、フィールドの値が指定された値または範囲内に収まっているかどうかを検査するオプションを設定できます。詳しくは、 155 ページの『データ型の値の検査』を参照してください。

  注: 対応する「フィールド」エントリーがリストを含む場合、この列のセルは修正できません。
- 役割 フィールドがマシン学習プロセスの「入力」(予測フィールド) または「対象」(予測されるフィールド) のどちらになるかをモデル作成ノードに指示するために使用します。「両方」 および 「なし」 も役割として利用できます。さらに、レコードを学習用、検定用、および検証用の独立したサンプルに

分割するために使用されるフィールドを示す 「データ区分」 も利用できます。値 「分割」 は各モデルがフィールドの可能な値それぞれに作成されるように設定します。 詳しくは、 156 ページの『フィールドの役割の設定』を参照してください。

他にもデータ型ノード ウィンドウを使用して指定できるさまざまなオプションがあります。

- データ型ノードがインスタンス化されたら、ツール・メニュー・ボタンを使用して、「単一フィールドを無視」 を選択することができます。「単一フィールドを無視」を選択すると、1 つの値だけを持つフィールドが自動的に無視されます。
- データ型ノードがインスタンス化されたら、ツール・メニュー・ボタンを使用して、「ラージ セットを無視」 を選択することができます。「ラージ セットを無視」を選択すると、メンバー数の多いセットが自動的に無視されます。
- データ型ノードがインスタンス化されたら、ツール・メニュー・ボタンを使用して、「連続型整数を順序型に変換」 を選択することができます。 詳しくは、トピック 149 ページの『連続型データの変換』を参照してください。
- ツール・メニュー・ボタンを使用して、フィルター・ノードを生成して選択したフィールドを除外することができます。
- サングラスの形をしたトグル ボタンを使用して、すべてのフィールドのデフォルトを「読み込み順」または「通過」に設定することができます。ソース・ノードの「データ型」タブのデフォルトでは、フィールドを通過しますが、データ型ノード自体はデフォルトで値を読み込みます。
- 「値の消去」 ボタンを使用して、このノード中でフィールド値に対して行った変更内容を消去し (継承されない値)、上流の操作から値を読み込み直すことができます。このオプションは、上流からの特定のフィールドに対して行った変更をリセットする場合に役立ちます。
- 「すべての値の消去」 ボタンを使用して、ノードに読み込まれたすべてのフィールドの値をリセットすることができます。このオプションにより、すべてのフィールドに対して 「値」 列に **「読み込み」** が設定されます。このオプションは、すべてのフィールドに対して値をリセットし、上流の操作から値とデータ型を読み込み直す場合に役立ちます。
- コンテキスト・メニューを使用して、1 つのフィールドから別のフィールドへの属性の 「コピー」 を選択できます。 詳しくは、トピック 157 ページの『データ型属性のコピー』を参照してください。
- 「未使用のフィールド設定を表示」 オプションを使用して、データ中にすでに存在していない、または過去にこのデータ型ノードに接続されていたフィールドのデータ型の設定を表示することができます。これは、変更されたデータ・セットのデータ型ノードを再利用する場合に役立ちます。

尺度

尺度 (以前の「データ型」または「使用タイプ」)、IBM SPSS Modeler におけるデータ・フィールドの使用法を記述します。入力ノードまたはデータ型の「データ型」タブで尺度を設定することができます。例えば、値 1 と 0 をとる整数フィールドの測定の尺度をフラグ型に設定することができます。通常は、1 = *True*、0 = *False* を示します。

ストレージと尺度の比較 **:** フィールドの尺度は、ストレージ・タイプとは異なります。これは、データが文字列、整数、実数、日付、時間、またはタイムスタンプのどれで保存されるかを示すことに注意してください。データ型ノードを使用してストリームの任意のポイントでデータ型を変更できますが、一方、ストレージは、IBM SPSS Modeler へのデータ読み込時において入力で決定する必要があります。 詳しくは、トピック 9 ページの『フィールドのストレージと形式の設定』を参照してください。

いくつかのモデリング・ノードは、それらの「フィールド」タブ上のアイコンによって入力フィールドおよび対象フィールドに対して許可される測定の尺度の種類を示します。

測定レベル・アイコン

表 20. 測定レベル・アイコン

アイコン	尺度
	デフォルト
	連続
	カテゴリー型
	フラグ
	名義
	順序
	不明
	集計棒グラフ
	地理空間

使用できる測定の尺度は、次の通りです。

- デフォルト ストレージ タイプと値が不明なデータは (まだ読み込まれていない場合など)、「<デフォルト>」と表示されます。
- 連続型 0-100 や 0.75-1.25 という範囲のように、数値を記述するために使用されます。連続値は、整数、実数、または日付/時間になります。
- カテゴリー型 個別値の正確な数値が不明な場合に、文字列値に使用されます。これはインスタンス化されていないデータ型で、すべてのストレージに関する情報やデータの使用法に関する情報がわかっていないことを示します。データが読み取られると、「ストリーム・プロパティー」ダイアログ・ボックスで指定された名義型フィールドの最大メンバー数に応じて、測定の尺度はフラグ型、名義型、またはデータ型不明 になります。
- フラグ型 `true` と `false`、`Yes` と `No`、`0` と `1` など、特性の有無を示す 2 つの異なる値を持つデータで使用されます。使用される値は異なる場合がありますが、一方の値は常に「true」値として、もう一方の値は「false」値として割り当てられる必要があります。 データは、テキスト、整数、実数、日付/時間、またはタイム・スタンプを表します。

- 名義型 複数の異なる値を持つデータを記述するために使用されます。それぞれの値は、「small/medium/large」などのセットのメンバーとして処理されます。名義データは、任意のストレージ (数値、文字列、日時) を持つことができます。測定の尺度を名義型にしても、自動的に値が文字列に変わることはないことに注意してください。
- 順序型 固有の順序のある複数の異なる値を持つデータを記述するために使用されます。例えば、給与区分や満足度ランキングは、順序型データとして類別できます。順序は、データ要素の普通のソート順により定義されます。例えば、1、3、5 は一連の整数のデフォルトのソート順ですが、HIGH、LOW、NORMAL (アルファベットの昇順) は一連の文字列の順序です。可視化、モデル設定および IBM SPSS Statistics など順位データを DISTINCT 型として認識する他のアプリケーションへの出力するため、順序型の測定の尺度によりカテゴリー・データ・セットを順位データとして定義できます。順序型フィールドは、名義型フィールドを使用できる場合はいつでも使用できます。また、任意のストレージ型 (実数、整数、文字列、日付、時間など) のフィールドは順序型として定義できます。
- データ型不明 上記のいずれのデータ型にもあてはまらない、値が 1 つのフィールド、または定義された最大数を超えるメンバーがある名義型のデータに使用されます。この測定の尺度は、不明にしないと、データ型が多くのメンバー (アカウント番号など) を使用して設定されてしまうような場合に効果的です。フィールドに 「データ型不明」 を選択すると、役割が自動的に 「なし」 に設定され、「レコード **ID**」 が唯一の代替となります。デフォルトのセットの最大サイズは、250 の一意な値です。この数値は、「ツール」メニューからアクセスできる「ストリームのプロパティー」ダイアログ・ボックスの「オプション」タブで調整または無効化することができます。
- 集合 **(Collection)** リストで記録される、地理空間データ以外のデータを識別するために使用されます。集合は実質的に深さが 0 のリスト フィールドであり、リストの要素が他のいずれかの尺度を持つものです。

  リストについて詳しくは、「SPSS Modeler Source, Process, and Output Nodes Guide」の『Source Nodes』セクションにある『List Storage and Associated Measurement Levels』というトピックを参照してください。
- 地理空間 地理空間データを識別するためにリスト ストレージ タイプとともに使用されます。リストは、 0 から 2 までのリストの深さを持つ「整数のリスト」フィールドまたは「実数のリスト」フィールドにすることができます。

  詳しくは、「SPSS Modeler Source, Process, and Output Nodes Guide」の『Type Node』セクションにある『Geospatial Measurement Sublevels』というトピックを参照してください。

手作業で尺度を指定するか、またはソフトウェアにデータを読み込ませ、その値に基づいて尺度を判断させることができます。

また、カテゴリー・データとして処理する必要のあるいくつかの連続型データ・フィールドがある場合、それらを変換するオプションを選択することができます。 詳しくは、トピック 149 ページの『連続型データの変換』を参照してください。

自動入力を使用するには

1. データ型ノードまたはソース・ノードの「データ型」タブのいずれかで、必要なフィールドの「値」列を「<読み込み>」に設定します。これで、下流にあるすべてのノードでメタデータが利用できるようになります。ダイアログ・ボックスのサングラス・ボタンを使用すると、すべてのフィールドを簡単に「<読み込み>」または「<通過>」に設定することができます。
2. 「値の読み込み」 ボタンをクリックして、データ・ソースから直接値を読み込むこともできます。

フィールドの尺度を手作業で設定するには

1. テーブル中のフィールドを選択します。
2. 「尺度」 列のドロップダウン・リストで、フィールドの尺度を選択します。
3. Ctrl-A または Ctrl キーを押しながらクリックして、複数のフィールドを選択してから、ドロップダウン・リストで尺度を選択することもできます。

地理空間のサブ尺度

地理空間の尺度はリスト ストレージ タイプとともに使用します。この尺度には、各種の地理空間データを識別するために使用する 6 つのサブ尺度が用意されています。

- ポイント - 都心部など、特定の場所を特定します。
- 多角形 - 領域の単一の境界とその場所 (国など) を識別する一連のポイント。
- 行ストリング - 折れ線や単に線とも呼びます。行ストリングは、線の経路を識別する一連のポイントです。行ストリングには、例えば固定の項目 (道路、河川、鉄道など) や、移動する物体の軌跡 (航空機の飛行経路や船舶の航路など) が該当します。
- 複数点 - 1 つの領域に対応する複数のポイントがデータの各行に含まれる場合に使用します。例えば、各行で都市の道路を表し、道路ごとに複数のポイントを使用して個別の街灯を示します。
- 多角形群 - データの各行が複数の多角形を含む場合に使用します。例えば、各行が国の輪郭を表す場合に、米国を複数の多角形として記述し、本土、アラスカ、ハワイなどの各地域を示すことができます。
- 複数行ストリング - データの各行が複数の線を含む場合に使用します。線は分岐できないため、複数行ストリングを使用することで一群の行を表すことができます。例えば、それぞれの国の可航水路や線路網などのデータが該当します。

これらのサブ尺度は、リスト ストレージ タイプで使用します。詳しくは、 12 ページの『リストス トレージおよび関連する尺度』を参照してください。

制限

地理空間データを使用する場合は、いくつか制限があることに注意してください。

- 座標系がデータの形式に影響する場合があります。例えば、投影座標系で x と y (および必要な場合は z) の座標値を使用しているが、地理座標系では経度と緯度 (および必要な場合は高度または深度) の座標値を使用している場合が該当します。

  座標系について詳しくは、「SPSS Modeler User's Guide」の『Working with Streams』セクションに記載されている 『Setting Geospatial Options for Streams』というトピックを参照してください。
- 行ストリングをそれ自体と交差させることはできません。
- 多角形は自己閉鎖でありません。それぞれの多角形で、最初のポイントと最後のポイントを同じポイントとして定義する必要があります。
- 多角形群ではデータの方向が重要です。時計回りの場合は内側が満たされていることになり、反時計回りの場合は中空になります。例えば、湖が存在する国土を記録する場合には、主要な陸地の境界を時計回りで記録し、それぞれの湖の形状を反時計回りで記録することが可能です。
- 多角形をそれ自体と交差させることはできません。このような交差には、例えば、図 8 のような形で連続した直線で多角形の境界をプロットしようとした場合が該当します。
- 多角形群をオーバーラップさせることはできません。
- 地理空間フィールドの場合、該当するストレージ タイプは「実数」と「整数」のみです (デフォルト設定は「実数」です)。

地理空間サブ尺度のアイコン

表 *21.* 地理空間サブ尺度のアイコン

アイコン	尺度
	ポイント
	多角形
	行ストリング
	複数点
	多角形群
	複数行ストリング

連続型データの変換

カテゴリー・データを連続型として処理すると、データの品質に重大な影響があります。例えば、それが対象フィールドである場合、2 値モデルでなく回帰モデルを作成します。この影響を回避するために、整数の範囲を、順序型またはフラグ型などのカテゴリー型に変換できます。

1. 道具の記号の付いた「操作および生成」メニュー・ボタンから、「連続型整数を順序型に変換」 を選択します。「変換値」ダイアログが表示されます。
2. 自動的に変換される範囲のサイズを指定します。これは、入力したサイズまでの範囲に適用されます。
3. 「**OK**」をクリックします。 影響を受ける範囲型は、フラグ型または順序型に変換され、データ型ノードの「データ型」タブに表示されます。

変換の結果

- 整数のストレージを持つ連続型フィールドを順序型に変更すると、下限値および上限値を拡張して、下限値から上限値までの整数値のすべてを含みます。例えば、範囲が 1 〜 5 の場合、値のセットは 1、2、3、4、5 です。
- 連続型フィールドをフラグ型フィールドに変更すると、下限値および上限値はフラグ型フィールドの偽の値および真の値となります。

インスタンス化とは ?

インスタンス化は、データ・フィールドのストレージ・タイプや値などの情報を読み込む、または指定するプロセスです。システム リソースを最適化するために、インスタンス化を行う作業はユーザーが指示する必要があります。ソース・ノードの「データ型」タブでオプションを指定するか、またはデータ型ノードにデータを流すことによって、ソフトウェアに値の読み込みを指示します。

- 不明なデータ型のデータは、インスタンス化されていないデータとも呼ばれます。ストレージ・タイプと値が不明なデータは、「データ型」タブの「尺度」列に「<デフォルト>」として表示されます。
- 文字列や数値など、フィールドのストレージに関する一部の情報が分かっている場合、そのデータは部分的にインスタンス化されていることになります。カテゴリー型や連続型は、部分的にインスタンス化された測定の尺度です。例えば、カテゴリーはフィールドがシンボル型であること (ただしそれが名義型、順序型またはフラグ型のどちらであるかわからないこと) を示します。
- データ型の詳細が、値も含めてすべて分かっている場合、この列には完全にインスタンス化された測定の尺度 (名義型、順序型、フラグ型、または連続型) が表示されます。注 : 連続型 は、部分的にインスタンス化されたデータ・フィールドと、完全にインスタンス化されたデータ・フィールドの両方で使用されます。連続型データは、整数または実数になります。

データ型ノードでデータ・ストリームを実行中、インスタンス化されていないデータ型は、初期のデータ値を基にして、部分的にインスタンス化されます。すべてのデータがノードを通過すると、値が「<通過>」に設定されている場合を除き、それらのデータが完全にインスタンス化されます。実行が中断された場合は、データは部分的にインスタンス化されたままになります。「データ型」タブがインスタンス化されたら、ストリーム中のその時点でフィールドの値は固定化されます。つまり、ストリームを再実行した場合も、上流の変更は特定のフィールドの値に影響しないということです。新しいデータや追加の操作に基づいて値を変更または更新するには、「データ型」タブで編集するか、フィールドの値を「<読み込み>」または「<読み込み +>」に設定する必要があります。

インスタンス化する場合

一般的に、データ・セットがさほど大きくなく、後でストリームにフィールドを追加する予定がない場合は、ソース・ノードでインスタンス化するのが便利です。ただし、次の場合には、別のデータ型ノードでインスタンス化するほうが便利です。

- データ・セットが巨大で、ストリームがデータ型ノードの前でサブセットをフィルタリングしている場合。
- ストリーム中でデータをフィルタリングしている場合。
- ストリーム中でデータが結合または追加されている場合。
- 処理の過程で新しいデータ・フィールドが作成される場合。

データ値

「データ型」タブの「値」 列を使用して、データから値を自動的に読み込んだり、個別のダイアログ・ボックスで尺度と値を指定することができます。

「値」ドロップダウン・リストに表示されるオプションを使用して、以下の表に示す自動入力用の指定を行うことができます。

表 22. 自動入力用の指定

オプション	関数
<Read>	ノードの実行時にデータが読み取られます。
<読み込み+>	データが読み取られ、現在のデータ (存在している場合) に追加されます。
<Pass>	データは読み取られません。
<Current>	現在のデータ値を保持します。
指定...	値および測定の尺度のオプションを指定するための個別のダイアログ ボックスが開きます。

データ型ノードを実行するか、または「値の読み込み」をクリックすると、選択内容に応じてデータ ソースからの値の自動入力および読み込みが行われます。これらの値は、「指定」オプションを使うか、または「フィールド」 列のセルをダブルクリックして、手作業で入力することもできます。

データ型ノードでフィールドを変更した後、ダイアログ ボックスのツールバーにある次のボタンを使用して値情報をリセットすることができます。

- 「値の消去」 ボタンを使用して、このノード中でフィールド値に対して行った変更内容を消去し (継承されない値)、上流の操作から値を読み込み直すことができます。このオプションは、上流からの特定のフィールドに対して行った変更をリセットする場合に役立ちます。
- 「すべての値の消去」 ボタンを使用して、ノードに読み込まれたすべてのフィールドの値をリセットすることができます。このオプションにより、すべてのフィールドに対して 「値」 列に 「読み込み」 が設定されます。このオプションは、すべてのフィールドに対して値をリセットし、上流の操作から値と尺度を読み込み直す場合に役立ちます。

「値」列のグレーのテキスト

データ型ノード内またはソース ノード内のいずれかで、「値」列のデータが黒色のテキストで表示されている場合は、そのフィールドの値が読み込まれ、そのノード内に格納されていることを示します。このフィールドに黒色のテキストが表示されない場合、そのフィールドの値は読み込まれず、さらに上流で決定されます。

グレーのテキストでデータが表示される場合があります。これは、SPSS Modeler が実際にデータの読み込みや格納を行うことなく、フィールドの有効値を識別または推測できる場合に発生します。これは、以下のノードのうちのいずれかを使用する場合に発生する可能性が高くなります。

- ユーザー入力ノード。データはノード内で定義されるため、あるフィールドの値がノード内に格納されていなかったとしても、そのフィールド値の範囲は、常に識別されます。
- Statistics ファイル ソース ノード。データ型を表すメタデータが存在する場合、このノードにより、データの読み込みや格納を行うことなく SPSS Modeler で値の可能な範囲を推論できるようになります。

いずれのノードでも、ユーザーが「値の読み込み」をクリックするまで、値はグレーのテキストで表示されます。

注: ストリーム内でデータをインスタンス化せず、データ値がグレーで表示される場合は、「検査」列で設定したデータ型の値の検査はいずれも適用されません。

「値」ダイアログ・ボックスの使用

「データ型」タブの 「値」 または 「欠損値」 列をクリックすると、事前定義された値のドロップダウン・リストが表示されます。このリストの「指定...」オプションを選択すると個別のダイアログ ボックスが表示され、選択したフィールドの値の読み込み、指定、ラベル付け、処理のオプションを設定できます。

コントロールの大半は、すべての種類のデータに共通しています。ここでは、これらの共通のコントロールを説明していきます。

尺度現在選択されている測定の尺度を表示します。データ利用目的に応じて設定を変更することができます。例えば、`day_of_week` というフィールドに個別の曜日を表す数字が格納されている場合に、各カテゴリーを個別に調査する棒グラフ・ノードを作成するために、名義型データに変更することができます。

ストレージ ストレージ タイプが判明している場合に、そのタイプを表示します。ストレージ・タイプは、選択した尺度の影響は受けません。ストレージ・タイプを変更するには、可変長ノードまたは固定長ノードの「データ」タブを使用するか、または置換ノードの変換関数を使用します。

モデル・フィールド モデル・ナゲットのスコアリングの結果として生成されたフィールドの場合、モデル・フィールドの詳細も表示することができます。詳細には、モデル作成時のフィールドの役割 (予測値、確率、傾向など) や対象フィールド名も含まれています。

値 選択したフィールドの値を決める方法を選択します。ここで選択した内容は、「データ型ノード」ダイアログ・ボックスの 「値」 列で行った選択内容に優先します。値の読み込みに関する選択項目には、次のようなものがあります。

- データから読み込み ノードの実行時に値を読み込む場合に選択します。このオプションは、「<読み込み>」と同じです。
- 通過 現在のフィールドのデータを読み込まない場合に選択します。このオプションは「<通過>」と同じです。
- 値とラベルの指定 このオプションは選択したフィールドの値とラベルを指定するために使用します。このオプションは、値の検査とともに使用して、現在のフィールドに対する知識に基づく値を指定します。このオプションを選択すると、フィールドの種類に応じた独自のコントロールが有効になります。値とラベルのオプションは、以降の各項目で個別に説明しています。

  注: 測定の尺度が「データ型不明」または <デフォルト> であるフィールドの場合、値やラベルを指定することはできません。
- データから値を拡張 現在のデータに、ここで入力した値を追加する場合に選択します。例えば、field_1 の範囲が (0,10) の場合に、値の範囲として (8,16) を入力した場合、元の最小値を削除せずに 16 を追加して範囲が拡張されます。新しい範囲は (0,16) になります。このオプションを選択すると、自動入力オプションが自動的に「<読み込み+>」になります。

リストの最大長 地理空間または集合のいずれかの尺度を持つデータのみに使用できます。リストの最大長を設定するには、リストに入れることができる要素の数を指定します。

最大文字列長 データ型不明のデータのみに使用できます。このフィールドは、SQL を生成してテーブルを作成するときに使用します。データの最大文字列の値を入力します。これにより、テーブルに生成される列が、その文字列に対して十分な大きさになります。文字列長の値が使用不可の場合は、デフォルトの文字列サイズが使用されます。このサイズは、データに対して適切でない場合があります (例えば、値が小さすぎる場合は、テーブルにデータを書き込むときにエラーが発生する場合があります。また、値が大きすぎる場合は、パフォーマンスに悪影響が及ぶ場合があります)。

値の検査 値が指定した連続型、フラグ型、または名義型の規則にしたがっているかどうかの検査方法を選択します。このオプションは、「データ型ノード」ダイアログ・ボックスの 「検査」 列に対応しており、ここでの設定内容がダイアログ・ボックスでの設定に優先されます。「値とラベルを指定」オプションとともに使用すると、値の検査によりデータ内の値を必要な値に一致させることができます。例えば、値に 1,0 と指定した後、検査オプションを使用して、1 と 0 以外のすべての値を持つレコードを破棄することができます。

空白を定義 選択すると、データの欠損値や空白を宣言するために使用する以下のコントロールがアクティブになります。

- 欠損値 このテーブルを使用して、99 や 0 などの特定の値を空白として定義します。この値は、フィールドのストレージ・タイプに適当なものでなければなりません。
- 範囲 欠損値の範囲を指定する場合に使用します (1 歳から 17 歳、65 歳超など)。境界値が空欄のままの場合、その範囲は無制限になります。例えば、下限に 100 が指定されていて上限がない場合、100 以上のすべての値は欠'1値になります。各境界値は範囲にふくまれます。例えば、下限が 5 で上限が 10

の範囲は、範囲定義に 5 と 10 も含まれます。欠損値範囲は、任意のストレージ・タイプに対して定義できます。これには、日付/時刻および文字列もふくまれます (値が範囲内かどうかを決定するために、アルファベット順のソート順序が使用されます)。

- ヌル/空白文字 システムのヌル値 (データ内では $null$ と表示されます) と空白文字 (表示可能な文字を含まない文字列値) を空白として指定することもできます。

  注: 内部で別に格納され、特定のケースで別に処理されるにもかかわらず、データ型ノードも空白文字列のような空の文字列を分析のために扱います。

注: 未定義または $null$ として空白をコード化するには、置換ノードを使用します。

説明 このテキスト ボックスを使用してフィールド ラベルを指定します。ラベルは、「ストリームのプロパティー」ダイアログ ボックスでの選択内容に応じて、グラフ、テーブル、出力、モデル ブラウザーなどのさまざまな場所に表示されます。

連続型データの値およびラベルの指定

連続型尺度は数値型フィールドに使用されます。連続型ノードには、3 種類のストレージ・タイプがあります。

- 実数
- 整数
- 日付/時刻

これらの 3 種類の連続型フィールドを編集するには、同じダイアログ・ボックスが使用されます。ただし、ストレージ・タイプは参照用にのみ表示されます。

値の指定

次のコントロールは連続型フィールドに固有のもので、値の範囲を指定するのに使用します :

下限: 値の範囲の下限を指定します。

上限: 値の範囲の上限を指定します。

ラベルの指定

集計範囲フィールドの任意の値のラベルを指定できます。ラベル ボタンをクリックして、値ラベルを指定する個別のダイアログボックスを開きます。

値とラベル・サブダイアログ・ボックス: 集計範囲フィールドの「値を指定」ダイアログ・ボックスの「ラベル」 をクリックして、範囲中の任意の値のラベルを指定できる新しいダイアログ・ボックスを開きます。

このテーブルの 「値」 と 「ラベル」 列で値とラベルのペアを定義できます。現在定義されているペアがここに表示されます。空のセルをクリックして値と対応するラベルを入力すると、新しいラベルのペアを追加できます。注 : このテーブルに値/値とラベルのペアを追加しても、フィールドに新しい値は追加されません。その代わり、フィールド値のメタデータが作成されるだけです。

データ型ノードで指定したラベルは、「ストリームのプロパティー」ダイアログ・ボックスの選択内容に応じて、ツールヒントや出力ラベルなどとしてさまざまな場所で表示されます。

数値型データおよび順序型の名前とラベルの指定

名義型 (セットが他) および順序型 (順序セット型) の尺度は、データの値がセットのメンバーとして個別に使われることを表しています。セット型では、文字列、整数、実数、または日付/時刻のストレージ・タイプを利用することができます。

次のコントロールは名義型フィールドおよび順序型フィールドに固有のもので、値とラベルを指定するのに使用します。

値**:** テーブルの「値」列により、現在のフィールドに関する知識に基づいて値を指定することができます。このテーブルを使用してフィールドに期待値を入力し、「値の検査」ドロップダウン・リストを使用してこれらの値に対するデータ・セットの整合性を検査することができます。矢印および削除ボタンを使用して、既存の値を変更したり、値の並び替えや削除などの作業を行えます。

ラベル**:** 「ラベル」列により、セット内のそれぞれの値にラベルを指定することができます。ラベルは、「ストリームのプロパティー」ダイアログ・ボックスの選択内容に応じて、グラフ、テーブル、出力およびモデル・ブラウザーなどの様々な場所で表示されます。

フラグ型の値の指定

フラグ型フィールドは、2 つの DISTINCT 値を持つデータを表示するために使用されます。フラグ型では、文字列、整数、実数、または日付/時刻のストレージ・タイプを利用することができます。

真 **(true):** 条件を満たす場合のフィールドのフラグ値を指定します。

偽 **(false):** 条件を満たさない場合のフィールドのフラグ値を指定します。

ラベル**:** フラグ型フィールド内のそれぞれの値のラベルを指定します。ラベルは、「ストリームのプロパティー」ダイアログ・ボックスの選択内容に応じて、グラフ、テーブル、出力およびモデル・ブラウザーなどの様々な場所で表示されます。

集合データの値の指定

集合フィールドは、リストに含まれる非地理空間データを表示するために使用します。

集合の「尺度」に設定できる項目は「尺度のリスト」のみです。デフォルトではこの尺度が「不明」に設定されていますが、別の値を選択して、リスト内の要素の尺度を設定することができます。以下のいずれかのオプションを選択できます。

- 不明
- 連続
- 名義
- 順序
- フラグ

地理空間データの値の指定

地理空間フィールドは、リストに含まれる地理空間データを表示するために使用します。

地理空間の「尺度」の場合は、以下のオプションを設定して、リスト内の要素の尺度を設定することができます。

データ型 地理空間フィールドのサブ尺度を選択します。使用可能なサブ尺度は、リスト フィールドの深さによって決まります。デフォルトは、ポイント (深さ 0)、行ストリング (深さ 1)、多角形 (深さ 1) です。

サブ尺度について詳しくは、 148 ページの『地理空間のサブ尺度』を参照してください。

リストの深さについて詳しくは、 12 ページの『リストス トレージおよび関連する尺度』を参照してください。

座標系 このオプションは、尺度を地理空間以外から地理空間に変更した場合にのみ使用可能です。座標系を地理空間データに適用するには、このチェック ボックスを選択します。デフォルトでは、「ツール」 > 「ストリームのプロパティー」 > 「オプション」 > 「地理空間」ペインで設定された座標系が表示されます。別の座標系を使用するには、「変更」ボタンをクリックして「座標系の選択」ダイアログボックスを表示し、必要な座標系を選択します。

座標系について詳しくは、「SPSS Modeler User's Guide」の『Working with Streams』セクションに記載されている 『Setting Geospatial Options for Streams』というトピックを参照してください。

欠損値の定義

「データ型」タブの 「欠損値」 列は、欠損値の処理がフィールドに定義されているかどうかを示します。次のように設定できます。

オン **(*):** 欠損値の処理がこのフィールドに定義されていることを示します。下流の置換ノードを使用して、または「指定」オプションを使用した明示的な指定によって (下記参照) 行うことができます。

オフ**:** フィールドに欠損値の処理が定義されていません。

指定**:** このフィールドで欠損値として見なされる明示的な値を宣言できるダイアログを表示します。

データ型の値の検査

各フィールドの「検査」オプションを有効にすると、そのフィールド中のすべての値を調べて、値が現在のデータ型または「値を指定」ダイアログ・ボックスで指定した値に適合しているかどうかが検査されます。これは、1 回の操作で、データ・セットをクリーン・アップして、データ・セットのサイズを減らすために役立ちます。

「データ型ノード」ダイアログ・ボックス内の 「検査」 列の設定によって、データ型の制限を超えた値が検出されたときの処理方法が決まります。フィールドの検査の設定を変更するには、そのフィールドの 「検査」 列にあるドロップダウン・リストを使用します。すべてのフィールドに対して検査を行うには、「フィールド」 列をクリックして Ctrl-A を押します。次に 「検査」 列中の任意のフィールドのドロップダウン・リストを使用してください。

「検査」では、次の設定を行うことができます。

なし**:** 値は検査されずに通過します。これはデフォルト設定です。

無効化**:** 制限外の値をシステムのヌル値 (`$null$`) に変更します。

強制**:** 完全にインスタンス化された尺度のフィールドに対して、その値が指定した範囲を超えていないかどうかが検査されます。指定範囲外の値は、次のルールに基づいて各尺度における有効な値に変換されます。

- フラグ型の場合、真 (true) または偽 (false) 以外の値は、偽 (false) の値に変換されます。
- セット型 (名義型または順序型) の場合は、すべての未知の値は、セットの値の最初のメンバーに変換されます。
- 範囲の上限より大きい数値は、上限値に置き換えられます。
- 範囲の下限より小さい数値は、下限値に置き換えられます。

- 範囲が設定されているにもかかわらず、ヌル値が検出された場合は、その範囲の中間の値が指定されます。

破棄: 不正な値が検出された場合は、レコード全体が破棄されます。

警告: すべてのデータを読み込む際に、不正な項目数がカウントされ、件数が「ストリームのプロパティー」ダイアログ・ボックスにレポートされます。

中止: 最初に不正な値が検出された時点で、ストリームの実行が終了します。エラーは、「ストリームのプロパティー」ダイアログ・ボックスに報告されます。

フィールドの役割の設定

フィールドの役割は、フィールドがモデル構築でどのように使用されるかを指定します。例えば、フィールドが入力か対象 (予測対象) か、などです。

注: データ区分、度数、レコード ID の役割は、それぞれ 1 つのフィールドだけに適用できます。

次の役割があります。

入力。このフィールドは、マシン学習に対する入力 (予測変数フィールド) として使用されます。

目標: このフィールドは、マシン学習の出力または対象 (モデルが予測しようとするフィールドの 1 つ) として使用されます。

両方。このフィールドは、Apriori ノードで入力と出力の両方として使用されます。他のすべてのモデル作成ノードでは、このフィールドは無視されます。

なし: このフィールドはマシン学習では無視されます。尺度が「データ型不明」に設定されているフィールドは、自動的に「役割」 列が「なし」 に設定されます。

データ区分: データを、学習、テスト、および (オプションで) 検証の各目的用の異なるサンプルに分割するためのフィールドを示します。このフィールドは、(「フィールド値」ダイアログ・ボックスで定義されているように) 2 個または 3 個の値を取るセット型としてインスタンス化されなければなりません。最初の値は、学習用サンプル、2 番目の値はテスト用サンプル、3 番目に値がある場合は、検証用サンプルを表しています。その他の値は無視され、フラグ・フィールドを使用できません。分析でデータ区分を使用するには、適当なモデル構築または分析ノードの「モデルのオプション」タブでデータ分割を有効にする必要があることに注意してください。データ分割が有効になっている場合、データ区分フィールドではヌル値をもつレコードは無視されます。複数のデータ区分フィールドがストリーム内で定義されていた場合、該当する各モデリング・ノードの「フィールド」タブで単一のデータ区分フィールドを指定する必要があります。使用するデータ中に適当なフィールドがまだ存在していない場合、データ区分ノードまたはフィールド作成ノードを使用すると新規に作成できます。 詳しくは、トピック 185 ページの『データ区分ノード』を参照してください。

分割。(名義型、順序型、フラグ型フィールドのみ) フィールドの可能な値それぞれにモデルが作成されるよう指定します。

度数: (数値型フィールドのみ) この役割を設定すると、フィールド値をレコードの度数の重みの因子として使用できます。この機能は C&R ツリー、CHAID、QUEST および線型モデルにのみサポートされます。他のすべてのノードはこの役割を無視します。度数の重みは、この機能をサポートするモデル作成ノードの「フィールド」タブで 「度数の重みを使用」 オプションを設定すると使用できます。

「レコード **ID**」。一意のレコード ID として使用されます。この機能は多くのノードによって無視されますが、線型モデルによってサポートされており、IBM Netezza データベース内マイニング・ノードに必要です。

データ型属性のコピー

値、検査オプション、および欠損値などのデータ型の属性を、あるフィールドから別のフィールドに簡単にコピーすることができます。

1. 属性をコピーするフィールドを右クリックします。
2. コンテキスト・メニューから、「コピー」 を選択します。
3. 属性を変更するフィールドを右クリックします。
4. コンテキスト・メニューから、「形式を選択して貼り付け」 を選択します。注: Ctrl キーを押しながらクリックするか、またはコンテキスト・メニューの「フィールドの選択」オプションを使用することにより、複数のフィールドを選択することもできます。

新しいダイアログ・ボックスが表示されます。ここから、貼り付ける特定の属性を選択することができます。複数のフィールドに貼り付ける場合は、ここで選択したオプションがすべての対象フィールドに適用されます。

次の属性を貼り付け。あるフィールドから別のフィールドに貼り付ける属性を、下のリストから選択します。

- データ型。尺度貼り付ける場合に選択します。
- 値。フィールドの値を貼り付ける場合に選択します。
- 欠損値。欠損値の設定を貼り付ける場合に選択します。
- 検査。値の検査オプションを貼り付ける場合に選択します。
- 役割。フィールドの役割を貼り付ける場合に選択します。

フィールド形式の「設定」タブ

「テーブル」中の「形式」タブとデータ型ノードは、現在のまたは未使用フィールドのリストおよび各フィールドの書式設定オプションを示します。フィールド形式テーブルの各列の説明を次に示します。

フィールド。選択したフィールド名が表示されます。

形式。 この列のセルをダブルクリックすると、ダイアログ・ボックスが表示され、個別にフィールドの書式設定を指定できます。詳しくは、トピック 158 ページの『フィールド形式のオプションの設定』を参照してください。ここで指定した書式設定は、ストリームのプロパティー全体で指定した書式設定に上書きされます。

注: Statistics のエクスポート・ノードと Statistics の出力ノードは、フィールドごとの書式設定がメタデータに含まれている *.sav* ファイルをエクスポートします。IBM SPSS Statistics *.sav* ファイル形式でサポートされない設定がフィールドごとの書式に指定された場合、ノードは IBM SPSS Statistics のデフォルト形式を使用します。

表示位置。この列を使用して、テーブルの列内で値をどのように表示するかを指定します。デフォルトの設定は 「自動」 で、シンボル値を左寄せ、数値を右寄せで表示します。デフォルト値を変更するには、「左」、「右」、または 「中央」 を選択します。

列幅。デフォルトでは、フィールドの値に基づいて列幅が自動的に計算されます。自分で値を指定するには、テーブル・セルをクリックしてドロップダウン・リストから新規の幅を選択します。ここに表示されない任意の幅を入力するには、「フィールド」または「形式」列のテーブル・セルをダブルクリックして、フィールド設定のサブダイアログ・ボックスを開きます。代わりに、セルを右クリックして 「形式を設定」 を選択することもできます。

現在のフィールドを表示、デフォルトでは、このダイアログ・ボックスには現在アクティブなフィールドが表示されます。使われていないフィールドを表示する場合は、代わりに 「未使用のフィールド設定を表示」 を選択します。

コンテキスト・メニュー。このタブで利用できるコンテキスト・メニューには、さまざまな選択項目やオプションが用意されています。列内で右クリックすると、このメニューが表示されます。

- すべて選択。すべてのフィールドを選択します。
- すべて選択解除。選択を解除します。
- フィールドの選択。データ型またはストレージの種類を基準にしてフィールドを選択します。選択できるオプションには、「カテゴリー型の選択」、「連続型の選択」 (数値)、「データ型不明の選択」、「文字列の選択」、「数字の選択」、または 「日付/時刻の選択」 があります。 詳しくは、トピック 145 ページの『尺度』を参照してください。
- 形式を設定。フィールドごとに日付、時刻や小数のオプションを指定するには、サブダイアログ・ボックスを表示します。
- 表示位置の設定。選択したフィールドの表示位置を設定します。「自動」、「中央」、「左」、または「右」 を選択することができます。
- 列幅の設定。選択したフィールドのフィールド幅を設定します。データから幅を読み込むには 「自動」 を指定します。あるいはフィールド幅を 5、10、20、30、50、100 または 200 に設定することもできます。

フィールド形式のオプションの設定

フィールド形式は、データ型またはテーブル・ノード「形式」タブからサブダイアログ・ボックスで指定します。このダイアログ・ボックスの表示前に複数のフィールドを選択した場合、選択した最初のフィールドの設定がすべてに使用されます。ここで指定した後 「**OK**」 をクリックして、これらの設定を「形式」タブで選択したすべてのフィールドに適用します。

次の各オプションがフィールドごとに利用可能です。これらの設定の多くは、「ストリームのプロパティー」ダイアログ・ボックスでも設定できます。 フィールド レベルの設定は、ストリームのデフォルト設定をオーバーライドします。

日付のフォーマット。日付ストレージ (記憶域) フィールドが使用する、または文字列が CLEM 日付関数によって日付として解釈された場合に使用する日付の形式を選択します。

時間のフォーマット。時間ストレージ (記憶域) フィールドが使用する、または文字列が CLEM 時間関数によって時間として解釈された場合に使用する時間の形式を選択します。

数値の表示フォーマット。標準 (**####.###**)、科学的 (**#.###E+##**)、または通貨表記形式 (**$###.##**) から選択できます。

小数点記号。桁区切り記号として、カンマ (,)、またはピリオド (.) を選択します。

グループ化記号。数値の表示フォーマットで、値をグループ化するのに使用する記号を選択します (例: *3,000.00* のカンマ)。オプションには、なし、ピリオド、カンマ、スペース、およびロケール定義 (現在のロケールがデフォルトとして使用されている場合)。

小数点以下の表示 **(**標準、科学的、通貨、エクスポート**)**。数値の表示フォーマットを表すために、実数を表示するときに使用する、小数点以下の桁数を指定します。このオプションは、各表示形式ごとに別々に指定します。「小数点以下のエクスポート」設定は、フラット・ファイルのエクスポートにのみ適用され、ストリームのプロパティーを上書きするということに注意してください。フラット・ファイル・エクスポートのストリームのデフォルトは、ストリームのプロパティー内の「小数点以下の表示」設定に指定した値です。XML エクスポート・ノードによってエクスポートされる小数点以下の桁数は、常に 6 です。

表示位置。列内で値をどのように表示するかを指定します。デフォルトの設定は 「自動」 で、シンボル値を左寄せ、数値を右寄せで表示します。デフォルト値を変更するには、「左」、「右」、または「中央」を選択します。

列幅。デフォルトでは、フィールドの値に基づいて列幅が自動的に計算されます。リスト・ボックスの右にある矢印を使用して、5 つの間隔における任意の幅を指定できます。

フィールドのフィルタリングまたは名前の変更

ストリーム内の任意の場所でフィールドの名前変更またはフィールドを除外することができます。例えば、患者 (レコード レベル・データ) の カリウム値 (フィールド レベル・データ) を重要視していない場合、「K」 (カリウム値) フィールドを除外できます。個々のフィルター・ノード、または入力ノードか出力ノードの「フィルター」タブを使用して実行することができます。どのノードからアクセスしているかに関係なく、機能は同じです。

- 可変長ファイル、固定長ファイル、Statistics ファイル、XML、または拡張インポートなどの入力ノードから、IBM SPSS Modeler にデータを読み込むフィールドの名前を変更したり、フィルタリングを行うことができます。
- フィルター・ノードを使用して、ストリーム中の任意の場所でフィールドの名前を変更したり、フィルタリングすることができます。
- Statistics エクスポート・ノード、Statistics 変換ノード、Statistics モデル・ノードおよび Statistics 出力ノードから、IBM SPSS Statistics の命名規則に従ったフィールドをフィルタリングしたり、名前を変更することができます。 詳しくは、トピック 388 ページの『IBM SPSS Statistics 用のフィールドの名前変更またはフィルタリング』を参照してください。
- 上記のノードいずれかの「フィルター」タブを使用して、複数の回答セットを定義または編集することができます。 詳しくは、トピック 161 ページの『複数回答セット編集』を参照してください。
- 最後にフィルター・ノードを使用して、ある入力ノードから別のノードへフィールドをマップすることができます。

フィルタリング・オプションの設定

「フィルター」タブ中のテーブルには、入力時の各フィールド名、および出力時の各フィールド名が表示されます。このテーブル中のオプションを使用して、重複しているフィールドや、下流の操作で不要なフィールドの、フィールド名を変更したり、フィールドをフィルタリングすることができます。

- フィールド**:** 現在接続しているデータ・ソースからの入力フィールドを表示します。
- フィルター**:** すべての入力フィールドのフィルター・ステータスを表示します。フィルタリングされているフィールドはこの列に、フィールドが下流に渡されないことを示す赤い X 印が表示されます。選択

したフィールドの 「フィルター」 列をクリックすると、フィルタリングの設定と解除を行えます。また、Shift キーを押しながら選択することにより、複数のフィールドのオプションを同時に設定することができます。

- フィールド: フィルター・ノードから出力されるフィールドを表示します。重複した名前は赤で表示されます。フィールド名を編集するには、この列をクリックして新しい名前を入力します。または、「フィルター」 列をクリックして重複するフィールドを無効にすることにより、フィールドを削除することもできます。

テーブル中のすべての列は、列見出しをクリックしてソートすることができます。

現在のフィールドを表示 :フィルター・ノードに接続しているデータ・セットのフィールドを表示する場合に選択します。このオプションはデフォルトで選択されており、フィルター・ノードを使用するもっとも一般的な方法となっています。

未使用のフィールド設定を表示 :フィルター・ノードに接続したことがある (今は接続していない) データ・セットのフィールドを表示する場合に選択します。このオプションは、フィルター・ノードをあるストリームから別のストリームにコピーしたり、フィルター・ノードを保存して再ロードするような場合に適しています。

「フィルター」ボタンのメニュー

ダイアログ・ボックスの左上にある「フィルター」ボタンをクリックして、多くのショートカットやその他のオプションを提供するメニューにアクセスします。

次の処理を選択できます。

- すべてのフィールドを削除
- すべてのフィールドを含める
- すべてのフィールドの切り替え
- 重複を削除。このオプションを選択すると、重複する名前のすべての出現が (初出の名前も含めて) 削除されることに注意してください。
- その他のアプリケーションに準拠させるため、フィールドおよび複数の回答セットの名前を変更。 詳しくは、トピック 388 ページの『IBM SPSS Statistics 用のフィールドの名前変更またはフィルタリング』を参照してください。
- フィールド名の短縮
- フィールドおよび複数回答セット名の匿名化
- 入力フィールド名を使用
- 複数回答セット編集。 詳しくは、トピック 161 ページの『複数回答セット編集』を参照してください。
- デフォルト・フィルター・ステートの設定

ダイアログ・ボックスの上部にある矢印のトグル ボタンを使用して、デフォルトでフィールドを含める、またはフィールドを破棄するかを指定することもできます。これは、大きいデータ・セットで、ほんのわずかなフィールドだけが下流で必要な場合に役立ちます。例えば保持しておきたいフィールドのみを選択して、その他のすべてのフィールドを (個別にではなく、破棄するフィールドをすべて選択して) 破棄するよう指定できます。

フィールド名の短縮

「フィルター」ボタンのメニュー (「フィルター」タブの左上部) から、フィールド名の短縮を選択できます。

最大長**:** フィールド名の長さを制限するための文字数を指定します。

数字の桁数**:** フィールド名を短縮するとフィールド名が重複する場合は、名前をさらに短縮して数字を追加することにより区別します。使用する数字の桁数を指定できます。矢印ボタンを使用して、数を調節してください。

例として、デフォルトの設定 (最大文字数 = 8、数字の桁数 = 2) を使用した場合に医療データ・セットのフィールド名がどのように短縮されるかを次の表に示します。

表 *23.* フィールド名の短縮

フィールド名	短縮されたフィールド名
Patient Input 1	Patien01
Patient Input 2	Patien02
Heart Rate	HeartRat
BP	BP

フィールド名の匿名化

「フィルター」タブの左上部にある「フィルター」ボタン・メニューをクリックして 「フィールド名の匿名化」 を選択し、「フィルター」タブを含むノードのフィールド名を匿名化します。匿名化したフィールド名は、一意の数字の基づいた値の接頭辞を含む文字列から成り立っています。

匿名化対象**:** 「選択されたフィールドのみ」を選択します。これにより、「フィルター」タブで既に選択されているフィールドの名前だけが匿名化されます。デフォルトは 「すべてのフィールド」 で、全フィールド名を匿名化します。

フィールド名プレフィックス**:** 匿名化されたフィールド名に対するデフォルトのプレフィックスは **anon_** です。別の接頭辞が必要な場合は、「カスタム」を選択して独自の接頭辞を入力してください。

複数回答セットを匿名化**:** フィールドと同じ方法で、複数回答セットの名前を匿名化します。 詳しくは、トピック『複数回答セット編集』を参照してください。

元のフィールド名で保存するには、「フィルター」ボタン・メニューから 「入力フィールド名を使用」 を選択します。

複数回答セット編集

「フィルター」タブの左上部にある「フィルター」ボタン・メニューをクリックして 「複数回答セット編集」 を選択して、複数の回答セットを追加または編集します。

例えば、どの博物館に行くのか、またはどの雑誌を読むのかに関する調査の回答を求める場合など、複数回答セットを使用して、各ケースに複数の値を持つデータを記録します。複数回答セットは、Data Collection ソース・ノードまたは Statistics ファイル・ソース・ノードを使用して IBM SPSS Modeler にインポートし、フィルター・ノードを使用して IBM SPSS Modeler に定義することができます。

「新規作成」 をクリックして新しい複数回答セットを作成するか、または 「編集」 をクリックして既存の設定を変更します。

名前とラベル: セットの名前と説明を指定します。

タイプ: 複数回答の質問を、次の 2 つのいずれかの方法で処理することができます。

- 複数二分法設定: 個々のフラグ型フィールドがそれぞれの回答に対して作成されます。10 冊の雑誌がある場合、10 のフラグ型フィールドがあり、それぞれに真 *(true)* または偽 *(false)* を表す 0 または 1 などの値が含まれます。カウントされた値を使用して、真 (true) としてカウントされる値を指定することができます。この方法は、回答者が適用されるすべてのオプションを選択できるようにしたい場合に役に立ちます。
- 複数カテゴリー設定: 名義型フィールドが、各回答に対して回答者からの最大回答数まで作成されます。各名義型フィールドには、「*Time*」の *1*、「*Newsweek*」の *2*、および「*PC Week*」の *3* などの考えられる回答を表す値が含まれます。この方法は、例えば回答者にもっともよく読む 3 冊の雑誌を選択するよう質問する場合など、回答数を制限する場合に最も役に立ちます。

設定のフィールド: 右側のアイコンを使用して、フィールドを追加または削除します。

コメント

- すべての複数回答セットのすべてのフィールドは同じ保存タイプを持たなければいけません。
- セットは含まれるフィールドとは異なります。例えば、セットを削除しても含まれるフィールドは削除されず、これらのフィールド間のリンクのみが削除されます。セットは削除のポイントから上流では表示されますが、下流では表示されません。
- フィルター・ノードを使用してフィールド名を変更した場合 (タブで直接変更した場合、または「フィルター」メニューの「IBM SPSS Statistics 用に名前変更」、「短縮」、「匿名化」のいずれかのオプションを選択して変更した場合)、多重回答セットで使用されているこれらのフィールドへの参照もすべて更新されます。ただし、フィルター・ノードで削除される複数回答セットのフィールドは、複数回答セットからは削除されません。このようなフィールドは、ストリーム内では表示されませんが、複数回答セットによって参照されます。このことは、例えばエクスポート時に検討されます。

フィールド作成ノード

IBM SPSS Modeler の強力な機能の 1 つとして、データ値を変更して、既存のデータから新しいフィールドを作成できることがあげられます。長期間にわたるデータ・マイニング・プロジェクトでは、Web ログ・データから顧客 ID を抽出したり、トランザクション・データや人口統計データから顧客の生涯価値を作成するような、新しいデータの作成が一般的に行われます。これらの変換はすべて、さまざまなフィールド設定ノードを使用して行うことができます。

さまざまなノードで、新規フィールドを作成することができます。

フィールド作成ノードで、1 つまたは複数の既存フィールドから、データ値を変更するか、新しいフィールドを作成します。これで、 タイプ式、フラグ、名義、ステート、カウント、および条件式の各フィールドが作成されます。

データ分類ノードにより、あるカテゴリー値のセットが別のセットに変換されます。データ分類は、カテゴリーを再編成したり、分析用のデータをグループ化しなおす場合に役立ちます。

データ分割ノードで、既存の 1 つまたは複数の連続型 (数値範囲) フィールドの値に基づいて、自動的に新しい名義型 (セット型) フィールドを作成します。例えば、連続型収入フィールドを、平均からの偏差による収入グループを含む、新しいカテゴリー・フィールドに変換することができます。新規フィールドのビンを作成すると、分割点に基づいてフィールド作成ノードを生成することができます。

フラグ設定ノードで、1 つ以上の名義型フィールドに定義されたカテゴリー値に基づいた、複数のフラグ型フィールドが派生します。

再構成ノードで、名義型またはグラグ型フィールドを、これから別のフィールドの値で埋めることができるフィールドのグループへ変換します。例えば、*credit*、*cash*、および *debit* の値の *payment type* という名前のフィールドがある場合、3 つの新しいフィールド (*credit*、*cash*、*debit*) が作成されます。その各々には、実際の支払の値を含めることができます。

時系列ノードにより、以前レコードのフィールドのデータを含む、新規フィールドが作成されます。時系列ノードは、多くの場合、時系列データなどの継続的なデータに使用されます。時系列ノードを使用する前に、ソート・ノードを使用して、データをソートしておくこともできます。

フィールド作成ノードの使用

フィールド作成ノードを使用して、1 つまたは複数の既存フィールドから、6 種類の新規フィールドを作成することができます。

- 式: 新規フィールドは任意の CLEM 式の結果です。
- フラグ型: 新規フィールドは、指定した条件を示すフラグ型です。
- 名義: 新規フィールドは名義型であり、そのメンバーは指定した値のグループです。
- 状態。新規フィールドは、2 つの状態 (ステート) のどちらかです。これら 2 つの状態は指定した条件によって切り替えられます。
- カウント: 新規フィールドは、条件を満たした回数を基準に決まります。
- 条件式: 新規フィールドは、条件の値に応じて、2 つの式のどちらかの値になります。

これらのノードには、それぞれの「フィールド作成ノード」ダイアログ・ボックスに、特別な一連のオプションが用意されています。これらのオプションは、以降の各項目で説明しています。

以下を使用すると、行の順序が変わる場合があるので注意してください。

- SQL プッシュバックによるデータベース内での実行
- リモートの IBM SPSS Analytic Server による実行
- 組み込みの IBM SPSS Analytic Server で実行する関数の使用
- リストの作成 (例については 166 ページの『リスト フィールドまたは地理空間フィールドの作成』を参照)
- 空間処理関数で説明されている関数の呼び出し

フィールド作成ノードの基本オプションの設定

フィールド作成ノードのダイアログ・ボックスの上部には、必要なフィールド作成ノードのデータ型を選択するさまざまなオプションが用意されています。

モード**:** 複数のフィールドを作成するかどうかに応じて、「単一」または「複数」を選択します。「複数」を選択した場合、ダイアログ・ボックスには複数の派生フィールド用のオプションが表示されます。

派生フィールド。単純なフィールド作成ノードに対して、作成して各レコードに追加するフィールド名を指定します。デフォルトのフィールド名は Derive*N* です。ここで、*N* は現在のセッション中でここまでに作成されたフィールド作成ノード数を表します。

データ型。ドロップダウン・リストから、CLEM 式や名義型などのフィールド作成ノードのデータ型を選択します。それぞれのデータ型に対応したダイアログ・ボックスで指定した条件に基づいて、新規フィールドが作成されます。

ドロップ・ダウン・リストでオプションを選択すると、各データ型のプロパティーに応じて、フィールド作成ノードのダイアログ・ボックスに異なるオプションが表示されます。

フィールドのデータ型。連続型、カテゴリー型、またはフラグ型など、新しい派生フィールドの尺度を選択します。このオプションは、すべてのフィールド作成ノードで共通です。

注 : 新規フィールドの作成には、しばしば特殊関数や数式が必要なことがあります。これらの式の作成を支援するために、すべての種類のフィールド作成ノードのダイアログ・ボックスから利用することができ、式の検査機能やすべての CLEM 式の一覧を提供する CLEM 式ビルダーが用意されています。

複数フィールドの作成

フィールド作成ノード内でモードを 「複数」 に設定すると、同じノード内で同じ条件に基づいて複数のフィールドを作成できます。この機能を利用すれば、データ・セット中の複数のフィールドで同じ変換処理を行う場合に、時間を節約することができます。例えば、初任給と以前の経歴に基づいて現在の給与を予測する回帰モデルを作成する場合、3 つの非対称変数すべてに対数変換を適用するために役立ちます。各変換に新しいフィールド作成ノードを追加する代わりに、すべてのフィールドに一度に同じ関数を適用できます。新しいフィールドの作成元となるすべてのフィールドを選択し、フィールドのカッコ内で `@FIELD` 関数を使用してフィールド作成式を入力します。

注 : `@FIELD` 関数は、複数のフィールドを同時に作成するために重要なツールです。このツールを使用すれば、正確なフィールド名を指定することなく現在のフィールドの内容を参照することができます。例えば、複数のフィールドに対して対数変換を適用するために用いられる CLEM 式は、`log(@FIELD)` になります。

「複数」 モードを選択すると、ダイアログ・ボックスに次のオプションが表示されます。

フィールド・リスト**:** フィールド・ピッカーを使用して、新規フィールドを作成するフィールドを選択します。選択した各フィールドに対して、1 つの出力フィールドが生成されます。注 : 選択するフィールドがそれぞれ同じストレージ・タイプである必要はありません。ただし、すべての フィールドに対して条件が有効でなければ、フィールド作成操作は失敗してしまいます。

フィールド名拡張子**:** 新しいフィールド名に追加する拡張子を入力します。例えば、*Current Salary* の対数を含む新しいフィールドの場合、フィールド名に拡張子 *log_* を追加すると、*log_Current Salary* が生成されます。ラジオ・ボタンを使用して、拡張子をフィールドの接頭辞 (先頭) として追加するか、接尾辞 (最後) として追加するかを指定します。デフォルトのフィールド名は Derive*N* です。ここで、*N* は現在のセッション中でここまでに作成されたフィールド作成ノード数を表します。

単一モードのフィールド作成ノードと同様に、新しいフィールドの作成に使用する CLEM 式を作成する必要があります。選択したフィールド作成操作の種類に応じて、条件を作成するためのさまざまなオプションがあります。これらのオプションは、以降の各項目で説明しています。CLEM 式を作成するには、CLEM 式フィールドに直接入力するか、または計算機ボタンをクリックして Clem 式ビルダーを使用してください。複数のフィールドの操作を参照する場合は、`@FIELD` 関数を使用してください。

複数フィールドの選択

フィールド作成ノード (複数モード)、レコード集計ノード、ソート・ノード、線グラフ・ノード、時系列ノードなどの、複数の入力フィールドに対して操作を実行するすべてのノードで、次に示す「フィールドの選択」ダイアログ・ボックスを使用して複数のフィールドを簡単に選択することができます。

ソート項目**:** 次のオプションを使って、表示する利用可能フィールドをソートすることができます。

- ファイル順**:** データ・ストリームから現在のノードにデータが渡された順序にフィールドをソートします。
- 「名前」 アルファベット順でフィールドをソートします。
- データ型**:** 尺度でソートしたフィールドを表示します。特定の尺度のフィールドを選択する場合に役立ちます。

リストからフィールドを 1 回に 1 つずつ選択するか、または Shift キーまたは Ctrl キーを押しながら複数のフィールドを選択します。また、リストの下のボタンを使用して、尺度に基づいて複数のフィールドを選択したり、テーブル中のすべてのフィールドを選択または選択解除することができます。

CLEM 式作成オプションの設定

フィールド作成ノード (CLEM 式型) は、CLEM 式の結果を基にして、データ・セット内の各レコード用に新規フィールドを作成します。この式を条件式にすることはできません。条件式に基づいて値を作成するには、フラグ型または条件式型のフィールド作成ノードを使用します。

CLEM 式 新規フィールドの値を作成するために CLEM 言語を使用して式を指定します。

注**:** SPSS Modeler は作成されたリスト フィールドに使用するサブ尺度を認識できないため、集合および地理空間の尺度の場合は、「指定**...**」をクリックして「値」ダイアログ ボックスを開き、必要なサブ尺度を設定することができます。詳しくは、『作成するリスト値の設定』を参照してください。

地理空間フィールドの場合、該当するストレージ タイプは「実数」と「整数」のみです (デフォルト設定は「実数」です)。

作成するリスト値の設定

フィールド作成ノードの「CLEM 式」の「フィールドのデータ型」ドロップダウン リストから「指定**...**」を選択すると、「値」ダイアログボックスが表示されます。このダイアログボックスでは、集合または地理空間のいずれかの CLEM 式の「フィールドのデータ型」尺度に使用するサブ尺度の値を設定します。

尺度 「集合 **(Collection)**」または「地理空間」のいずれかを選択します。他の尺度を選択すると、編集可能な値がないというメッセージがダイアログ ボックスに表示されます。

集合

集合の「尺度」に設定できる項目は「尺度のリスト」のみです。デフォルトではこの尺度が「不明」に設定されていますが、別の値を選択して、リスト内の要素の尺度を設定することができます。以下のいずれかのオプションを選択できます。

- 不明
- カテゴリー型
- 連続
- 名義
- 順序
- フラグ

地理空間

地理空間の「尺度」の場合は、以下のオプションを選択して、リスト内の要素の尺度を設定することができます。

データ型 地理空間フィールドのサブ尺度を選択します。使用可能なサブ尺度は、リスト フィールドの深さによって決まります。デフォルト値は以下のとおりです。

- ポイント (深さ 0)
- 行ストリング (深さ 1)
- 多角形 (深さ 1)
- 複数点 (深さ 1)
- 複数行ストリング (深さ 2)
- 多角形群 (深さ 2)

サブ尺度について詳しくは、「SPSS Modeler Source, Process, and Output Nodes Guide」の『Type Node』セクションにある『Geospatial Measurement Sublevels』というトピックを参照してください。

リストの深さについて詳しくは、「SPSS Modeler Source, Process, and Output Nodes Guide」の『Source Node』セクションにある『List Storage and Associated Measurement Levels』というトピックを参照してください。

座標系 このオプションは、尺度を地理空間以外から地理空間に変更した場合にのみ使用可能です。座標系を地理空間データに適用するには、このチェック ボックスを選択します。デフォルトでは、「ツール」 > 「ストリームのプロパティー」 > 「オプション」 > 「地理空間」ペインで設定された座標系が表示されます。別の座標系を使用するには、「変更」ボタンをクリックして「座標系の選択」ダイアログボックスを表示し、データに一致する座標系を選択します。

座標系について詳しくは、「SPSS Modeler User's Guide」の『Working with Streams』セクションに記載されている 『Setting Geospatial Options for Streams』というトピックを参照してください。

リスト フィールドまたは地理空間フィールドの作成

場合によっては、リスト項目として記録すべきデータが、間違った属性で SPSS Modeler にインポートされることがあります。例えば、x 座標と y 座標、経度と緯度などが .csv ファイルの個々の行として別々の地理空間フィールドになっている場合があります。この場合は、個々のフィールドを結合して単一のリスト フィールドにしなければなりません。これを行う方法の 1 つとして、フィールド作成ノードを使用する方法があります。

注**:** 地理空間データを結合する場合は、どのフィールドが x (経度) フィールドで、どのフィールドが y (緯度) フィールドかを知っておく必要があります。また、結果として得られるリスト フィールドの要素の順序が、地理空間座標の標準形式である [x, y] ([経度, 緯度]) になるようにデータを結合する必要があります。

リスト フィールドを作成する簡単な例を以下のステップに示します。

1. ストリームでフィールド作成ノードをソース・ノードに接続します。
2. フィールド作成ノードの「設定」タブで、「データ型」リストから「**CLEM** 式」を選択します。
3. 「フィールドのデータ型」で、地理空間以外のリストの場合は「集合 **(Collection)**」を選択し、そうでなければ「地理空間」を選択します。デフォルトでは、 SPSS Modeler は「最良推測」の手法を使用して正しいリストの詳細情報を設定します。「指定**...**」を選択すると、「値」ダイアログ ボックスが開きます。集合の場合、このダイアログを使用して、データに関する詳細情報をリスト内に入力できます。地理空間の場合、このダイアログを使用して、データ型を設定したり、データの座標系を指定したりできます。

   注**:** 地理空間の場合、指定する座標系は、データの座標系と完全に一致する必要があります。そうでないと、地理空間の機能で不正な結果が生成されます。
4. 「**CLEM** 式」ペインに、データを結合して正しいリスト形式にするための式を入力します。または、計算機ボタンをクリックして Clem 式ビルダーを開きます。

   リストを作成する式の簡単な例としては、[x, y] があります。この x と y は、データ ソース内の個別のフィールドです。作成される新しいフィールドは、リストです。このリストでは、各レコードの値が、そのレコードの連結された x 値と y 値になります。

   注**:** このように結合してリストにするフィールドは、ストレージ タイプが同じでなければなりません。

リストおよびリストの深さについて詳しくは、 12 ページの『リストス トレージおよび関連する尺度』を参照してください。

フィールド作成ノード (フラグ型) のオプションの設定

フィールド作成ノード (フラグ型) は、高血圧や使用されていない顧客アカウントなどの、特定の条件を示すために使用されます。フラグ・フィールドは各レコードに対して作成され、真 (true) の条件を満たす場合に、フィールドに真 (true) のフラグ値が追加されます。

真 **(true)** の値**:** 下のフィールドで指定した条件と一致するレコードの、フラグ型フィールドに入れる値を指定します。デフォルトは T です。

False 値**:** 下のフィールドで指定した条件に一致しない レコードのフラグ型フィールドに入れる値を指定します。デフォルトは F です。

真 **(true)** となる条件**:** 各レコードの特定の値を評価して、真 (True) の値または偽 (False) の値 (上記) を与えるために使用する CLEM 条件を指定します。偽 (false) ではない数値に対しては、真 (true) の値がレコードに与えられることに注意してください。

注 : 空文字列を返すには、"" のように中に何も入れずに引用符を 2 つ指定します。空文字列は、例えばテーブル中で真 (true) の値をより際だたせるために、偽 (true) の値として使用されます。同様に、数値として扱われる値を文字列値として使用する場合にも、引用符で囲む必要があります。

例

IBM SPSS Modeler 12.0 以前のリリースでは、値をカンマで区切った複数の回答を単一のフィールドにインポートしていました。以下に例を示します。

```
museum_of_design,institute_of_textiles_and_fashion
museum_of_design
archeological_museum
$null$
national_art_gallery,national_museum_of_science,other
```

分析の目的でこのデータを準備するには、それぞれの回答に次のような式を使用して個別のフラグ型フィールドを生成するために、hassubstring 関数を使用できます。

```
hassubstring(museums,"museum_of_design")
```

フィールド作成ノード (名義型) のオプションの設定

各レコードがどの条件を満たすかを判断するには、フィールド作成ノード (名義型) を使用して一連の CLEM 条件を実行します。各レコードが条件を満たすと、値 (どの条件のセットを満たしたかを示す) が新しく作成されたフィールドに追加されます。

デフォルト値: 一致する条件がない場合に、新規フィールドで使用する値を指定します。

フィールド設定値: 特定の条件を満たした場合に、新規フィールドに入れる値を指定します。リスト中の各値には、対応する条件があり、この条件は隣接する列で指定します。

値の設定条件 (真の場合に値を設定): セット型フィールド中の各メンバーに対する条件を指定します。Clem 式ビルダーを使用して、利用可能な関数とフィールドを選択します。矢印および削除ボタンを使用して、条件を並べ替えたり、削除することができます。

データ・セット内の特定のフィールドの値が、条件に該当するかどうかを調べます。各条件が調べられると、どの条件を満たしたかを示すために (満たした場合)、上のフィールドで指定した値が新規フィールドに割り当てられます。どの条件も満たさなかった場合は、デフォルト値が使用されます。

フィールド作成ノード (ステート型) のオプションの設定

フィールド作成ノード (ステート型) は、フィールド作成ノード (フラグ型) に似ています。フィールド作成ノード (フラグ型) は、現在のレコードが *1* つ の条件を満たすかどうかを示す値を設定します。これに対し、フィールド作成ノード (ステート型) では、*2* つの独立した 条件がどのように満たされるかに応じてフィールドの値が変化します。つまり、各条件を満たすと値が変わります (オンまたはオフになる)。

初期ステート: 新規フィールドの各レコードに初期ステートとしてオンまたはオフのどちらの値を与えるかを選択します。この値は、各条件を満たす場合に変わることに注意してください。

「オン」の値: オンの条件を満たす場合の新規フィールドの値を指定します。

スイッチ「オン」の条件式: 条件が真 (true) の場合にステートをオンに変更する CLEM 条件を指定します。計算機ボタンをクリックして、式ビルダーを開きます。

「オフ」の値: オフの条件を満たす場合の新規フィールドの値を指定します。

スイッチ「オフ」の条件式: 条件が偽 (false) の場合にステートをオフに変更する CLEM 条件を指定します。計算機ボタンをクリックして、式ビルダーを開きます。

注 : 空の文字列を指定するには、"" のように引用符を 2 つ入力してください。引用符の間には何も指定しないでください。同様に、数値として扱われる値を文字列値として使用する場合にも、引用符で囲む必要があります。

フィールド作成ノード (カウント型) のオプションの設定

フィールド作成ノード (カウント型) は、データ・セット内の数値型フィールドの値に一連の条件を適用するために使用します。各条件を満たすたびに、作成されたカウント型フィールドの値が設定された増分の値ずつ増やされます。このデータ型のフィールド作成ノードは時系列データの場合に役に立ちます。

初期値: 実行時に新規フィールド用に使用する値を設定します。初期値は数値の定数にする必要があります。値を増やしたり減らすには、矢印ボタンを使用します。

増分条件: CLEM 条件を指定します。この条件が満たされた場合、「増分」で指定した数値に基づいて作成値が変更されます。計算機ボタンをクリックして、式ビルダーを開きます。

増分: カウントを増やすための値を設定します。数値の定数または CLEM 式の結果のどちらかを使用できます。

リセット条件: この条件を満たす場合に、作成値を初期値にリセットする条件を指定します。計算機ボタンをクリックして、式ビルダーを開きます。

フィールド作成ノード (条件式型) のオプションの設定

フィールド作成ノード (条件式 (If-Then) 型) は一連の If-Then-Else 文を使用して、新規フィールドの値を作成します。

If: 実行時に各レコードを評価する CLEM 条件を指定します。条件が真 (数値のときは偽以外) の場合は、Then の式によって下のフィールドに指定された値が新規フィールドに与えられます。計算機ボタンをクリックして、式ビルダーを開きます。

Then: 上記の If 文が true (または false 以外) の場合に新規フィールドに入れる値または CLEM 式を指定します。計算機ボタンをクリックして、式ビルダーを開きます。

Else: 上記の If 文が false の場合に新規フィールドに入れる値または CLEM 式を指定します。計算機ボタンをクリックして、式ビルダーを開きます。

フィールド作成ノードを使用して値を再コード化する

フィールド作成ノードは、例えばカテゴリー型の値を持つ文字列フィールドを数値名義 (セット) 型フィールドに変換して値を再コード化するために使用できます。

1. 「データ型」については、該当するフィールドのタイプ (名義型、フラグ型など) を選択します。
2. 値を再コード化するための条件を指定します。例えば、`Drug='drugA'` の場合は 1、`Drug='drugB'` の場合は 2、などのように設定することができます。

置換ノード

置換ノードは、フィールド値の置換やストレージの変更に使用されます。`@BLANK(FIELD)` のような、CLEM 条件に基づいて値を置換することができます。また、すべての空白値やヌル値を特定の値に置換することもできます。置換ノードは、データ型ノードと組み合わせて、欠損値を置き換えるためによく使用されます。例えば、`@GLOBAL_MEAN` のような式を指定することによって、空白をフィールドの平均値で置き換えることができます。この式は、すべての空白値をグローバル・ノードで算出した平均値に置き換えます。

対象フィールド: フィールド・ピッカー (テキスト・フィールドの左側にあるボタン) を使用して、値を調査、置換するデータ・セットのフィールドを選択します。デフォルトでは、下の「条件」および「置換値」の指定、および関連する式に基づいて、値が置換されます。また、下の「置換」オプションを使用して、置換の代替手法を選択することもできます。

注: ユーザーが定義した値で置換するフィールドを複数選択する場合は、フィールドのデータ型が同じでなければなりません (すべて数値またはすべてシンボル値)。

置換: 次のいずれかの方法を使用して、選択したフィールドの値を置換する場合に選択します。

- 条件を指定: このオプションを選択すると、「条件」フィールドと Clem 式ビルダーがアクティブになります。これを使用して、指定した値と置換するための条件として使用する式を作成します。
- 常時: 選択したフィールドのすべての値を置換します。例えば、このオプションを使用して、CLEM 式の (to_string(income)) により、income のストレージを文字列に変換することができます。
- 空白値: 選択したフィールドの、ユーザーが指定したすべての空白値を置換します。空白を選択するには、標準の条件 @BLANK(@FIELD) を使用します。注 : ソース・ノードまたはデータ型ノードを使用して、空白値を定義することができます。
- ヌル値: 選択したフィールドの、すべてのヌル値を置換します。Null 値を選択するには、標準の条件 NULL(@FIELD) が使用されます。
- 空白値とヌル値: 選択したフィールドの空白値とシステム ヌル値の両方を置換します。このオプションは、ヌルが欠損値として定義されているかどうか不明な場合などに役立ちます。

条件: 「条件を指定」オプションを選択した場合は、このオプションを使用することができます。このテキスト・ボックスには、選択したフィールドを評価するための CLEM 式を指定します。計算機ボタンをクリックして、式ビルダーを開きます。

置換値: 選択したフィールドに新しい値を設定する CLEM 式を指定します。テキスト・ボックスに undef と入力することにより、値をヌル値で置き換えることもできます。計算機ボタンをクリックして、式ビルダーを開きます。

注: 選択したフィールドが文字型の場合は、文字型の値で置換する必要があります。文字型フィールドに対して、置換値にデフォルトの 0 または他の数値を使用すると、エラーが発生してしまいます。

以下を使用すると、行の順序が変わる場合があるので注意してください。

- SQL プッシュバックによるデータベース内での実行
- リモートの IBM SPSS Analytic Server による実行
- 組み込みの IBM SPSS Analytic Server で実行する関数の使用
- リストの作成 (例については 166 ページの『リスト フィールドまたは地理空間フィールドの作成』を参照)
- 空間処理関数で説明されている関数の呼び出し

置換ノードを使ったストレージの変換

置換ノードの「置換」条件を使えば、単一または複数のフィールドのストレージ・タイプを簡単に変換することができます。例えば変換関数 to_integer を使用すると、CLEM 式の to_integer(income) を使用して *income* を文字列から整数に変換することができます。

Clem 式ビルダーを使えば、利用できる変換関数を表示したり、CLEM 式を自動的に作成することができます。「関数」ドロップダウン・リストで、「変換」 を選択すると、ストレージ変換関数が一覧表示されます。利用できる変換関数を次に示します。

- to_integer(ITEM)
- to_real(ITEM)
- to_number(ITEM)
- to_string(ITEM)
- to_time(ITEM)
- to_timestamp(ITEM)
- to_date(ITEM)
- to_datetime(ITEM)

日付と時刻の値の変換 : 変換の関数と日付や時刻の値のような、入力に特別な型が必要なその他の関数は、「ストリームのオプション」ダイアログ・ボックスに指定されている現在の形式に依存します。例えば、値が *Jan 2003*、*Feb 2003* などの文字列フィールドを日付ストレージへ変換する場合、ストリームのデフォルトの日付形式として 「**MON YYYY**」 を選択します。

変換関数は、フィールド作成ノードのフィールド作成計算時の一時変換でも、利用できます。また、フィールド作成ノードを使用して、カテゴリー値を含む文字列フィールドの読み取りなど、他の操作も実行できます。 詳しくは、トピック 169 ページの『フィールド作成ノードを使用して値を再コード化する』を参照してください。

データ分類ノード

データ分類ノードにより、あるカテゴリー値のセットを別のセットに変換することができます。データ分類ノードは、カテゴリーを縮小したり、分析用にデータをグループ化し直したりする場合に役立ちます。例えば、「商品」 の値を、キッチン用品、バス用品、および電化製品の 3 種類のグループに分類しなおすことができます。しばしばこの操作は、値をグループ化してデータ分類ノードを生成することにより、棒グラフ・ノードから直接行われます。 詳しくは、トピック 254 ページの『棒グラフ・ノードの使用方法』を参照してください。

データ分類は、1 つまたは複数のシンボル値フィールドに対して実行することができます。また、既存のフィールドを新しい値で置き換えたり、新規フィールドを生成することもできます。

データ分類ノードをいつ使用するか

データ分類ノードを使用する前に、その作業により適している他のフィールド設定ノードがないかどうかを検討してください。

- 数値範囲型 (ランクやパーセンタイルなど) をセット型に自動的に変換するには、データ分割ノードを使用する必要があります。詳しくは、トピック 176 ページの『データ分割ノード』を参照してください。
- 数値範囲型をセット型に手作業で分類するには、フィールド作成ノードを使用する必要があります。例えば、給与の値を特定の給与範囲カテゴリーに分類する場合は、フィールド作成ノードを使用して各カテゴリーを手作業で定義する必要があります。
- *Mortgage_type* のようなカテゴリー・フィールドの値に基づいて 1 つまたは複数のフィールドを作成するには、フラグ設定ノードを使用する必要があります。
- カテゴリー型フィールドを数値型ストレージに変換するには、フィールド作成ノードを使用できます。例えば、*No* と *Yes* 値を、それぞれ 0 と 1 に変換することができます。 詳しくは、トピック 169 ページの『フィールド作成ノードを使用して値を再コード化する』を参照してください。

データ分類ノードのオプション設定

データ分類ノードの使用には、次の 3 つのステップがあります。

1. まず、複数のフィールドを再分類するのか、または 1 つのフィールドを再分類するのかを選択します。
2. 次に、分類したものを既存のフィールドに再コード化するのか、または新しいフィールドを作成するのかを選択します。
3. 最後に、データ分類ノードのダイアログ・ボックスのオプションを使用して、セットを適切にマップします。

   モード**:** 1 つのフィールドのカテゴリーを再分類するには「単一」を選択します。複数のフィールドを同時に変換する場合は、「複数」 を選択します。

   データ分類先**:** 元の名義型フィールドをそのまま保持して、再分類した値を含む追加フィールドを作成する場合は、「新規フィールド」 を選択します。元のフィールドの値に新しい分類値を上書きする場合は、「既存のフィールド」 を選択します。これは本質的には、「置換」操作になります。

   モードおよび置換オプションを指定したら、ダイアログ・ボックスの下半分に表示されるオプションを使用して変換フィールドを選択し、新しい分類値を指定する必要があります。これらのオプションは、選択したモードによって異なります。

   データ分類フィールド**:** 右にあるフィールド選択ボタンを使用して、1 つ (単一モード) または複数 (複数モード) のカテゴリー型フィールドを選択します。

   新規フィールド名**:** 記録値を入れる新しい名義型フィールド名を指定します。このオプションは、単一モードで 「新規フィールド」 が選択されている場合にだけ利用できます。「既存のフィールド」 が選択されている場合は、元のフィールド名がそのまま保持されます。複数モードで作業を行う場合は、このオプションに代わって、各新規フィールドに追加する拡張子を指定するオプションが表示されます。 詳しくは、トピック 173 ページの『複数フィールドのデータ分類』を参照してください。

   データ分類値**:** このテーブルにより、古いセット値からここに指定した値へのマッピングを明確に行うことができます。

   - 元の値**:** 選択したフィールドの既存の値が表示されます。
   - 新しい値**:** 新しいカテゴリー値を入力するか、またはドロップダウン・リストから選択します。分布図からの値を使用して、データ分類ノードを自動的に生成する場合、これらの値はドロップダウン・リストに含まれます。これにより、既存の値を既知の値のセット素早くに関連付けることができます。例えば、ネットワークやロケールに基づいて、医療機関は異なる方法で診断をグループ化することがあります。合併または買収の後、すべての機関は新規のみならず既存のデータさえも一貫した方法で分類することが要求されます。長いリストから各対象値を手作業で入力しなくても、値の基本リストから IBM SPSS Modeler に読み込み、「診断」 フィールドの分布図を実行し、分布図から直接このフィールドのデータ分類ノード (値) を生成できます。このプロセスにより、「新しい値」ドロップダウン・リストのすべての対象「診断」値が利用可能になります。
4. 上で選択した、1 つまたは複数のフィールドの元の値を読み込むには、「取得」 をクリックします。
5. 元の値を、まだマップされていないフィールドの 「新しい値」列に貼り付けるには、「コピー」をクリックします。マップされていない元の値が、ドロップダウン・リストに追加されます。
6. 「新しい値」列のすべての指定内容を消去するには、「新規消去」をクリックします。注 : このオプションをクリックしても、ドロップダウン・リストの値は消去されません。

7. 元のそれぞれの値に対して、連続する整数を自動的に生成するには、「自動」 をクリックします。この場合、整数値 (1.5、2.5 などの実数値ではない) しか生成することはできません。

例えば、商品名に対する連続した商品 ID 番号や、大学の講義番号などを自動的に生成することができます。この機能は、IBM SPSS Statistics の自動再コード変換に対応しています。

未指定の場合に使用する値: このオプションは、新しいフィールドで未指定の値を置換するために使用されます。「当初の値」 を選択して元の値をそのまま保持することも、デフォルト値を指定することもできます。

複数フィールドのデータ分類

複数フィールドのカテゴリー値を同時にマップするには、「複数」 モードを選択します。複数モードにすると、「データ分類」ダイアログ・ボックスには、次の設定項目が表示されます。

データ分類フィールド: 右側にあるフィールド選択ボタンを使用して、変換するフィールドを選択します。すべてのフィールドを一度に選択することも、名義型やフラグ型のように、同じ種類のフィールドだけを選択することもできます。

フィールド名拡張子: 複数のフィールドを同時に再コード化する場合は、個別のフィールド名を指定するよりも、新しいフィールドすべてに共通の拡張子を付ける方が効率的です。「_recode」のような拡張子を指定して、拡張子を元のフィールド名の前に付けるか、または後に付けるかを選択してください。

再分類されたフィールドのストレージと尺度

データ分類ノードでは、常に再コード化操作により名義型フィールドが作成されます。そのため、状況によっては 「既存のフィールド」 の再分類を行う際に、フィールドの尺度が変更されてしまうこともあります。

新しいフィールドのストレージ (データがどのように使用される かではなく、データがどのように格納される か) は、次の「設定」タブの設定に基づいて算出されます。

- 未指定の値に対してデフォルト値を使用するように設定されている場合、新しい値とデフォルト値の両方が調べられ、適切なストレージ・タイプが決められます。例えば、すべての値が整数と判断された場合は、フィールドのストレージ・タイプは整数になります。
- 未指定の値に対して元の値を使用するように設定されている場合、ストレージ・タイプは元のフィールドのストレージに基づいて決められます。すべての値が、元のフィールドのストレージとして解析された場合は、そのストレージがそのまま保持されます。それ以外の場合は、古い値と新しい値の両方を考慮して、もっとも適切なストレージ・タイプが決められます。例えば、整数のセット { 1, 2, 3, 4, 5 } を 4 => 0, 5 => 0 で再分類すると新しい整数のセット { 1, 2, 3, 0 } が生成されますが、4 => "Over 3", 5 => "Over 3"で再分類すると、文字列のセット { "1", "2", "3", "Over 3" } が生成されます。

注 : 元のデータ型がインスタンス化されていない場合は、新しいデータ型もインスタンス化されません。

匿名化ノード

ノードのモデル下流に含まれるデータと連携している場合、匿名化ノードで、フィールド名、フィールド値のどちらかまたは両方を隠すことができます。これにより、権限を持たないユーザーが従業員の記録や患者の治療記録など機密データを閲覧する危険なく、生成されたモデルを自由に(例えばテクニカル・サポートへ)分散させることができます。

ストリームの匿名化ノードの場所によっては、他のノードに変更する必要があります。例えば選択ノードの上流に匿名化ノードを挿入する時、選択ノードの選択基準を匿名化された値に実行する場合に変更する必要があります。

匿名化で使用される方法は、様々な要素によって決まります。フィールド名、および連続型尺度以外のすべてのデータ値について、データは次の形式の文字列に置換されます。

prefix_Sn

*prefix_* は、ユーザー指定の文字列またはデフォルトの文字列 anon_ で、*n* は、0 から開始してそれぞれの一意値まで増加する整数の値です (anon_S0 や anon_S1 など)。

数値の範囲は文字列より整数または実数値に対応しているため、範囲型のフィールド値を変換する必要があります。フィールド値はその範囲を異なる範囲に変換することによってのみ匿名化することができ、元のデータを隠します。範囲内にある値 *x* の変換は、次のように行われます。

A*(*x* + B)

ここで、

A は、0 より大きい換算係数です。

B は値に追加する翻訳オフセットです。

例

換算係数 *A* が 7 に、翻訳オフセット *B* が 3 に設定されているフィールド *AGE* の場合、*AGE* の値は次のように変換されます。

7*(AGE + 3)

匿名化ノードのオプションの設定

ここでは、どのフィールドで値をより下流に隠すかを選択することができます。

データ・フィールドを匿名化ノードから上流にインスタンス化した後、匿名化処理を実行することができます。データ型ノードまたは入力ノードの「データ型」タブで 「値の読み込み」 をクリックすると、データをインスタンス化することができます。

フィールド: 現在のデータ・セットのフィールドの一覧を表示します。フィールド名がすでに匿名化されている場合、匿名化された名前がここで表示されます。

尺度。フィールドの尺度。

値を匿名化: 1 つ以上のフィールドを選択し、この列をクリックして「はい」を選択すると、デフォルトの接頭辞である **anon_** を使用してフィールド値が匿名化されます。「指定」を選択してダイアログ・ボックスを表示し、このダイアログ・ボックスで独自の接頭辞を入力することも、フィールド値の変換で乱数とユーザー指定値のいずれを使用するかを指定することもできます (連続型 のフィールド値の場合)。連続型フィールド・タイプまたは連続型でないフィールド・タイプは、同じ操作では指定できません。各フィールド タイプで個別に指定する必要があります。

現在のフィールドを表示: 匿名化ノードに接続しているデータ・セットのフィールドを表示する場合に選択します。デフォルトでは、このオプションが選択されます。

未使用のフィールド設定を表示: ノードに接続していた (現在は接続していない) データ・セットのフィールドを表示する場合に選択します。このオプションは、ノードをあるストリームから別のストリームにコピーしたり、ノードを保存して再ロードするような場合に適しています。

フィールド値の匿名化方法の指定

「値を置換」ダイアログ・ボックスで、フィールド値の匿名化にデフォルトの接頭辞を使用するか、またはユーザー指定の接頭辞を使用するか選択することができます。ダイアログ・ボックスの 「**OK**」 をクリックし、選択したフィールドに対し、「設定」タブの 「はい」 を選択して値の匿名化の設定を変更します。

フィールド値プレフィックス: 匿名化されたフィールド値に対するデフォルトのプレフィックスは **anon_** です。別の接頭辞が必要な場合は、「カスタム」を選択して独自の接頭辞を入力してください。

「値の変換」ダイアログ・ボックスは、連続型フィールドに対してのみ表示され、フィールド値の変換では乱数またはユーザー定義の値使用するのかを指定することができます。

無作為:「無作為」オプションを選択し、変換に乱数を使用します。デフォルトでは、「ランダム シードの設定」 が選択されます。「シード」 フィールドで値を指定するか、デフォルト値を使用します。

固定: 「固定」オプションを選択し、変換に乱数を使用します。

- スケール: フィールドが変換中に複製される数です。最小値は 1 で、最大値は通常 10 ですが、あふれを防止するために低くなる場合があります。
- 翻訳: 変換中、フィールド値に追加される数です。最小値は 0 で、最大値は通常 1000 ですが、あふれを防止するために低くなる場合があります。

フィールド値の匿名化

「設定」タブで匿名化のために選択されたフィールドには、匿名化された値が含まれます。

- 匿名化ノードを含むストリームを実行する場合
- 値をプレビューする場合

値をプレビューするには、「値を匿名化」タブの 「値を匿名化」 ボタンをクリックします。ドロップダウン・リストから色を選択します。

尺度が連続型の場合、次の項目が表示されます。

- 元の範囲の最小値および最大値
- 値の変換に使用された方程式

尺度が連続型以外のものである場合、画面にはそのフィールドの元の値および匿名化された値が表示されます。

黄色の背景色で表示された場合、前回値が匿名化されたため選択されたフィールドの設定が変更されたか、匿名化された値が正常でないなど、匿名化ノードのデータ上流に変更が行われたことを示します。値の現在の設定が表示された場合、再度 「値を匿名化」 ボタンをクリックし、現在の設定にしたがって値の新しい設定を生成します。

値を匿名化: 選択したフィールドに匿名化された値を作成し、テーブルに表示します。連続型フィールドにランダム シードを使用している場合、繰り返しこのボタンをクリックすると、クリックごとに異なる値のセットが作成されます。

値の消去: テーブルから元の値および匿名化された値を消去します。

データ分割ノード

データ分割ノードにより、既存の 1 つまたは複数の連続型 (数値範囲) フィールドの値に基づいて、自動的に新しい名義型フィールドを作成することができます。例えば、連続型収入フィールドを、平均からの同じ偏差による収入グループを含む、新しいカテゴリー・フィールドに変換することができます。または、2 つのフィールド間の当初のアソシエーションの強度を保存するために、カテゴリーの「スーパバイザ」フィールドを選択できます。

データ分割は、次を含む多くの理由で、有用です。

- アルゴリズムの要件。Naive Bayes やロジスティック回帰などの一定のアルゴリズムには、カテゴリー入力が必要です。
- パフォーマンス。 多項ロジスティックなどのアルゴリズムは、入力フィールドの異なる値の数が減らされると、より適正に実行されます。例えば、各ビンの当初の値ではなく、中央値または平均値を使用します。
- データのプライバシー。給与などの慎重な扱いが必要な個人情報は、プライバシーを保護するために、実際の数字でなく、一定の範囲内の数字として報告できます。

さまざまなデータ分割方法を使用することができます。新規フィールドのビンを作成すると、分割点に基づいてフィールド作成ノードを生成することができます。

データ分割ノードをいつ使用するか

データ分割ノードを使用する前に、その作業により適している他の技法がないかどうかを検討してください。

- あらかじめ定義された給与範囲など、カテゴリーの分割点を手作業で指定するには、フィールド作成ノードを使用します。詳しくは、トピック 162 ページの『フィールド作成ノード』を参照してください。
- 既存のセットの新しいカテゴリーを作成するには、データ分類ノードを使用します。 詳しくは、トピック 171 ページの『データ分類ノード』を参照してください。

欠損値の処理

データ分割ノードは欠損値を次のように処理します。

- ユーザー定義の空白。変換時に、空白として指定された欠損値が含まれます。例えば、データ型ノードを使用して空白値を示すために -99 を指定した場合、この値がデータ分割処理に含まれます。データ分割処理中に空白値を無視するには、置換ノードを使用して空白値をシステムのヌル値に置き換える必要があります。
- システム欠損値 **($null$)**。データ分割処理時にヌル値は無視され、変換後もヌル値のまま保持されます。

「設定」タブには、利用できる技術に関するオプションが用意されています。「表示」タブには、以前にこのノードに流されたデータに対して確立された分割点が表示されます。

データ分割ノードのオプション設定

データ分割ノードで次の手法を利用して、自動的にビン (カテゴリー) を生成することができます。

- 固定幅のデータ分割
- 分位 (等カウントまたは合計)
- 平均と標準偏差
- ランク

- カテゴリーの「スーパバイザ」フィールドに関連する最適化

このダイアログ・ボックスの下半分に表示されるオプションは、上で選択したデータ分割方法によって異なります。

ビン・フィールド:変換保留中の連続型 (数値範囲) フィールドがここに表示されます。データ分割ノードにより、複数のフィールドを同時にデータ分割することができます。右側のボタンを使用して、フィールドを追加または削除してください。

データ分割手段: 新規フィールドのビン (カテゴリー) の分割点を判断する方法を選択します。後続のトピックで、各ケースで使用できるオプションについて説明します。

ビンの閾値: ビンのしきい値をどのように計算するかを指定します。

- 常に再計算: 分割点およびビンの割り当てを、ノード実行時に常に再計算します。
- 可能な場合は「ビン値」タブから読み込む: 分割点およびビンの割り当てを必要に応じて計算します (新しいデータを追加する場合など)。

次のトピックでは、利用できるデータ分割方法とオプションについて説明していきます。

固定幅のデータ分割

データ分割手段として 「固定幅」 を選択した場合、ダイアログ・ボックスには新しい種類のオプション群が表示されます。

名前の拡張子: フィールドの生成に使用する拡張子を指定します。デフォルトの拡張子は「*_BIN*」になります。また、拡張子をフィールド名の先頭に追加するか (接頭辞)、または最後に追加するか (接尾辞) を指定することもできます。例えば、「*income_BIN*」と言う名前の新規フィールドを生成することができます。

ビン幅: ビンの「幅」を計算するための値 (整数または実数) を指定します。例えば、デフォルト値の 10 を使用して、フィールド 「年齢」 を分割することができます。「年齢」の範囲は 18 から 65 までであるため、生成されるビンは以下の表のようになります。

表 *24. 18* から *65* までの範囲の年齢のビン

ビン **1**	ビン **2**	ビン **3**	ビン **4**	ビン **5**	ビン **6**
>=13 to <23	>=23 to <33	>=33 to <43	>=43 to <53	>=53 to <63	>=63 to <73

ビンの開始点は、検出されたもっとも低い値から、ビン幅 (指定された) の半分を減算したものになります。例えば上記のビンでは、区間の開始点として 13 が使用されます。この値は、18 *[*最も低いデータ値*]* - 5 *[0.5 (*ビン幅 *10)]* = 13 という式によって計算された値です。

ビン数 : このオプションを使用して、新規フィールドの固定幅ビン (カテゴリー) 数を判断するために使用する整数を指定します。

ストリーム中でデータ分割ノードを実行した後に、データ分割ノードの 「プレビュー」 タブをクリックして、生成されたビンのしきい値を表示できます。 詳しくは、トピック 181 ページの『生成されたビンのプレビュー』を参照してください。

分位 (等カウントまたは合計)

データ分割手段は、検出されたレコードを 100 分位 (または 4 分位、10 分位他) グループへの分割に使用できる名義型フィールドを作成します。そのため、各グループは同じ番号のレコードを含むか、または各

グループの合計が等しくなります。レコードは、指定されたビン・フィールドの値に基づいて昇順でランク付けされます。したがって、選択されたビンの変数の一番低い値はランク 1 に割り当てられ、次のレコード・セットはランク 2 というように割り当てられます。各ビンのしきい値は、データと使用されている分位方法に基づいて自動的に生成されます。

分位名の拡張子。標準のパーセンタイルを使用して生成されるフィールドに対して使用する拡張子を指定します。デフォルトの拡張子は、*_TILE* に *N* を付けたものになります。*N* は分位数です。また、拡張子をフィールド名の先頭に追加するか (接頭辞)、または最後に追加するか (接尾辞) を指定することもできます。例えば、「*income_TILE4*」と言う名前の新規フィールドを生成することができます。

ユーザー設定の分位の拡張子。カスタム分位範囲に使用する拡張子を指定します。デフォルトは「*_TILEN*」 です。この場合、*N* がカスタムの数値で置換されない ことに注意してください。

利用できる分位を次に示します。

- **4** 分位。それぞれが 25% のケースを含む、4 のビンを生成します。
- **5** 分位。それぞれが 20% のケースを含む、5 のビンを生成します。
- **10** 分位。それぞれが 10% のケースを含む、10 のビンを生成します。
- **20** 分位。それぞれが 5% のケースを含む、20 のビンを生成します。
- パーセンタイル。それぞれが 1% のケースを含む、100 のビンを生成します。
- カスタム **N**。ビンの数を指定するために選択します。例えば、3 を指定すると、それぞれが 33.3 % のケースを含む 3 つのカテゴリー (2 つの分割点) が生成されます。

データ内の離散型の値が指定された分位数より少ない場合は、すべての分位が使用されません。このような場合、新しい分布は元のデータ分布を反映する可能性があります。

分位方法。レコードをビンに割り当てるときに使用する方法を指定します。

- レコード件数。各便に等しい数のレコードが割り当てられるよう調べます。
- 合計値。各ビンの中の値の合計が等しくなるようなビンにレコードを割り当てるよう調べます。例えば営業成績を対象とする場合、この方法を使用して見通しをレコードごとの値に基づき、最上位のビンで最も高い見通しの値により 10 分位グループに割り当てることができます。例えば、製薬会社は、医師を書いた処方箋の数に基づいて 10 分位にランク付けするかもしれません。各 10 分位にはほぼ同数の処方箋が含まれますが、これらの処方箋に貢献している医師の数は、10 分位に集中したほとんどの処方箋を書く医師により、同じではありません。

同順位。分割点の両側の値が同じ場合、結果は「タイ」状態になります。例えば、10 分位を割り当てており、10 % 以上のレコードがビンフィールドに対して同じ値を持っている場合、しきい値を一方または他方に合わせない限り、その値はすべて同じビンに適合しません。タイを次のビンに持ち上げるか現在のビンにとどめておくことができますが、一部のビンが想定値以上の値を持つことになった場合でも、同じ値を持つすべてのレコードが同じビンに分類されるようにします。次のビンのしきい値も、タイを解決する方法に従って同じ数値のセットに対して別のやり方で値を割り当てられるように、結果に合わせて調整することもできます。

- 隣のビンへ追加。タイ値を次のビンに移動するように選択します。
- 現在のまま保持。タイ値を現在の (低い) ビンに保持します。この方法は生成されるビンの数を結果として少なくします。
- 無作為割当。タイ値をビンに無作為に割り当てる場合に選択します。各ビンのレコード数が等しい数になるようにします。

例:レコード カウントによる分位

以下の表は、レコード・カウントによって各分位に分類する際に、単純化されたフィールド値がどのように 4 分位としてランク付けされるかを示しています。この結果は選択したタイオプションによって異なります。

表 *25.* レコード・カウントによる分位の例：

値	隣のビンへ追加	現在のまま保持
10	1	1
13	2	1
15	3	2
15	3	2
20	4	3

ビンあたりの項目数は、次のように算出されます。

```
total number of value / number of tiles
```

この例では、ビンあたりの望ましい項目数は 1.25 になります (5 つの値 / 4 分位)。値 13 (値番号 2) は望ましいカウントしきい値の 1.25 にまたがっているため、選択されているタイ・オプションに応じて扱いが異なります。「隣のビンへ追加」モードの場合、ビン 2 に追加されます。「現在のまま保持」モードの場合はビン 1 に残り、ビン 4 に割り当てられる値の範囲が、既存のデータ値の範囲外の値になります。その結果、3 つのビンだけが生成され、以下の表に示すように、各ビンのしきい値が調整されます。

表 *26.* データ分割例の結果：

ビン	下限	上限
1	>=10	<15
2	>=15	< 20
3	>=20	<=20

注：分位によるデータ分割スピードは、並行処理を可能にします。

ケースのランク付け

データ分割手段として 「ランク」 を選択した場合、ダイアログ・ボックスには新しい種類のオプション群が表示されます。

ランク付けでは、指定されたオプションに応じて、ランク、ランクの比率、および数値フィールドのパーセンタイル値を含む新規フィールドが作成されます。

ランク順序。「昇順」 (最小値を 1 とする) または「降順」 (最大値を 1 とする) を選択します。

順位。上で指定したように、ケースを昇順または降順でランク付けする場合に選択します。新規フィールドの値の範囲は 1 から *N* までです。*N* は、元のフィールドでの個別値の数です。同順位の値には、そのランクの平均が与えられます。

ランクの比率。新規フィールドの値が、ランクを非欠損ケースの重みの合計で除算した値になるように、ケースをランク付けする場合に選択します。小数点付き順位の範囲は 0 から 1 までです。

パーセンテージの小数点付き順位。各ランクが、有効な値を持つレコード数で除算された後、100 倍されます。パーセンテージの小数点付き順位の範囲は 1 から 100 までです。

拡張。すべてのランクオプションで、カスタム拡張子を作成し、拡張子をフィールド名の先頭に追加するか (接頭辞)、または最後に追加するか (接尾辞) を指定することができます。例えば、「*income_P_RANK*」と言う名前の新規フィールドを生成することができます。

平均/標準偏差

データ分割手段として 「平均/標準偏差」 を選択した場合、ダイアログ・ボックスには新しい種類のオプション群が表示されます。

この方法では、指定したフィールドの分布の平均および標準偏差の値に基づいて、バンド カテゴリーを持つ 1 つまたは複数の新規フィールドが生成されます。以下から、使用する偏差数を選択してください。

名前の拡張子 **:** フィールドの生成に使用する拡張子を指定します。デフォルトの拡張子は「*_SDBIN*」になります。また、拡張子をフィールド名の先頭に追加するか (接頭辞)、または最後に追加するか (接尾辞) を指定することもできます。例えば、「*income_SDBIN*」と言う名前の新規フィールドを生成することができます。

- **+/- 1** 標準偏差**:** これを選択すると、3 つのビンが生成されます。
- **+/- 2** 標準偏差**:** これを選択すると、5 つのビンが生成されます。
- **+/- 3** 標準偏差**:** これを選択すると、7 つのビンが生成されます。

例えば「+/-1 標準偏差」を選択すると、以下の表に示す計算によって 3 つのビンが生成されます。

表 *27.* 標準偏差のビンの例 :

ビン **1**	ビン **2**	ビン **3**
x < (Mean - Std. Dev)	(Mean - Std. Dev) <= x <= (Mean + Std. Dev)	x > (Mean + Std. Dev)

正規分布では、68% のケースが平均の 1 標準偏差に、95% が 2 標準偏差に、そして 99% が 3 標準偏差にあてはまります。ただし、標準偏差に基づいてバンド カテゴリーを作成すると、一部のビンが実際のデータ範囲外に定義されたり、取り得るデータ値の範囲外 (例えば、負の給与範囲など) になる可能性があることに注意してください。

最適カテゴリー化

データ分割対象のフィールドが別のカテゴリー・フィールドと強力に関連付けられている場合は、2 つのフィールド間の当初のアソシエーションの強度を保つ方法でビンを作成するために、カテゴリー・フィールドを「スーパバイザ」フィールドとして選択することができます。

例えば、クラスター分析を使用して、住宅ローンの不履行率に基づき、最初のクラスターの最も高い割合で状態を分類するとします。この場合、「スーパバイザ」フィールドとしてのモデルによって生成されたビン・フィールドや所属クラスター・フィールドとして「期日経過率」および「請戻権喪失率」を選択します。

名前の拡張子: 生成されるフィールドで使用する拡張子と、その拡張子をフィールド名の先頭 (接頭辞) または最後 (接尾辞) に付加するかどうかを指定します。例えば、「pastdue_OPTIMAL」および「inforeclosure_OPTIMAL」と言う名前の新規フィールドを生成することができます。

スーパバイザ フィールド: ビンの作成に使用されるカテゴリー フィールド。

大規模データセットとのパフォーマンスを改善するプレビン フィールド: 最適なデータ分割を効率化するために前処理を実行するかどうかを指定します。単純な、監視されないデータ分割方法を使用して大量のビンへ値が振り分けられ、各ビン内で値はその平均により表現されて、監視されるデータ分割に進む前に、重みに応じてケースが調整されます。実際的な問題として、この方法は、正確さを犠牲にしても速度が速いほうを採用し、大規模データ・セット向けに推奨されます。このオプションを使用する場合、事前処理の実行後に変数が終了するビンの最大数を指定することもできます。

比較的小さなケース度数のビンを大きな近傍ビンとマージ: しきい値が大きいほど結合が多くなります。

切り取りポイントの設定値

「分割点の設定」ダイアログ・ボックスで、最適データ分割アルゴリズムの詳細設定を指定することができます。これらの設定は、アルゴリズムに対象フィールドを使用してビンを計算する方法を指示します。

ビンの終点**:** 下位または上位の終点が包括的 (下位 <= *x*) か、排他的 (下位< *x*) かを指定できます。

最初のビンと最後のビン**:** 最初と最後のビン両方に対し、 ビンに境界がないか (正または負の方向に無限に拡張)、最低または最高のデータ・ポイントによる境界があるかを指定できます。

生成されたビンのプレビュー

データ分割ノードの「ビンの値」タブで、生成されたビンのしきい値を表示できます。「生成」メニューを使用して、あるデータ・セットから別のデータ・セットへこのしきい値を適用するのに使用できる、フィールド作成ノードも生成できます。

分割フィールド **:** ドロップダウン・リストから、表示するフィールドを選択します。表示されるフィールド名は、元のフィールド名を使用しています。

分位 **:** ドロップダウン・リストを使用して、10 や 100 などの、表示する分位数を選択します。このオプションは、分位方法 (等カウントまたは等合計) でビンが生成された場合にだけ利用できます。

ビンの閾値 **:** そのビンに分類されたレコード数も表示されます。最適化されたデータ分割方法の場合のみ、各ビンのレコード数が、全体の割合として表示されます。しきい値は、ランク付けデータ分割方法を使用中の場合に適用できません。

値の読み込み **:** データ・セットからビンに分けられた値を読み込みます。ストリームに新しいデータが流されると、しきい値は上書きされます。

フィールド生成ノードの生成

「生成」メニューを使用して、現在のしきい値に基づいたフィールド作成ノードを作成できます。確立したビンのしきい値を、あるデータ・セットから別のデータ・セットに適用する場合に、このオプションが役立ちます。また、大きいデータ・セットに対して作業を行う場合、いったんこれらの分割点が明らかになれば、データ分割操作よりもフィールド作成操作の方がより効率的に、速く実行できます。

RFM 分析ノード

リーセンシ、フリクエンシ、マネタリー (RFM) の分析ノードを使用すると、最後に購入したのがどのくらい最近か (リーセンシ)、どのくらい頻繁に購入するか (フリクエンシ)、トランザクション全体でいくら消費したか (マネタリー) を検証することによって、最も良い顧客となると考えられるのはどの顧客かを量的に決定することができます。

RFM 分析の推論は、製品またはサービスを購入する顧客がサイド購入する可能性が高いということです。カテゴリー化された顧客データは、多くのビンに分割され、分割基準は必要に応じて調整されます。それぞれのビンで、顧客はスコアに割り当てられます。これらのスコアは結合され、全体の RFM スコアを提供します。このスコアは、それぞれの RFM パラメータに作成されたビンの顧客の所属を表します。この分割されたデータは、例えば最も頻繁に取引し、支出の高い顧客を識別することによってニーズを満たすことができます。また、詳細なモデル作成および分析のためにストリーム内に渡される場合があります。

ただし、RFM スコアを分析しタンク付けする機能は役に立つツールですが、使用する場合は特定の要素に注意する必要があります。高くランク付けされた対象の顧客を勧誘する場合がありますが、これらの顧客の過剰な勧誘は不快感を呼び、取引の繰り返しが実際は失敗してしまう恐れがありますので注意してください。また、低いスコアの顧客は無視することはせず、より良い顧客を開拓することができることを記憶しておく価値があります。それに対し、市場によっては高いスコアだけが必ずしも良好な販売の可能性を反映するわけではありません。例えば、リーセンシを表すビン 5 の最近購入した顧客は、車やテレビなど効果で長持ちする商品を販売する者にとっては、対象となる顧客ではありません。

注 : データの保存方法によっては、RFM 分析ノードを RFM レコード集計ノードに先行してデータを使用可能な形式に変換する必要があります。例えば、入力データは顧客ごとに 1 行の顧客の形式である必要がありますが、顧客のデータがトランザクション・フォームである場合、上流で RFM レコード集計ノードを使用してリーセンシ、フリクエンシ、マネタリーのフィールドを作成する必要があります。 詳しくは、トピック 87 ページの『RFM レコード集計ノード』を参照してください。

IBM SPSS Modeler の RFM レコード集計ノードおよび RFM 分析ノードを設定して独立した分割を使用します。最新性、頻度、金額値の各尺度のデータを、これらの値および尺度に関係なくランク付けし、分割します。

RFM 分析ノードの設定

リーセンシー: フィールド・ピッカー (テキスト・ボックスの右側にあるボタン) を使用して、リーセンシのフィールドを選択します。このフィールドは日付、タイムスタンプまたは単純な数値です。日付またはタイム・スタンプが最も新しいトランザクションの日付を示す場合、最も高い値が最新のものであると見なします。数値が指定されている場合、数値は最新のトランザクションから経過した時間を表し、最も低い値が最新であると見なします。

注 : RFM レコード集計ノードが RFM 分析ノードに先行する場合、RFM レコード集計ノードに生成されたリーセンシ、フリクエンシ、マネタリーのフィールドが RFM 分析ノードの入力として選択されます。

度数: フィールド・ピッカーを使用して、使用するフリクエンシのフィールドを選択します。

マネタリー: フィールド・ピッカーを使用して、使用するマネタリーのフィールドを選択します。

ビン数: それぞれの 3 つの出力タイプに対し、作成するビン数を指定します。デフォルトは 5 です。

注 : ビン数の最小値は 2 で、最大値は 9 です。

重み: デフォルトでは、スコア計算時最も高い重要度がリーセンシのデータに与えられ、次にフリクエンシ、マネタリーの順に与えられます。必要に応じて、これらのフィールドに影響する重みを修正して、高い重要度を与えるフィールドを変更します。

RFM スコアは次のように計算されます。(リーセンシ スコア x リーセンシの重み) + (フリクエンシ スコア x フリクエンシの重み) + (マネタリー・スコア x マネタリーの重み)

同順位。同じ (タイ) のスコアがどのように分割されるかを指定します。以下のオプションがあります。

- 隣のビンへ追加: タイ値を次のビンに移動するように選択します。
- 現在のまま保持: タイ値を現在の (低い) ビンに保持します。この方法は生成されるビンの数を結果として少なくします。(デフォルトの設定です。)

ビンの閾値: ノードが実行された場合に RFM スコアおよびビンの割り当てが常に再計算されるかどうか、必要な場合にのみ計算されるか (例えば、新しいデータが追加されて場合) を指定します。「可能ならビンの値から読み込む」 を選択すると、「ビンの値」タブでさまざまなビンの上限および下限の分割点を編集することができます。

実行時、RFM 分析ノードは処理されていないリーセンシ、フリクエンシ、マネタリーのフィールドを分割し、次の新しいフィールドをデータ・セットに追加します。

- リーセンシ スコア。リーセンシのランク (ビン値)
- フリクエンシ スコア。フリクエンシのランク (ビン値)
- マネタリー・スコア。マネタリーのランク (ビン値)
- RFM スコア。リーセンシ、フリクエンシ、マネタリー・スコアの重みの合計

最後のビンに外れ値を追加: このチェック・ボックスを選択した場合、低い側のビンより下にあるレコードは低い側のビンに追加され、最も高いビンの上にあるレコードは最も高いビンに追加されます。このチェック・ボックスを選択しなかった場合は、ヌル値が割り当てられます。このチェック・ボックスは、「可能ならビンの値から読み込む」 が選択されている場合にのみ使用できます。

RFM 分析ノードの分割

「ビンの値」タブを使用すると、生成されたビンのしきい値を表示でき、またある場合は修正することもあります。

注 : このタブでは、「設定」タブの 「可能ならビンの値から読み込む」が選択されている場合にのみ、値を修正できます。

分割フィールド: ドロップダウン・リストから、ビンに分割するフィールドを選択します。「設定」タブで選択された値を使用できます。

ビンの値のテーブル: 生成された各ビンのしきい値がここに表示されます。「設定」タブで 「可能ならビンの値から読み込む」 を選択すると、関連するセルをダブルクリックして、各ビンの上限および下限の分割点を修正することができます。

値の読み込み: データ・セットから分割された値を読み込み、ビンの値のテーブルを作成します。「設定」タブで 「常に再計算」 を選択した場合は、新しいデータがストリームで実行されると、ビンのしきい値が上書きされます。

アンサンブル・ノード

アンサンブル・ノードでは、2 つまたはそれ以上のモデル・ナゲットを組み合わせて個々のモデルのいずれかから取得するよりも、より正確な予測を取得します。複数モデルの予測を組み合わせることにより、個々のモデルの制限を回避でき、全体の精度がより高くなります。こうして組み合わされたモデルは通常、少なくとも最良のモデルと同じくらい、あるいはしばしばそれ以上のパフォーマンスを実現します。

ノードのこの組み合わせは、自動分類、自動数値および自動クラスターの自動モデル作成ノードで自動的に作成されます。

アンサンブル・ノードを使用した後、分析ノードまたは評価ノードを使用して、各入力モデルと組み合わせた結果の制度を比較することができます。これを実行するには、「アンサンブル モデルにより生成されたフィールドを除外」 オプションがアンサンブル・ノードの「設定」タブで選択されていないことを確認します。

出力フィールド

各アンサンブル・ノードは、結合したスコアを含むフィールドを生成します。名前は、特定の対象フィールドに基づき、フィールドの尺度 (フラグ型、名義 (セット) 型、または連続型 (範囲) 型) によって *$XF_*、*$XS_*、または *$XR_* の接頭辞が付きます。例えば、対象フィールドがフラグ型で *response* という名前の場合、出力フィールド名は *$XF_response* となります。

信頼度または傾向フィールド**:** フラグ型フィールドと名義型フィールドの場合は、追加の信頼度または傾向フィールドがアンサンブル法に基づいて作成されます。これを以下の表に示します。

表 *28.* アンサンブル法でのフィールド作成：

アンサンブル法	フィールド名
票決 確信度-重み付き票決 未調整傾向-重み付き票決 調整済み傾向-重み付き票決 最高確信度勝ち取り	*$XFC_<field>*
平均未調整傾向	*$XFRP_<field>*
平均調整済み傾向	*$XFAP_<field>*

アンサンブル・ノードの設定

集合体の対象フィールド**:** 2 つまたはそれ以上の上流モデルで対象フィールドとして使用される単一フィールドを選択します。上流モデルはフラグ型、名義型または連続型対象を使用することができますが、少なくとも 2 つのモデルが同じ対象を共有してスコアを結合する必要があります。

アンサンブル モデルにより生成されたフィールドを除外**:** 出力から、アンサンブル・ノードに使用する個々のモデルで生成されたすべての追加フィールドを削除します。すべての入力モデルの結合スコアにのみ関心がある場合、このチェック・ボックスを選択します。例えば分析ノードまたは評価ノードを使用して結合スコアのと各入力モデルの制度を比較する場合、このオプションが選択解除されていることを確認します。

使用可能な設定は、対象として選択されたフィールドの尺度によって異なります。

連続型対象

連続型対象の場合、スコアは平均が算出されます。これはスコアの結合にのみ使用できる方法です。

スコアまたは推定を平均化する場合、アンサンブル・ノードでは標準誤差の計算を使用して、測定されたまたは推定された値と真の値との間の差異を算出し、これらの推定がどれくらい近いかを示します。デフォルトでは、新しいモデルに標準誤差の計算が生成されます。ただし、再生成する場合など、既存のモデルのチェック・ボックスの選択を解除することができます。

カテゴリー対象

カテゴリー対象の場合、それぞれの予測値が選択される回数を集計し、最も高い合計数を持つ値を選択することによって動作する票決など、多くの方法がサポートされています。例えば、5 つのモデルのうち 3 つ

ではい と予測され、残り 2 つでいいえ と予測される場合、はい が 3 対 2 の票決で勝ちます。代わりに、各予測の信頼度または傾向値に基づいて、票決に重みを付けることができます。また、各予測の確信度または傾向値に基づいて、票決に重み付けすることができます。重みは集計され、最も大きな合計の値が再度選択されます。最後の予測の確信度は、勝った値の重みの合計をアンサンブルに含まれるモデルの数で割った値です。

すべてのカテゴリー・フィールド**:** フラグ型フィールドおよび名義型フィールドの場合、次の方法がサポートされます。

- 票決
- 確信度-重み付き票決
- 最高確信度勝ち取り

フラグ型フィールドのみ**:** フラグ型フィールド場合のみ、傾向に基づいた次の方法も使用できます。

- 未調整傾向重み付き票決
- 調整済み傾向重み付き票決
- 平均未調整傾向
- 平均調整済み傾向

可否同数**:** 票決方法の場合、可否同数の解決方法を指定することができます。

- 無作為選択**:** 可否同数の値の 1 つが無作為に選択されます。
- 最高確信度**:** 最高確信度で予測された可否同数の値が勝ちます。これは、予測されたすべての値の最高確信度と必ずしも同じとは限りません。
- 未調整または調整済み傾向 **(**フラグ型フィールドのみ**):** 絶対傾向が次のように計算されている場合の、最大絶対傾向によって予測された可否同数の値。

```
abs(0.5 - propensity) *
2
```

または、調整済み傾向の場合は次のようになります。

```
abs(0.5 - adjusted propensity) * 2
```

データ区分ノード

データ区分ノードは、モデル構築の学習、テスト、および検証の各ステージ用に、データを独立したサブセット (サンプル) に分割するフィールドを生成するために使用されます。1 組のサンプルをモデルの生成に使用し、別の組のサンプルで生成したモデルをテストすることにより、そのモデルが、このデータに似た性質を持つより大きなデータ・セットにどの程度適用できるかについての良い目安を得ることができます。

データ区分ノードは、役割が 「データ区分」 に設定された名義型フィールドを生成します。適当なフィールドが既に存在している場合、データ型ノードを使用すると、そのフィールドをデータ区分として指定できます。この場合、新しいデータ区分ノードは必要ありません。2 つまたは 3 つの値を持つインスタンス化された設定フィールドをデータ区分として使用できますが、名義型フィールドは使用できません。 詳しくは、トピック 156 ページの『フィールドの役割の設定』を参照してください。

単一のストリーム内で複数のデータ区分フィールドを定義できますが、その場合、データ分割を使用する各モデリング・ノードごとに「フィールド」タブでデータ区分フィールドを 1 つだけ選択しなければなりません。(1 つのデータ区分だけが存在している場合、データ分割を有効にすると、そのデータ区分が必ず自動的に使用されます)。

データ分割を有効にする**:** 分析でデータ区分を使用するには、適当なモデル構築または分析ノードの「モデルのオプション」タブでデータ分割を有効にする必要があります。このオプションの選択を解除すると、フィールドを削除しないでデータ分割を無効にできます。

データ範囲や場所などの他のいくつかの基準に基づいてデータ区分フィールドを生成する場合、フィールド作成ノードを使用できます。 詳しくは、トピック 162 ページの『フィールド作成ノード』を参照してください。

例: RFM ストリームを構築して、以前のマーケティング・キャンペーンに肯定的に応答した最近の顧客を識別する場合、販売会社のマーケティング部門ではデータ区分ノードを使用して、データを学習用データ区分および検定用データ区分に分割します。

データ区分ノードのオプション

データ区分フィールド**:** そのノードにより生成されるフィールドの名前を指定します。

データ区分**:** データを、2 組 (学習とテスト) または 3 組 (学習、テスト、および検定) のサンプルに分割できます。

- 学習とテスト**:** データを 2 つのサンプルに区分し、一方のサンプルを使用して学習し、もう一方のサンプルを使用してテストできるようになります。
- 学習、テスト、および検定**:** データを 3 つのサンプルに区分し、1 つのサンプルを使用してモデルを学習し、2 つ目のサンプルを使用してモデルのテストと調整を行い、3 つめのサンプルを使用して結果を検証できるようになります。3 組にすると、結果として各データ区分のサイズが小さくなりますが、作業するデータ・セットが非常に大きければ、ほとんどの場合に最適です。

データ区分のサイズ。各データ区分の相対的なサイズを指定します。各データ区分のサイズの合計が 100% より小さい場合、データ区分に含まれないレコードは、破棄されます。例えば、ユーザーが 1000 万個のレコードを持っており、学習データ区分のサイズを 5%、テストを 10% に指定した場合、そのノードの実行後、およそ 500,000 個が学習レコードに、 100 万個がテスト・レコードに割り当てられ、残りは破棄されます。

値**:** データ中の各データ区分サンプルを表すために使用される値を指定します。

- システム定義の値 **(**「**1**」、「**2**」および「**3**」**)** を使用**:** 各データ区分を表すのに整数値を使用します。例えば、全てのレコードが、学習データ区分に含まれる全てのサンプルは、データ区分フィールドの値が 1 になります。これにより、ロケール間でのデータのポータビリティが保証され、データ区分フィールドが別の場所にインストールされた場合にも、ソートの順序が維持されます 1 は学習区分を表します) 。ただし、値の解釈にやや手間がかかります。
- ラベルをシステム定義の値の後に結合する**:** ラベルを整数値に結合します。例えば、学習データ区分レコードの値は `1_Training` になります。こうすることにより、人間がデータを見たとき、それぞれの値が何を表しているかが解りやすくなります。しかも、ソートの順序も維持されたままです。ただし、値は、特定のロケールに固有になります。
- ラベルを値として使用**:** 整数値を持たないラベルを使用します。例えば、学習 です。この場合、ラベルを編集して値を指定できるようになります。ただし、データはロケール固有になり、データ区分列の再インストールすると、値はインストール先の言語のソート順序で処理されるため、本来の「意味上の」順序とは一致しなくなる場合があります。

シード。 「ランダム シードの設定」を選択した場合のみ、使用することができます。無作為なパーセンテージに基づいてレコードをサンプリングまたはデータ区分している場合、このオプションで、別のセッションに同じ結果を複製できるようになります。乱数ジェネレータに使用される開始値を指定することで、ノー

ドが実行されるごとに毎回同じレコードが割り当てられることが保証されます。自動的に無作為な値を生成するには、希望のシード値を入力するか、「生成」 ボタンを入力します。このオプションが選択されないと、ノードが実行されるごとに異なるサンプルが生成されます。

注: データベースから読み込まれたレコードで「シード」オプションを使用する場合は、ノードを実行するたびに同じ結果になるように、サンプリングの前にソート ノードが必要になることがあります。この理由は、ランダム シードがレコードの順序に依存しているためです。各レコードがリレーショナル・データベース内で同じ位置に留まる保証はありません。詳しくは、トピック 88 ページの『ソート・ノード』を参照してください。

一意のフィールドを使用してデータ区分を割り当てる: 「ランダム シードの設定」を選択した場合のみ、使用することができます。(ティア 1 のデータベースのみ) SQL プッシュバックを使用して、レコードをデータ区分に割り当てるには、このチェック ボックスにチェックマークを付けます。ドロップダウン・リストから、一意の値を持つフィールド (ID フィールドなど) を選択肢、レコードが無作為にかつ繰り返し割り当てられるようにします。

データベースの階層については、データベース・ソース・ノードの説明に記載されています。 詳しくは、トピック 19 ページの『データベース・ソース・ノード』を参照してください。

条件抽出ノードの生成

「データ区分」ノードの「生成」メニューを使用すると、各データ区分ごとに「条件抽出ノード」を自動的に生成できます。例えば、すべてのレコードを学習データ区分に選択して、このデータ区分のみを使用して、さらに評価または分析を続けることができます。

フラグ設定ノード

フラグ設定ノードは、1 つ以上の名義型フィールド用に定義されたカテゴリー値を基にして、フラグ型フィールドを派生させるために使用します。例えば、データ・セットには、高、正常、低 という値を持つ名義型フィールド *BP* (血圧) を含まれる場合があります。データの操作を簡単にするために、高血圧用のフラグ型フィールドを作成し、そこで患者が高血圧であるかどうかを示すことができます。

フラグ設定ノードのオプションの設定

セット型フィールド: 測定の尺度が名義型 (セット型) であるすべてのデータ・フィールドがリストされます。リストから 1 つのフィールドを選択して、セット内の値を表示します。これらの値の中から選択して、フラグ・フィールドを作成することができます。利用可能な名義型フィールドとその値を表示するには、上流のソースまたはデータ型ノードを使用して、データを完全にインスタンス化する必要があります。詳しくは、トピック 144 ページの『データ型ノード』を参照してください。

フィールド名拡張子: 新しく作成するフラグ・フィールドの接頭辞または接尾辞として追加する拡張子を指定する場合に選択します。デフォルトでは、フィールド名_フィールド値のように、元のフィールド名にフィールドの値を組み合わせた新規フィールド名が自動的に作成されます。

利用できるセット値: 上のフィールドで選択したセット内の値が表示されます。フラグを生成する対象になる 1 つ以上の値を選択します。例えば、フィールド *blood_pressure* の値が高、中、および低の場合、高を選択して右のリストに追加することができます。この場合、高血圧を示す値があるレコードに対して、フラグ・フィールドが作成されます。

フラグ型フィールドを作成: 新しく作成されたフラグ型フィールドのリストが表示されます。フィールド名拡張子コントロールを使用して、新しいフィールドの命名に関するオプションを指定することができます。

真 **(true)** の値**:** フラグを設定するときにノードが使用する真 (true) の値を指定します。デフォルトの値は、**T** です。

False 値**:** フラグを設定するときにノードが使用する偽 (false) の値を指定します。デフォルトの値は、**F** です。

集計キー**:** 下のフィールドで指定するキー・フィールドを基にしてレコードをグループ化する場合に選択します。「集計キー」を選択した場合、真 (true) に設定されたレコードが *1* 件でも存在すると、グループ内のすべてのフラグ型フィールドが「オン」になります。フィールド・ピッカーを使用して、レコードを集計するために使用するキー・フィールドを指定してください。

再構成ノード

再構成ノードを使用し、名義型フィールドまたはフラグ型フィールドの値に基づいて複数のフィールドを生成することができます。新規に生成されたフィールドは、他のフィールドからの値、または数値フラグ (0 または 1) を含むことができます。このノードの機能は、フラグ設定ノードの機能と似ています。ただし、より柔軟なノードです。このノードを使用すると、他のフィールドからの値を使用して (数値フラグを含む) どのようなタイプのフィードでも生成できるようになります。したがって、レコード集計や下流の他のノードを使用する操作が可能になります。(フラグ設定ノードにより、ワン・ステップでフィールドのレコード集計が可能となり、これはフラグ型フィールドを生成する場合に便利です。)

例えば、次のようなデータ・セットは、預金と手形という値を持つ名義型フィールド アカウントを含んでいます。開設時の残高と現在の残高は各アカウントに記録され、顧客は各タイプの複数のアカウントを持っています。特定のアカウント・タイプを持つ顧客がいるか、いた場合、各アカウント・タイプの残高はいくらか、ということを知りたくなります。再構成ノードを使用して、アカウントの各値に対してフィールドを生成し、値として *Current_Balance* (現在の残高) を選択します。新しい各フィールドには、あるレコードの現在の残高が書き込まれます。

表 *29.* 再構成前のサンプル・データ：

CustID	Account	Open_Bal	Current_Bal
12701	ドラフト	1000	1005.32
12702	貯金	100	144.51
12703	貯金	300	321.20
12703	貯金	150	204.51
12703	ドラフト	1200	586.32

表 *30.* 再構成後のサンプル・データ：

CustID	Account	Open_Bal	Current_Bal	Account_Draft_ Current_Bal	Account_Savings _Current_Bal
12701	ドラフト	1000	1005.32	1005.32	$null$
12702	貯金	100	144.51	$null$	144.51
12703	貯金	300	321.20	$null$	321.20
12703	貯金	150	204.51	$null$	204.51
12703	ドラフト	1200	586.32	586.32	$null$

レコード集計ノードと共に再構成ノードを使用

多くのケースで、再構成ノードとレコード集計ノードを一対として使用したい場合があります。先の例では、ある顧客 (ID 12703) が 3 つのアカウントを持っていました。レコード集計ノードを使用して各アカウントタイプの全体残高を計算します。キーとなるフィールドは *CustID* であり、レコード集計フィールドは新規に再構成されたフィールド、*Account_Draft_Current_Bal* と *Account_Savings_Current_Bal* です。結果を次の表に示します。

表 *31.* 再構築とレコード集計後のサンプルデータ：

CustID	Record_Count	Account_Draft_Current_ Bal_Sum	Account_Savings_Current_ Bal_Sum
12701	1	1005.32	$null$
12702	1	$null$	144.51
12703	3	586.32	525.71

再構成ノードのオプション設定

利用可能なフィールド**:** 測定の尺度が名義型 (セット型) またはフラグ型 であるすべてのデータ・フィールドがリストされます。リストから 1 つのフィールドを選択してセット型 (またはフラグ型) に値を表示します。次に、この値の中から再構成フィールドを生成するための値を選択します。使用可能なフィールドとその値を表示するには、上流のソースまたはデータ型ノードを使用して、データを完全にインスタンス化する必要があります。 詳しくは、トピック 144 ページの『データ型ノード』を参照してください。

使用可能な値**:** 上のフィールドで選択したセット内の値が表示されます。再構成フィールドを生成する対象になる 1 つ以上の値を選択します。例えば、フィールド 血圧 の値が高、中、および低の場合、高を選択して右のリストに追加することができます。これで、高 の値を持つレコードに対して指定した値 (以下参照) を持つフィールドが生成されます。

再構成フィールドの作成**:** 新しく作成された再構成フィールドのリストが表示されます。デフォルトでは、フィールド名_フィールド値のように、元のフィールド名にフィールドの値を組み合わせた新規フィールド名が自動的に作成されます。

フィールド名を含める**:** 新しいフィールド名から接頭辞としての元のフィールド名を削除するために選択解除します。

他のフィールドから値を使用**:** 再構築されたフィールドに書き込まれる値を持つ 1 つ以上のフィールドを指定します。フィールド・ピッカー・ボタンを使用して 1 つ以上のフィールドを選択します。選択された各フィールドに対して、新しいフィールドが 1 つ生成されます。値フィールドの名前が、再構成されたフィールド名に追加されます (*BP_High_Age* や *BP_Low_Age* など)。各新規のフィールドは、元の値フィールドのデータ型を引き継いでいます。

値フラグの作成**:** 選択すると、他のフィールドからの値を使用せずに、新規のフィールドに数値フラグ (0=偽、1=真) を書き込みます。

行列入替ノード

デフォルトでは、列とフィールドおよび行は、レコードと観測値が書き込まれます。必要に応じて、行列入替ノードを使用して、行にあるデータと列にあるデータを入れ替えて、フィールドをレコードに、レコードをフィールドにすることができます。例えば、各時系列データが列ではなく行に書き込まれている時系列データを持っている場合、分析前にそのデータを入れ替えることができます。

行列入替ノードのオプションの設定

「行列入替方法」ドロップダウンで、行列入替ノードに実行させる方法を「フィールドとレコードの両方」、「レコードからフィールドへ」、または「フィールドからレコードへ」から選択します。これら 3 つの方法の各設定について、以下のセクションに記述しています。

制約事項: 「レコードからフィールドへ」方式および「フィールドからレコードへ」方式は、Windows 64 ビット、Linux 64 ビット、および Mac でのみサポートされています。

フィールドとレコードの両方

新規フィールド名は、指定された接頭辞に基づいて自動的に生成することも、データ内の既存のフィールドから読み込むこともできます。

接頭辞を使用: このオプションは、新しいフィールド名を指定した接頭辞 (Field1、Field2、など) に基づいて自動的に生成します。必要に応じて接頭辞をカスタマイズできます。このオプションを使用する場合は、元のデータの行数に無関係に、生成するフィールド数を指定する必要があります。例えば、「新規フィールド数」 が 100 に設定されると、最初の 100 行を超えるすべてのデータは破棄されます。元のデータが 100 行に満たない場合は、一部のフィールドはヌルとなります。(必要に応じてフィールド数を増やすことができますが、この設定の目的は、100 万のレコードを 100 万のフィールドに行列入替することを防ぐことにあり、もしそのような入れ替えを行うと管理不能となります。)

例えば、行内に系列データがあり、各月の独立したフィールド (列) があると想定します。各系列が個別のフィールドに、各月が行になるように、行列の入れ替えを実行できます。

フィールドから読み込み: 既存のフィールドからフィールド名を読み込みます。このオプションを使用すると、新規フィールド数が指定された最大数を上限として、データによって決定します。選択されたフィールドの各値は、出力データの新しいフィールドとなります。選択されたフィールドには (整数、文字列、日付など) 任意のストレージ・タイプがありますが、フィールド名の重複を避けるために、選択されたフィールドの各値は一意である必要があります (つまり、値の数は行数と一致する必要があります)。フィールド名が重複している場合、警告が表示されます。

- 値の読み込み: 選択されたフィールドがインスタンス化されていない場合、このオプションを選択して新しいフィールド名のリストを編成します。フィールドがすでにインスタンス化されている場合、この手順は必要ではありません。
- 読み込む値の最大数: データからフィールド名を読み込む場合、上限値を指定してあまりに多いフィールドが作成されることを回避します。(前述のとおり、100 万のレコードを 100 万のフィールドに入れ替えると、管理できない結果が生じます。)

例えば、データの最初の列が各シリーズの名前を指定している場合、これらの値を入れ替えられたデータのフィールド名として使用することができます。

入れ替え:デフォルトでは、連続型 (数値範囲) フィールドのみが入れ替わります (整数または実数)。オプションで、数値型フィールドのサブセットを選択するか、代わりに文字列フィールドを入れ替えることができます。ただし、入れ替えるフィールドはすべて同じストレージ・タイプでなければなりません (数値または文字列のいずれか一方。両方ではありません)。これは、入力フィールドを混在させると、各出力列内に混在した値が生成され、フィールドの値がすべて同じストレージでなければならないという規則に反するためです。その他のストレージ・タイプ (日付、時間、タイムスタンプ) を入れ替えることはできません。

- すべての数値型: すべての数値型フィールド (整数または実数ストレージ) を入れ替えます。出力の行数は、下のデータの数値型フィールド数に一致します。
- すべての文字列: すべての文字列フィールドを入れ替えます。

- ユーザー設定**:** 数値型フィールドのサブセットを選択することができます。出力の行数は、選択したフィールド数に一致します。このオプションは数値型フィールドに対してのみ使用できます。

行 **ID** 名**:** ノードで作成された行 ID フィールドの名前を指定します。このフィールドの値は、元のデータのフィールド名によって決まります。

ヒント**:** 行から列へ時系列データを入れ替える際、元のデータに各測定値の期間にラベルを付ける日付、月、または年などの行が含まれる場合は、データの最初の行のラベルを含めるのではなく、これらのラベルをフィールド名として IBM SPSS Modeler に読み込みます (前述の例で説明したとおり、元のデータの月または日付をフィールド名として表示します)。これにより、各列のラベルと値の混在を回避します (ストレージ・タイプが列内で混在しないため、数値を文字列として読み込むことを強制します)。

レコードからフィールドへ

フィールド**:** 「フィールド」リストには、行列入替ノードに入力するすべてのフィールドが含まれます。

インデックス**:** 「インデックス」セクションを使用してインデックス・フィールドとして使用するフィールドを選択します。

フィールド**:** 「フィールド」セクションを使用してフィールドとして使用するフィールドを選択します。

値**:** 「値」セクションを使用して値フィールドとして使用するフィールドを選択します。

集約関数**:** インデックスに対して複数のレコードがある場合、レコードを 1 件に集約する必要があります。「集約関数」ドロップダウンを使用して、以下の関数のいずれかを利用するレコードの集約方法を指定します。集約はすべてのフィールドに影響を及ぼすことに注意してください。

- 平均値**:** キー・フィールドの各組み合わせの平均値を返します。平均値は、中心傾向の尺度であり、算術平均です (ケース数で割った合計)。
- 合計**:**キー・フィールドの各組み合わせの合計値を返します。合計は、欠損値のないすべてのケースに対する変数の値の合計です。
- 最小値**:** キー・フィールドの各組み合わせの最小値を返します。
- 最大値**:** キー・フィールドの各組み合わせの最大値を返します。
- 中央値**:** キー・フィールドの各組み合わせの中央値を返します。中央値は、外れ値に対して敏感でない、中心化傾向の測定値です。それに対して平均値は、いくつかの極端に大きい、または小さい値に影響されます。50 番目のパーセンタイルまたは 2 番目の四分位でもあります。
- カウント**:** キー・フィールドの各組み合わせの非ヌル値のカウントを返します。

フィールドからレコードへ

フィールド**:** 「フィールド」リストには、行列入替ノードに入力するすべてのフィールドが含まれます。

インデックス**:** 「インデックス」セクションを使用してインデックス・フィールドとして使用するフィールドを選択します。

値**:** 「値」セクションを使用して値フィールドとして使用するフィールドを選択します。値フィールドを選択しない場合、割り当てられていないすべての数値フィールドが値として使用されます。ただし、数値ではないフィールドが使用できる場合、割り当てられていないすべての文字列フィールドが使用されます。

時系列ノード

時系列ノードは、多くの場合、時系列データなどの継続的なデータに使用されます。時系列ノードを使用して、前のレコードのフィールドのデータを含む新規フィールドを作成します。時系列ノードを使用する際には、あらかじめ特定のフィールドでソートされたデータがあれば便利です。このためには、ソート・ノードを使用します。

時系列ノードのオプションの設定

選択したフィールド**:** フィールド・ピッカー (テキスト・ボックスの右側にあるボタン) を使用して、時系列データを取得するフィールドを選択します。選択した各フィールドを使用して、データ・セット中のすべてのレコードに対する新しいフィールドが作成されます。

オフセット**:** 時系列フィールド値を抽出する最新レコードが、現在のレコードのいくつ前にあるかを指定します。例えば、「オフセット」を 3 に設定すると、各レコードがこのノードを通過するときに、3 つ前のレコードのフィールド値が現在のレコードに追加されます。「スパン」の設定を使用して、いくつ前のレコードまでの値を抽出するか指定します。オフセット値を調整するには、矢印を使用します。

スパン**:** 値を抽出する元になる前のレコードの数を指定します。例えば、「オフセット」を 3 に設定し、「スパン」を 5 に設定した場合は、ノードを通過する各レコードに対し、「選択したフィールド」リストで指定した各フィールドごとに 5 つのフィールドが追加されます。つまり、ノードがレコード 10 を処理するときには、レコード 7 からレコード 3 までのフィールドが追加されます。スパン値を調整するには、矢印を使用します。

時系列がない場合**:** 時系列値がないレコードの処理方法を、次の 3 つのオプションから選択します。このようなレコードは、時系列として使用する前のレコードがない、データ・セットの先頭数レコードなどがあてはまります。

- レコードを破棄**:** 選択したフィールドで時系列値を利用できない場合は、そのレコードを破棄します。
- 未定義の時系列を保持**:** 時系列値がない場合もレコードを保持します。この場合、時系列値には未定義の値が入れられ、`$null$` として表示されます。
- 次の値を入れる**:** 時系列値がないレコードに対して使用する値または文字列を指定します。デフォルトは、システム ヌル値の *undef* です。ヌル値は文字列 `$null$` で表されます。

置換値を選択する場合は、適切な処理を行うために、次の規則にしたがってください。

- 選択するフィールドは、それぞれ同じストレージ・タイプでなければなりません。
- 選択したすべてのフィールドのストレージ・タイプが数値の場合、置換値は整数でなければなりません。
- 選択したすべてのフィールドのストレージ・タイプが実数の場合、置換値は実数でなければなりません。
- 選択したすべてのフィールドのストレージ・タイプがシンボル値の場合、置換値は文字列でなければなりません。
- 選択したすべてのフィールドのストレージ・タイプが日付/時間の場合、置換値は日付/時間フィールドでなければなりません。

上記の条件を満たさない場合は、時系列ノードの実行時にエラーが発生してしまいます。

フィールド順序ノード

フィールド順序ノードにより、下流のフィールドを表示するために使用する順序を定義することができます。この順序は、テーブル、リスト、およびフィールド・ピッカーなど、さまざまな場所のフィールドの表示に適用されます。この操作は、さまざまなデータ・セットにおいて、特定のフィールドをより参照しやすくする場合などに役立ちます。

フィールド順序ノードのオプションの設定

ファイルを並べ替えるには、次の 2 つの方法があります。ユーザー指定の順序と自動ソートです。

カスタム配列

「ユーザー指定の順序」 を選択すると、フィールド名とデータ型のテーブルが有効になります。このテーブルから、すべてのフィールドを参照したり、矢印ボタンを使用して独自の並び順を作成することができます。

フィールドを並び替えるには

1. テーブル中のフィールドを選択します。複数のフィールドを選択するには、Ctrl キーを押しながらフィールドを選択します。
2. 単純な矢印ボタンをクリックすると、フィールドが 1 行上または下に移動します。
3. 線の付いた矢印ボタンを使用すると、フィールドがリストの最上位または最下位に移動します。
4. ここに含まれていないフィールドの順序を指定するには、**「他のフィールド」**として示されている区切り行を上または下に移動します。

「他のフィールド」の詳細

他のフィールド**:** 区切り行「他のフィールド」の目的は、テーブルを 2 つに分割することです。

- この区切り行の上に表示されているフィールドは、このノードの下流におけるフィールド表示の並び順の一番上に表示されます (テーブルに表示されているように)。
- この区切り行の下に表示されているフィールドは、このノードの下流におけるフィールド表示の並び順の一番下に表示されます (テーブルに表示されているように)。

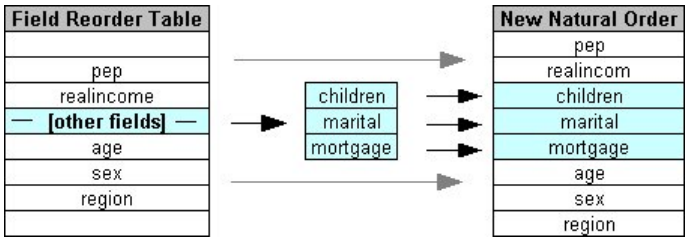


図 *6.* 新しいフィールドの並び順に「他のフィールド」がどのように組み込まれるかを示す図

- ここに表示されていない他のフィールドは、区切り行の位置で示される、「最上位のフィールド」と「最下位のフィールド」の間に配置されます。

他のカスタム・オプションには、次のようなものがあります。

- 各列見出しの矢印をクリックすると、その列が昇順または降順にソートされます (「データ型」、「名前」、または 「ストレージ」)。列でソートする場合、ここで指定されていないフィールド (**「他のフィールド」**行で示されます) は、最後に通常の順序でソートされます。

- フィールド順序ノードから未使用のフィールドをすべて削除するには、「未使用を消去」 をクリックします。未使用のフィールドは、テーブル中に赤で表示されます。これは、上流の操作でそのフィールドが削除されたことを表しています。
- 任意の新規フィールド (稲妻のアイコンは、新規または未指定のフィールドを示します) の順序を指定します。「**OK**」 または 「適用」 をクリックすると、アイコンが消えます。

注 : ユーザー指定の順序を適用した後に、上流でフィールドが追加された場合、その新規フィールドはカスタム リストの最後に追加されます。

自動ソート

ソート用パラメーターを指定する場合は、「自動ソート」 を選択します。ダイアログ・ボックスに、自動ソート用のオプションが表示されます。

ソート項目**:** フィールド順序ノードに読み込まれるフィールドのソート方法を選択します。矢印ボタンは、ソートが昇順に行われるか、または降順に行われるかを示しています。目的の並び順を選択してください。

- 名前
- タイプ
- ストレージ

自動ソートが行われた後に、フィールド順序ノードの上流にフィールドが追加された場合、そのフィールドも設定内容に応じて自動的に適切な位置に配置されます。

時間区分ノード

SPSS Modeler バージョン 17.1 以前で使用可能であったオリジナルの時間区分ノードは、Analytic Server (AS) との互換性がなく、SPSS Modeler リリース 18.0 で廃止されました。

それに置き換わる時間区分ノードには、オリジナルの時間区分ノードからの変更点が多数含まれています。この新しいノードは、Analytic Server または SPSS Modeler 自体と一緒に使用できます。

時間区分ノードは、間隔を指定し、推定または予測のための新しい時間フィールドを作成するために使用します。秒単位から年単位まで、すべての時間区分がサポートされます。

新しい時間フィールドを作成するには、このノードを使用します。新しい時間フィールドのストレージ タイプは、選択した入力時間フィールドと同じになります。AS 時間区分ノードにより、以下の項目が生成されます。

- 「フィールド」タブで「時間フィールド」として指定されたフィールド (選択した接頭辞/接尾辞が付加されます)。デフォルトの接頭辞は `$TI_` です。
- 「フィールド」タブで「次元フィールド」として指定されたフィールド。
- 「フィールド」タブで「集合するフィールド」として指定されたフィールド。

選択した区分または周期 (測定値が範囲内に収まる分や秒など) に従って複数の追加フィールドを生成することもできます。

時間区分 - フィールド オプション

新しい時間区分の作成元データを選択するには、 時間区分ノードの「フィールド」タブを使用します。

フィールド ノードに対するすべての入力フィールドが、その尺度タイプのアイコンと共に表示されます。時間フィールドの尺度タイプは、いずれも「連続型」です。入力として使用するフィールドを選択してください。

時間フィールド 新しい時間区分の作成元となる入力フィールドが表示されます。単一の連続型フィールドのみ使用することができます。このフィールドは、区分を変換するための集計キーとして時間区分ノードによって使用されます。新規フィールドのストレージ タイプは、選択された入力時間フィールドと同じになります。整数フィールドを選択した場合は、時間インデックスと見なされます。

ディメンション フィールド **(Dimensions fields)** オプションで、ここにフィールドを追加して、フィールド値に基づく個々の時系列を作成できます。単純な例として、地理空間データの場合は、ディメンションとしてポイント フィールドを使用できます。この例の場合は、時間区分ノードからのデータ出力が、ポイント フィールドのそれぞれのポイント値に対して時系列にソートされます。

ディメンションは、フラット化された多次元データ (TM1 ノードで生成されるデータに類似) を使用するときや、地理空間などのより複雑なデータ型をサポートする場合に理想的です。基本的には、「次元フィールド」の使用法は、SQL クエリーの **`Group By`** 節と同等か、レコード集計ノードの「キー・フィールド」に類似していると考えることができます。ただし、実際には、「次元フィールド」の方が、行と列の従来のデータよりも複雑なデータ構造を処理できるため、より洗練されています。

集合するフィールド 時間フィールドの期間の変更の一環として集計するフィールドを選択します。ここで選択したフィールドのみが、「指定されたフィールドのカスタム設定」テーブルの「ビルド」タブで使用可能になります。指定されなかったフィールドはすべて、ノードから出力されるデータから除外されます。そのため、「フィールド」リストに残されたフィールドはすべてデータから除外されます。

時間区分 - ビルド オプション

尺度タイプに基づいて、時間間隔を変更するためのオプションや、データのフィールドの集計方法を指定するには、「ビルド」タブを使用します。

データを集計すると、既存の日付フィールド、時刻フィールド、タイムスタンプ フィールドは、生成された各フィールドによって置き換えられ、出力から除外されます。その他のフィールドは、このタブで指定したオプションに基づいて集計されます。

時間区分 系列を作成するための間隔と周期を選択します。

デフォルト設定 各種のデータに適用するデフォルトの集計方法を選択します。デフォルトは、尺度を基に適用されます。例えば、連続型フィールドは合計を使用して集計され、一方で名義型フィールドは最頻値を使用します。デフォルトは、以下の 3 つの尺度に対して設定できます。

- 連続型 連続型フィールドに利用できる関数には、「合計」、「平均値」、「最小値」、「最大値」、「中央値」、「第 **1** 四分位数」、および「第 **3** 四分位数」があります。
- 名義型 オプションには「最頻値」、「最小値」、および 「最大値」 があります。
- フラグ型 オプションは「いずれかが真の場合は真」 または「いずれかが偽の場合は偽」のいずれかです。

指定されたフィールドのカスタム設定 個々のフィールドのデフォルトの集計設定に対する例外を指定できます。テーブルに対してフィールドを追加したり削除したりするには、右側にあるアイコンを使用します。フィールドに使用する集計関数を変更するには、該当する列のセルをクリックします。不明のフィールドはこのリストから除外され、テーブルに加えることはできません。

新規フィールド名拡張子 ノードによって生成されるすべてのフィールドに適用する「接頭辞」または「接尾辞」を指定します。

再投影ノード

地理空間データまたはマップ データの場合に座標の特定に使用する最も一般的な 2 つの方法は、投影座標系と地理座標系です。IBM SPSS Modeler では、Clem 式ビルダーの地理空間関数、時空間予測 (STP) ノード、マップ視覚化ノードなどの項目で投影座標系を使用します。そのため、地理座標系で記録されたデータをインポートする場合は、データの再投影の必要があります。可能な場合は、地理空間フィールド (地理空間の尺度を持つすべてのフィールド) が、インポート時ではなく使用時に自動的に再投影されます。自動的に再投影できないフィールドがある場合は、再投影ノードを使用して座標系を変更してください。この方法で再投影すると、誤った座標系を使用したためにエラーが発生した状態を修正できるということになります。

再投影によって座標系を変更する必要がある状態の例を以下のリストに示します。

- レコード追加 地理空間フィールドの座標系が異なる 2 つのデータセットを追加しようとすると、SPSS Modeler によって次のエラーメッセージが表示されます。**<*Field1*> と <*Field2*> の座標系に互換性がありません。一方または両方のフィールドを同じ座標系に再投影してください。**

  <*Field1*> および <*Field2*> は、エラーの原因になった地理空間フィールドの名前です。
- *if/else* 式 if/else ステートメントを含む式を使用し、式の両方の部分に地理空間フィールドまたは戻りの型を指定している場合、座標系が異なっていると SPSS Modeler によって次のエラーメッセージが表示されます。**条件式に、互換性のない戻りの型 <*arg1*> および <*arg2*> が含まれています**

  <*arg1*> および <*arg2*> は、座標系が異なる地理空間タイプを返す then 引数または else 引数です。
- 地理空間フィールドのリストの構成 多数の地理空間フィールドから構成されるリスト フィールドを作成するには、リスト式に指定するすべての地理空間フィールド引数の座標系が同じでなければなりません。同じではない場合は、次のエラーメッセージが表示されます。**<*Field1*> と <*Field2*> の座標系に互換性がありません。一方または両方のフィールドを同じ座標系に再投影してください。**

座標系について詳しくは、「SPSS Modeler User's Guide」の『Working with Streams』セクションに記載されている 『Setting Geospatial Options for Streams』というトピックを参照してください。

再投影ノードのオプションの設定

フィールド

地理フィールド
: デフォルトではこのリストは空です。地理空間フィールドを「再投影されるフィールド」リストからこのリストに移動して、それらのフィールドが確実に再投影されないようにすることができます。

再投影されるフィールド
: デフォルトでは、このノードに入力されるすべての地理空間フィールドがこのリストに含まれています。このリストにあるすべてのフィールドが、「座標系」領域で設定した座標系に再投影されます。

座標系

ストリームのデフォルト
: デフォルトの座標系を使用するには、このオプションを選択します。

指定　このオプションを選択すると、「変更」ボタンを使用して「座標系の選択」ダイアログボックスを表示し、再投影に使用する座標系を選択することができます。

座標系について詳しくは、「SPSS Modeler User's Guide」の『Working with Streams』セクションに記載されている 『Setting Geospatial Options for Streams』というトピックを参照してください。

第 5 章 グラフ作成ノード

グラフ作成ノードの共通の機能

IBM SPSS Modeler に取り入れたデータを調べるために、データ・マイニングのさまざまなフェーズでグラフやチャートが使用されます。例えば、散布図ノードや棒グラフ・ノードをデータ・ソースに接続して、データの型や分布を知ることができます。その後、レコードやフィールドを操作して、モデル作成操作用のデータを準備できます。また、新しく作成されたフィールド間の分布や相関関係を確認する場合にも、グラフがよく使用されます。

「グラフ」パレットには次のノードがあります。

グラフボード・ノードでは、単一のノードにさまざまな種類のグラフを提供しています。このノードを使用して、検証するデータ・フィールドを選択肢、選択したデータに使用できるグラフを選択できます。選択したフィールドに適していないグラフの種類は、ノードによって自動的に除外されます。

散布図ノードで、数値フィールド間の関係が示されます。作図は、点 (散布図) または折れ線を使用して作成できます。

棒グラフ・ノードで、ローンの種類や性別など、シンボル値（カテゴリー) の出現頻度を表示します。通常、棒グラフ・ノードを使用してデータの不均衡を表示しますが、そのデータはモデルの作成前にバランス・ノードを使用して修正できます。

ヒストグラム・ノードでは、数値フィールドの値の出現頻度が示されます。多くの場合、ヒストグラム・ノードは、操作やモデルの構築前にデータを調べるために使用されます。棒グラフ・ノードと同様、ヒストグラム・ノードにより、データ内の不均衡がしばしば明らかになります。

集計棒グラフ・ノードで、他の数値フィールドの値に相対的な数値フィールドの値の棒グラフを表示します（集計棒グラフ・ノードでは、ヒストグラムに似たグラフが作成されます）。集計棒グラフは、値が時間の経過とともに変化する変数やフィールドを表示する場合に役立ちます。3 次元グラフを使用して、分布をカテゴリー別に表示するシンボル値軸を追加することもできます。

線グラフ・ノードでは、1 つの *X* フィールドに対して複数の *Y* フィールドを表示する作図が作成されます。*Y* フィールドは色付きの線で作図され、それぞれ「スタイル」フィールドを「ライン」に、「X モード」フィールドを「ソート」に設定した散布図ノードに相当します。線グラフは、複数の変数の変動を長期にわたって調査するときに役立ちます。

Web グラフ・ノードで、複数のシンボル値 (カテゴリー) フィールドの値の関係の強さが示されます。このグラフでは、接続の強さを示すためにさまざまな幅の線が使用されます。Web グラフ・ノードを使用して、例えば、E コマース・サイトで購入されたさまざまな商品の関係を調査できます。

時系列ノードで、時系列データの 1 つ以上のセットを表示します。通常、最初に時間区分ノードを使用して *TimeLabel* フィールドを作成します。このフィールドは、*x* 軸にラベルを付けるために使用されます。

評価ノードは、予測モデルの評価と比較に用いられます。評価グラフで、モデルが特定の結果をどの程度予測するかを表示します。それによって、予測値と予測の信頼度に基づいたレコードがソートされます。そして、レコードが等サイズ (分位) のグループに分割され、各分位のビジネスに関する基準の値が、高い方から降順で作図されます。作図には、複数のモデルが異なる線で示されます。

マップ視覚化ノードは、複数の入力接続を受け入れて、地理空間データを一連の層としてマップに表示することができます。各層は単一の地理空間フィールドです。例えば、基本層を国のマップとし、その上に道路の層、川の層、町の層を設けることができます。

E 散布図 (ベータ) ノードで、数値フィールド間の関係が示されます。これは散布図ノードに類似していますが、オプションは異なり、出力にはこのノードに固有の新規グラフ インターフェースを使用します。新規グラフ機能を利用するには、このベータ レベル ノードを使用します。

t 分布 Stochastic Neighbor Embedding (t-SNE) は、高次元データの視覚化のためのツールです。t-SNE は、データ ポイントの類似性を確率に変換します。SPSS Modeler の t-SNE ノードは Python で実装されており、`scikit-learn`© Python ライブラリーを必要とします。

グラフ作成ノードをストリームに追加すると、ノードをダブルクリックしてオプションを指定するダイアログ・ボックスを開くことができます。大部分のグラフでは、1 つまたは複数のタブにさまざまな固有のオプションが用意されています。また、すべてのグラフに共通なタブ オプションも数多く用意されています。これらの共通オプションについて詳しくは、次の各セクションを参照してください。

グラフ作成ノードにオプションを環境設定すると、ダイアログ・ボックス内またはストリームの一部としてグラフを実行できます。生成されたグラフ・ウィンドウ内で、データを選択するか領域を指定し、効果的にデータの「サブセットを作成」することで、フィールド作成 (設定とフラグ) ノードと条件抽出ノードを生成することができます。例えば、強力なこの機能を使用し、外れ値を識別してそれを除外することができます。

外観、オーバーレイ、パネル、およびアニメーション

オーバーレイと外観

外観 (および重ね書き) は、視覚化に次元数を追加します。外観の効果 (グループ化、クラスター化、積み上げ) は、視覚化タイプ、フィールドまたは変数の種類、およびグラフ要素の種類と統計に依存します。例えば、色についてカテゴリー・フィールドを使用して、散布図のポイントをグループ化し、または積み上げ棒グラフ内の積み上げを作成します。また、色について連続した数値の範囲を使用して、散布図の各ポイントの範囲の値を示します。

外観およびオーバーレイを検証して、ニーズを満たすものを探す必要があります。次の説明で、適切な外観およびオーバーレイを選択できます。

注：すべての外観またはオーバーレイが、すべての視覚化の種類に使用できるわけではありません。

- 色**:** 色がカテゴリー・フィールドによって定義されている場合、個々のカテゴリーに基づいて、各カテゴリーに 1 色ずつ、視覚化を分割します。色が連続する数値の範囲を表す場合、範囲型フィールドの値に基づいて色が異なります。グラフィック要素 (バーやボックスなど) が複数のレコードやケースを表し、色に対して範囲型フィールドが使用されている場合、色は範囲型フィールドの平均値 によって異なります。
- 形状**:** 形状は、さまざまな形状の要素 (カテゴリーごとに 1 つ) に視覚化を分割するカテゴリー・フィールドによって定義されます。
- 透明**:** 透過性がカテゴリー・フィールドによって定義されている場合、個々のカテゴリーに基づいて、各カテゴリーに透過性レベルを 1 つずつ、視覚化を分割します。透過性が連続する数値の範囲を表す場合、範囲型フィールドの値に基づいて透過性が異なります。グラフィック要素 (バーやボックスなど) が複数のレコードやケースを表し、透過性に対して範囲型フィールドが使用されている場合、透過性は範囲型フィールドの平均値 によって異なります。最大値では、グラフィック要素は完全に透明です。最小値では、完全に不透明になります。
- データ・ラベル**:** データ・ラベルは、グラフィック要素に付加されるラベルを作成するために使用される値を持つ任意のタイプのフィールドによって定義されます。
- サイズ**:** サイズがカテゴリー・フィールドによって定義されている場合、個々のカテゴリーに基づいて、各カテゴリーに 1 つのサイズずつ、視覚化を分割します。サイズが連続する数値の範囲を表す場合、範囲型フィールドの値に基づいてサイズが異なります。グラフィック要素 (バーやボックスなど) が複数のレコードやケースを表し、サイズに対して範囲型フィールドが使用されている場合、サイズは範囲型フィールドの平均値 によって異なります。

パネルとアニメーション

パネル**:** ファセットとも呼ばれるパネリングによって、 グラフのテーブルを作成します。1 つのグラフがパネリング・フィールドの各カテゴリーに生成されますが、すべてのパネルは同時に表示されます。パネリングは、視覚化がパネル・フィールドの条件に従っているかどうかを確認する場合に役立ちます。例えば、ヒストグラムを性別によってパネリングし、度数分布が男性と女性で等しいかどうかを確認することができます。つまり、給与が性差の影響を受けているかどうかを確認することができます。パネリング向けに単一のカテゴリー・フィールドを選択します。

アニメーション**:** アニメーションは、複数のグラフがアニメーション・フィールドの値から作成されるという点類似していますが、これらのグラフは同時に表示されません。検証モードのコントロールを使用して、一連のグラフの出力およびフリップをアニメーション化します。さらに、パネリングとは異なり、アニメーションはカテゴリー・フィールドを必要としません。値が自動的に範囲に分割される連続フィールドを指定

することができます。検証モードのアニメーション・コントロールでさまざまなサイズの範囲を設定することができます。一部の視覚化では、アニメーションは提供されません。

「出力」タブの使用方法

すべての種類のグラフに対して、生成されたグラフのファイル名や表示に関する次のオプションを指定することができます。

注: 棒グラフ・ノードのグラフには、追加の設定があります。

出力名。ノードの実行時に生成されるグラフの名前を指定します。「自動」 は、出力を生成するノードの名前に基づいて名前を選択します。「ユーザー設定」 で別の名前を指定することもできます。

画面に出力。グラフを生成し、新しいウィンドウに表示する場合に選択します。

ファイルに出力。出力をファイルとして保存する場合に選択します。

- 出力グラフ。グラフ形式の出力を生成する場合に選択します。棒グラフ・ノードでのみ使用できます。
- 出力テーブル: テーブル形式の出力を生成する場合に選択します。棒グラフ・ノードでのみ使用できます。
- ファイル名: 生成されたグラフまたはテーブルのファイル名を指定します。「**...**」 ボタンを使用して、ディレクトリーを参照しながら特定のファイルを指定することもできます。
- ファイルの種類。ドロップダウン・リストでファイルの種類を指定します。「出力テーブル」 オプションが指定されている棒グラフ・ノードを除く、すべてのグラフ作成ノードで、使用可能なグラフ ファイルの種類は次のとおりです。

  - ビットマップ (.bmp)
  - PNG (.png)
  - 出力オブジェクト (.cou)
  - JPEG (.jpg)
  - HTML (.html)
  - その他の IBM SPSS Statistics アプリケーションで使用する ViZml ドキュメント (.xml)

  棒グラフ・ノードの 「出力テーブル」 オプションでは、使用できるファイルの種類は次のとおりです。

  - データ (タブ区切り) (*.tab)
  - データ (カンマ区切り) (*.csv)
  - HTML (.html)
  - 出力オブジェクト (.cou)

出力にページ番号を付ける: 出力を HTML として保存するとこのオプションが使用可能になり、各 HTML ページのサイズを制御することができます (棒グラフ・ノードの場合のみ該当します)。

ページ当たりの行数: 「出力のページ分割」を選択するとこのオプションが使用可能になり、各 HTML ページの長さを決定することができます。デフォルトは 400 行です。(棒グラフ・ノードの場合のみ該当します)。

「注釈」タブの使用方法

すべてのノードで使用されるこのタブには、ノード名の変更、カスタム・ツールヒントの提供、および長い注釈の保存などのオプションが用意されています。

3 次元グラフ

IBM SPSS Modeler の散布図および集計棒グラフには、情報を 3 次元に表示する機能があります。この機能は、モデル作成のためにサブセットを選択したり、新規フィールドを作成するために、データを視覚化する場合に役立ちます。

3 次元グラフを作成すると、グラフをクリックしてマウスでドラッグし、回転させてあらゆる角度で表示することができます。

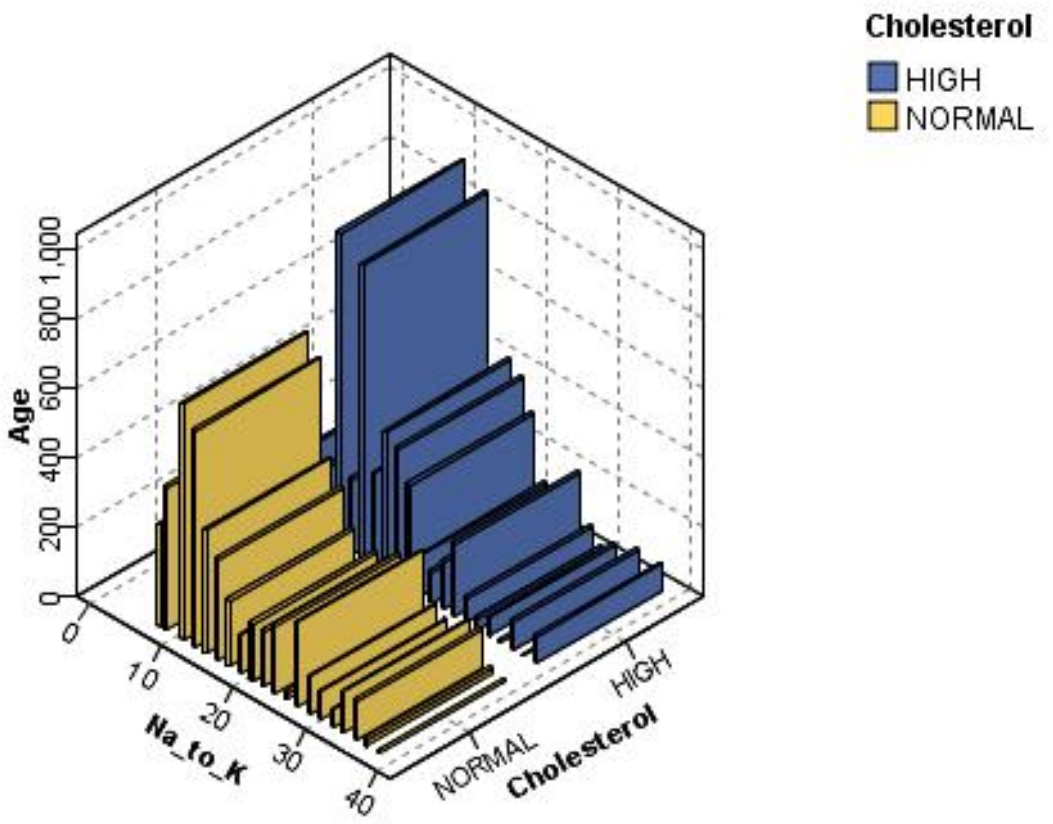


図 7. *x*、*y*、*z* 軸を持つ集計棒グラフ

IBM SPSS Modeler で 3 次元グラフを作成するには、次の 2 種類の方法があります。情報を 3 番目の軸に作図する方法 (真の 3 次元グラフ) と、グラフを 3 次元効果で表示する方法です。どちらの方法も、散布図や集計棒グラフで利用することができます。

情報を 3 番目の軸に作図するには

1. 「グラフ作成ノード」ダイアログ・ボックスで、「散布図」 タブをクリックします。
2. z 軸のオプションを有効にするには、「3 次元」ボタンをクリックします。
3. フィールド・ピッカー・ボタンを使用して、z 軸のフィールドを選択します。場合によっては、シンボル値フィールドしか選択できないこともあります。フィールド・ピッカーが、適切なフィールドを表示します。

グラフに 3 次元効果を追加するには

1. グラフを作成したら、出力ウィンドウの 「グラフ」 タブをクリックします。
2. 「3 次元」ボタンをクリックして、 3 次元グラフ ビューに切り替えます。

Graphboard ノード

グラフボード・ノードを使用すると、1 つのノードで多数の異なるグラフ出力 (棒グラフ、円グラフ、ヒストグラム、散布図、ヒート・マップなど) から選択することができます。まず、最初のタブで、必要なデータ・フィールドを選択すると、データに使用できるグラフの種類の選択肢が表示されます。選択したフィールドに適さないグラフの種類は、ノードによって自動的に除外されます。「詳細」タブでは、詳細な、またはより高度なグラフ・オプションを定義できます。

注**:** グラフボード・ノードを編集したりグラフの種類を選択したりするには、データを含むストリームにグラフボード・ノードを接続する必要があります。

どの視覚化テンプレート (およびスタイル・シートとマップ) を使用するかを制御するための 2 つのボタンが用意されています。

「管理」。コンピューターで視覚化テンプレート、スタイル・シート、およびマップを管理します。視覚化テンプレート、スタイル・シート、およびマップをローカル・マシンでインポート、エクスポート、名前変更、および削除できます。詳しくは、トピック 231 ページの『テンプレート、スタイル・シート、マップ・ファイルの管理』を参照してください。

「場所」。視覚化テンプレート、スタイル・シート、およびマップが保管されている場所を変更します。現在の場所は、ボタンの右側に表示されます。 詳しくは、トピック 230 ページの『テンプレート、スタイル・シート、マップの位置の設定』を参照してください。

グラフボード [基本」タブ

データを良好に表示できる視覚化の種類がわからない場合、「基本」タブを使用します。データを選択する場合、データに適切な視覚化のサブセットが提供されます。詳しくは、 218 ページの『グラフボードの例』を参照してください。

1. リストから 1 つまたは複数のフィールド (変数) を選択します。複数のフィールドを選択するには、Ctrl キーを押したままクリックします。

   フィールドの測定の尺度によって、使用できる視覚化の種類が決まります。リストでフィールドを右クリックし、オプションを選択して測定の尺度を変更できます。使用できる測定の尺度の種類に関する詳細は、 206 ページの『フィールド (変数) タイプ』を参照してください。

2. 視覚化の種類を選択します。使用できる種類の詳細は、 210 ページの『組み込まれている利用可能なグラフボード視覚化タイプ』を参照してください。
3. 特定の視覚化について、要約統計を選択することができます。統計が度数の統計量か連続フィールドから計算されているかによって、異なる統計のサブセットを使用できます。使用できる統計も、テンプレートによって異なります。使用できる統計の完全リストは次のステップの後にあります。
4. オプションの外観およびパネル・フィールドなど、さらにオプションを定義する場合、「詳細」をクリックします。 詳しくは、トピック 208 ページの『グラフボード [詳細」タブ』を参照してください。

連続型フィールドから計算された要約統計

- *Mean (*平均*).* 中心傾向の指標。算術平均 (合計をケース数で割った値) です。
- *Median (*中央値*).* この値より上と下それぞれにケースの半数ずつが該当することになる値。50 パーセンタイル。ケース数が偶数の場合の中央値は、昇順または降順にソートしたときに中央に来る 2 つのケースの平均です。中央値は、外れ値に対して敏感でない、中心傾向の指標です。それに対して平均値は、少数の極端に大きいまたは小さい値に影響されることがあります。
- *Mode (*最頻値*).* 最も多く出現する値。複数の値が最高の頻度で出現し、その頻度が同じである場合は、それぞれが最頻値となります。
- *Minimum (*最小値*).* 数値変数の最小値。
- *Maximum (*最大*).* 数値変数の最大値。
- 「範囲」。最小値と最大値の差異。
- 「中間域」。範囲の中間、つまり最小値との差が最大値との差と等しい値です。
- *Sum (*合計*).* 欠損値でない値を持つすべてのケースにわたる値の和 (合計)。
- 「累積合計」。値の累積合計。各グラフィック要素は、サブグループの合計と、以前のグループすべての総合計を示します。
- 「パーセント合計」。合計した変数に基づいたサブグループ内の、全グループの合計に対するパーセンテージ。
- 「累積パーセント合計」。合計したフィールドに基づいたサブグループ内の、全グループの合計に対する累積パーセンテージ。各グラフィック要素は、サブグループのパーセンテージと、以前のグループすべての総パーセンテージを示します。
- *Variance (*分散 *(*信頼性分析*)).* 平均値の周りの値の散らばりの指標。平均値からの偏差の平方和を、ケース数より 1 少ない値で割ったものに等しくなります。分散の測定単位は、変数自体の単位の 2 乗です。
- *Standard Deviation (*標準偏差*).* 平均値の周りの散らばりの指標。正規分布では、平均から 1 標準偏差以内にケースの 68% が含まれ、2 標準偏差以内にケースの 95% が含まれます。例えば平均年齢が 45 で標準偏差が 10 である場合、正規分布ではケースの 95% が 25 と 65 の間に含まれます。
- *Standard Error (*標準誤差*).* サンプル間で検定統計量の値がどの程度ばらついているかの指標。統計量のサンプル分布の標準偏差です。例えば、平均値の標準誤差はサンプル平均の標準偏差です。
- *Kurtosis (*尖度*).* 外れ値が存在する度合いの指標。正規分布の場合、尖度の統計値は 0 です。尖度が正の場合、そのデータの極端な外れ値は正規分布よりも多いことを示します。尖度が負の場合、そのデータの極端な外れ値は正規分布よりも少ないことを示します。
- *Skewness (*歪度*).* 分布の非対称性の指標。正規分布は対称であり、歪度の値は 0 です。歪度が正の大きな値である分布は、右側の裾が長くなります。歪度が負で絶対値が大きい分布は、左側の裾が長くなります。目安として、歪度が標準誤差の 2 倍より大きい場合は、対称分布からずれていると解釈します。

次の領域の統計で、サブグループごとに複数のグラフィック要素が作成される場合があります。区間、領域、または辺のグラフィック要素を使用する場合、領域統計では、範囲を示すグラフィック要素が作成されます。他のすべてのグラフィック要素は、2 つの要素を生成します。一方は範囲の始点を示し、もう一方は範囲の終点を示します。

- 「領域: 範囲」。最小値と最大値の間の値の範囲。
- 「領域: 平均値の 95 % の信頼区間」。母集団の平均値を含む 95 % の確率をもつ値の範囲。
- 「領域: 個別の 95 % の信頼区間」。個別ケースにおいて予測値を含む 95 % の確率をもつ値の範囲。
- 「領域: 平均値の上下 1 標準偏差」。平均値 の上下 1 標準偏差 間の値の範囲。
- 「領域: 平均値の上下 1 標準誤差」。平均値 の上下 1 標準誤差 間の値の範囲。

度数ベースの要約統計

- 「度数」。行/ケースの数。
- 「累積度数」。行/ケースの累積数。各グラフィック要素は、サブグループの度数と、以前のグループすべての度数合計を示します。
- 「度数のパーセント」。行/ケースの総合計に対する、各サブグループ内の行/ケースのパーセンテージ。
- 「度数の累積パーセント」。行/ケースの総合計に対する、各サブグループ内の行/ケースの累積パーセンテージ。各グラフィック要素は、サブグループのパーセンテージと、以前のグループすべての総パーセンテージを示します。

フィールド (変数) タイプ

フィールドのタイプおよびデータ型を示すアイコンが、変数リストの変数の隣に表示されます。また、アイコンは複数回答セットも示します。

表 32. 測定レベル・アイコン：

測定水準	数値	文字列	日付	時間
連続		n/a		
順序セット				
設定				

表 33. 多重回答セットのアイコン：

複数レスポンス設定タイプ	アイコン
複数回答セット、複数カテゴリー	

表 33. 多重回答セットのアイコン (続き):

複数レスポンス設定タイプ	アイコン
複数回答セット、複合二分	

測定水準

フィールドの尺度は、視覚化作成時に重要です。以下は尺度の説明です。リストでフィールドを右クリックし、オプションを選択して尺度を一時的に変更できます。多くの場合、フィールド、カテゴリー、連続の 2 つの幅広い分類のみを考慮する必要があります。

カテゴリー: 値やカテゴリーの数が限られているデータ (性別や宗教など)。カテゴリー・フィールドは、文字列 (英数字) 変数でも、数値コードを使用してカテゴリーを表す数値フィールドでもかまいません (例えば、0 = 男性、1 = 女性など)。質的データともいいます。セット型、順序セット型、フラグはすべてカテゴリー型フィールドです。

- セット: 本質的な順位のないカテゴリーを表す値を持つフィールド/変数 (従業員が所属する会社の部署など)。名義変数の例としては、地域やジップ・コードや所属宗教などがあります。名義型変数とも呼ばれます。
- 順序セット: 本質的な順位を持つカテゴリーを表す値を持つフィールド/変数 (「非常に不満」から「非常に満足」までのサービス満足度など)。順序セットの例としては、満足度や信頼度を表す得点や嗜好評価得点などがあります。順序変数とも呼ばれます。
- フラグ: 「はい」と「いいえ」や 1 と 2 など、2 つの異なる値を取るフィールド/変数。二分変数や 2 値変数とも呼ばれます。

連続: 間隔または比率尺度について測定したデータです。データ値は、値の順序と値の間の距離を示します。例えば、$72,195 の給料は、$52,398 の給料より高く、2 つの値の距離は $19,797 です。また、量的データ、スケールデータ、数値範囲データとも呼ばれます。

カテゴリー・フィールドは、視覚化のカテゴリーを定義し、通常、それぞれのグラフィック要素を描画、またはグラフィック要素をグループ化します。連続フィールドは、カテゴリー・フィールドのカテゴリー内で要約されることがよくあります。例えば、性別分類内にある収入のデフォルトの視覚化は、男性の平均収入と女性の平均収入を表示します。連続型フィールドの未加工値は、散布図として作図することもできます。例えば、散布図各ケースの現在の給与および初期の給与を表示できます。カテゴリー・フィールドを使用して、ケースを性別ごとに分類できます。

データの型

尺度は、データ型を決定するフィールドのプロパティーだけではありません。フィールドは、特定のデータ型でも保存されます。データ型は、文字列 (文字など、非数値型データ)、数値 (実数)、日付です。尺度と異なり、変数のデータ型を一時的に変更することはできません。データを元のデータ・セットに保存する方法を変更する必要があります。

多重回答グループ

一部のデータ・ファイル、マルチアンサー・セット と呼ばれる特殊な「フィールド」をサポートします。マルチアンサー・セットは、通常の意味での「フィールド」ではありません。マルチアンサー・セットは、

複数のフィールドを使用して質問に対する回答を記録し、回答者は複数の回答をすることができます。マルチアンサー・セットはカテゴリー・フィールドのように扱われ、カテゴリー・フィールドで可能なことのほとんどを複数回答セットでも行うことができます。

マルチアンサー・セットは、多重二分グループの場合と、多重カテゴリー・セットの場合があります。

「多重二分セット」。多重二分セットは通常、複数の二分フィールド (はい/いいえ、有/無、チェック・マークあり/チェック・マークなしのような 2 つの値のみが含まれる可能性があるフィールド) で構成されます。フィールドは厳密に二分というわけではありませんが、セット内のすべてのフィールドは同じようにコード化されます。

例えば、調査で「次のどのニュース ソースを信頼しますか?」という質問に 5 つの回答が設定されているとします。回答者は、各選択肢の隣のボックスをチェックして、複数の選択肢を選択することができます。5 つの回答は、データ・ファイルで 5 つのフィールドとなります。*No* (チェックなし) は 0、 *Yes* (チェック済み) は 1 にコード化されます。

多重カテゴリー・グループ: 多重カテゴリー・グループは、複数のフィールドで構成され、すべて同じ方法でコード化されます。多くの回答カテゴリーがある場合があります。例えば、「あなたの民族の財産を最もよく表す国民性を最大 3 つ挙げてください」という調査項目があるとします。非常に多くの回答が考えられますが、コード化のため、最も一般的な国民性 40 件に制限され、他の回答は「その他」のカテゴリーに分類されます。データ・ファイルでは、3 つの選択すると 3 つのフィールドに該当します。それぞれに 41 件のカテゴリーがあります (40 は国民性をコード化、1 つは「その他」カテゴリー)。

グラフボード [詳細」タブ

作成する視覚化の種類がわかっている場合、オプションの外観、パネルを追加する場合または視覚化をアニメーション化する場合、「詳細」タブを使用します。詳しくは、 218 ページの『グラフボードの例』を参照してください。

1. 「基本」タブで視覚化の種類を選択する場合、その種類が表示されます。「基本」タブで選択しない場合は、ドロップダウン・リストから選択します。視覚化の種類について詳しくは、 210 ページの『組み込まれている利用可能なグラフボード視覚化タイプ』を参照してください。
2. 視覚化の種類に必要なフィールド (変数) を指定するコントロールが、視覚化のサムネイル・イメージのすぐ右側にあります。これらのフィールドのすべてを指定する必要があります。
3. 特定の視覚化について、要約統計を選択することができます。棒グラフなど、透過度の表示にこれらの要約オプションのいずれかを使用することができます。要約統計量について詳しくは、 204 ページの『グラフボード [基本」タブ』を参照してください。
4. 1 つまたは複数のオプションの表示を選択することができます。視覚化に他のフィールドを追加することができる次元性を追加することができます。例えば、フィールドを使用すると、散布図のポイントのサイズを変更することができます。オプションの外観について詳しくは、 201 ページの『外観、オーバーレイ、パネル、およびアニメーション』を参照してください。透過度の表示は、スクリプトにはサポートされていません。
5. マップの視覚化を作成している場合、「マップ・ファイル」 グループが使用されるマップ・ファイルを表示します。デフォルトのマップ・ファイルがある場合、このファイルが表示されます。マップ・ファイルを変更するには、「マップ・ファイルを選択」 をクリックして「マップの選択」ダイアログ・ボックスを表示します。このダイアログ・ボックスでデフォルトのマップ・ファイルを指定することもできます。 詳しくは、トピック 209 ページの『マップ視覚化のためのマップ・ファイルの選択』を参照してください。

6. 1 つまたは複数のパネリングまたはアニメーションを選択することができます。パネリングとアニメーションのオプションについて詳しくは、 201 ページの『外観、オーバーレイ、パネル、およびアニメーション』を参照してください。

マップ視覚化のためのマップ・ファイルの選択

マップの視覚化テンプレートを選択する場合、マップを描画するための地理的情報を定義するマップ・ファイルが必要です。デフォルトのマップ・ファイルがある場合、マップの視覚化に使用されます。異なるマップ・ファイルを選択するには、「詳細」タブの 「マップ・ファイルを選択」 をクリックして「マップの選択」ダイアログ・ボックスを表示します。

「マップの選択」ダイアログ・ボックスを使用して、プライマリ マップ・ファイルと参照マップ・ファイルを選択します。マップ・ファイルは、マップを描画するための地理的情報を定義します。アプリケーションは標準 マップ・ファイルとともにインストールされます。他に使用する ERI シェープファイルがある場合、まずシェープファイルを SMZ ファイルに変換する必要があります。詳しくは、トピック 232 ページの『マップ・シェープファイルの変換と配布』を参照してください。マップを変換した後、「テンプレート ピッカー」ダイアログ・ボックスの 「管理...」をクリックし、「マップを選択」ダイアログ・ボックスで使用できるよう、管理システムにマップをインポートします。

マップ・ファイルを指定する際に考慮するポイントは以下のとおりです。

- すべてのマップ テンプレートには少なくとも 1 つのマップ・ファイルが必要です。
- 通常、マップ・ファイルはマップのキー属性をデータ キーにリンクさせます。
- テンプレートにデータ キーにリンクするマップ キーを必要としない場合、参照マップの要素を描画するための座標 (緯度や経度など) を指定する参照マップ・ファイルおよびフィールドが必要です。
- オーバーレイ・マップ テンプレートには、プライマリ マップ・ファイルと参照マップ・ファイルが必要です。参照マップが最初に描画されます。

属性や特徴など、マップの用語に関する詳細は、 233 ページの『マップの主要な概念』を参照してください。

「マップ・ファイル」。管理システム内にある任意のマップ・ファイルを選択できます。事前にインストールされたマップ・ファイルやインポートしたマップ・ファイルがあります。マップ・ファイル管理に関する詳細は、 231 ページの『テンプレート、スタイル・シート、マップ・ファイルの管理』を参照してください。

「マップ キー」。マップ・ファイルをデータ キーにリンクするキーとして使用する属性を指定します。

「このマップ・ファイルと設定をデフォルトとして保存」。選択したマップ・ファイルをデフォルトとして使用する場合は、このチェック・ボックスを選択します。デフォルトのマップ・ファイルを指定した場合、マップの視覚化を作成するごとにマップ・ファイルを指定する必要はありません。

「データ キー」。このコントロールは、「テンプレート選択」の「詳細」タブに表示されるのと同じ値をリストします。選択した特定のマップ・ファイルによりキーを変更する必要がある場合に便利です。

「視覚化ですべてのマップ・フィーチャーを表示」。このオプションにチェック・マークを付けると、一致するデータ キー値がない場合でも、マップ内のすべてのフィーチャーが視覚化でレンダリングされます。データが存在するフィーチャーだけを表示する場合は、このオプションをクリアします。「マッチしないマップ キー」 リストに表示されたマップ キーで指定されたフィーチャーは、視覚化には表示されません。

「マップとデータ値の比較」。マップ キーとデータ キーは相互にリンクして、マップの視覚化を作成します。マップ キーとデータ キーは、同じドメイン (国や地域など) から取得する必要があります。「比較」

をクリックして、データ キーとマップ キーの値が一致しているかどうかをテストします。表示されたアイコンにより、比較の状態が分かります。これらのアイコンについては、下記のとおりです。比較が実行され、マップ キーの値が一致しないデータ キーの値がある場合、データ キーの値が 「一致しないデータ キー」 リストに表示されます。「一致しないデータ キー」 リストでは、マッチするデータ キー値がないマップ キーを確認することもできます。「視覚化ですべてのマップ・フィーチャーを表示」にチェック・マークが付いていない場合、これらのマップ キーの値で識別されるフィーチャーは表示されません。

表 *34.* 比較アイコン：

アイコン	説明
	比較が実行されていません。「比較」 をクリックする前のデフォルトの状態です。データ キーとマップ キーの値が一致しているか分からないため、注意して続行する必要があります。
	比較が実行され、データ キーとマップ キーの値が一致しています。データ キーのすべての値で、マップ キーに識別された一致フィーチャーがあります。
	比較が実行され、データ キーとマップ キーの値が一致していないものがあります。一部のデータ キーについて、マップ キーに識別された一致フィーチャーがありません。注意して続行する必要があります。続行すると、マップの視覚化に含まれないデータ値があります。
	比較が実行され、一致しているデータ キーとマップ キーの値がありません。続行してもマップが表示されないため、異なるデータ キーまたはマップ キーを選択する必要があります。

組み込まれている利用可能なグラフボード視覚化タイプ

様々な視覚化タイプを作成することができます。組み込まれている次のタイプは、「基本」タブおよび「詳細」タブのどちらでも使用できます。テンプレートの一部の説明 (特にマップ テンプレート) は、「特殊テキスト」 を使用して「詳細」タブで指定されたフィールド (変数) を識別します。

表 *35.* 利用可能なグラフ タイプ：

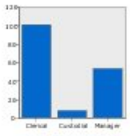
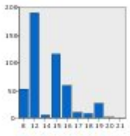
グラフのアイコン	説明	グラフのアイコン	説明
	**Bar** 連続する数値範囲の要約統計量を計算し、カテゴリー・フィールドについて、カテゴリーごとの結果を棒グラフで表示します。 必須: カテゴリー・フィールドおよび連続型フィールド。		度数の棒グラフ カテゴリー・フィールドのカテゴリーごとに、行またはケースの比率を棒グラフで表示します。分布グラフのノードを使用して、このグラフを作成することもできます。このノードには追加のオプションもあります。詳しくは、 253 ページの『棒グラフ・ノード』を参照してください。 必須:1 つのカテゴリー・フィールド。

表 35. 利用可能なグラフ タイプ (続き) :

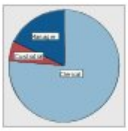
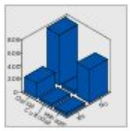
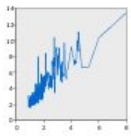
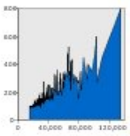
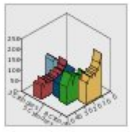
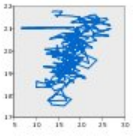
グラフのアイコン	説明	グラフのアイコン	説明
	円グラフ 連続する数値フィールドの合計値を計算し、その合計値のカテゴリー・フィールドのカテゴリーごとの比率を、円を分割して表示します。 必須: カテゴリー・フィールドおよび連続型フィールド。		度数の円グラフ カテゴリー・フィールドのカテゴリーごとに、行またはケースの比率を円を分割して表示します。 必須:1 つのカテゴリー・フィールド。
	**3-D** 棒 連続する数値フィールドの要約統計量を計算し、2 つのカテゴリー・フィールドまたはカテゴリー変数の間のカテゴリーの交差結果を表示します。 必須:カテゴリー・フィールドおよび連続型フィールドのペア。		**3-D** 円グラフ 3-D 効果が追加されている以外は、円グラフと同じです。 必須: カテゴリー・フィールドおよび連続型フィールド。
	**Line** 別のフィールドの値についてフィールドの要約統計量を計算し、値を接続する線を引きます。作図グラフのノードを使用して、折れ線グラフを作成することもできます。このノードには追加のオプションもあります。詳しくは、240 ページの『散布図ノード』を参照してください。 必須:任意の種類のフィールドのペア。		面グラフ 別のフィールドの値についてフィールドの要約統計量を計算し、値を接続する面を描きます。エリアが、下部の色づけされたスペースの線と似ていると言う点で、線とエリアの誤差は最小となります。ただし、色外観を使用すると、線が分割されたり、エリアが積み上げられた状態になります。 必須:任意の種類のフィールドのペア。
	**3-D** 面グラフ あるフィールドの値と作図して、カテゴリー・フィールドごとに分割して表示します。カテゴリーごとに面要素が表示されます。 必須:カテゴリー・フィールドと任意の種類のフィールドのペア。		パス フィールドの値を他の値と作図して、元のデータ・セットの出現順に値を線でつないで表示します。順番は、パスとライン間の大きな違いです。 必須:任意の種類のフィールドのペア。

表 35. 利用可能なグラフ タイプ (続き) :

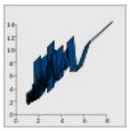
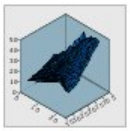
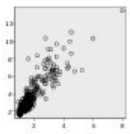
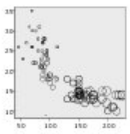
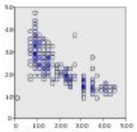
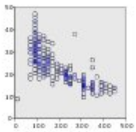
グラフのアイコン	説明	グラフのアイコン	説明
	リボン 別のフィールドの値についてフィールドの要約統計量を計算し、値を接続するリボンを引きます。帯は、基本的には 3-D 効果付きの線です。本格的な 3-D グラフではありません。 必須:任意の種類のフィールドのペア。		表面 3 つのフィールドの値をそれぞれの値と作図して、値同士を面でつないで表示します。 必須:任意の種類の 3 つのフィールド。
	散布図 フィールドの値を、他の値と作図して表示します。フィールド間の関係を (存在する場合に) 際立たせることができます。作図グラフのノードを使用して、散布図を作成することもできます。このノードには追加のオプションもあります。詳しくは、 240 ページの『散布図ノード』を参照してください。 必須:任意の種類のフィールドのペア。		バブル・プロット 基本的な散布図と同様に、フィールドの値を、他の値と作図して表示します。散布図との違いは、3 つ目のフィールドの値が、各ポイントのサイズを変更するのに使われる点です。 必須:任意の種類の 3 つのフィールド。
	分割散布図 基本的な散布図と同様に、フィールドの値を、他の値と作図して表示します。散布図との違いは、類似値がグループのビンに分割され、色またはサイズの外観を使用してそれぞれのビンに含まれるケース数が示される点です。 必須:連続型フィールドのペア。		六角分割散布図 ビン分割散布図の説明を参照してください。違いは、基礎となるビンの形で、丸ではなく六角形をしています。結果として作成される六角ビン分割散布図は、分割散布図と似たものになります。ただし、基礎となるビンの形状のために、ビンごとの値の数はグラフごとに異なります。 必須:連続型フィールドのペア。

表 35. 利用可能なグラフ タイプ (続き) :

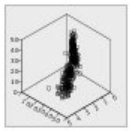
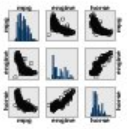
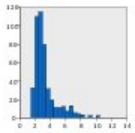
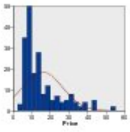
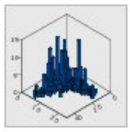
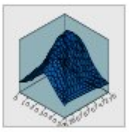
グラフのアイコン	説明	グラフのアイコン	説明
	**3-D Scatterplot (3-D** 散布図**)** 3 つのフィールドを、相互にプロットして表示します。フィールド間の関係を (存在する場合に) 際立たせることができます。作図グラフのノードを使用して、3-D 散布図を作成することもできます。このノードには追加のオプションもあります。詳しくは、 240 ページの『散布図ノード』を参照してください。 必須:任意の種類の 3 つのフィールド。		散布図行列 **(SPLOM)** あるフィールドの値を、各フィールドの他の値と作図して表示します。SPLOM は、散布図のテーブルに似ています。SPLOM には、フィールドごとのヒストグラムも含まれます。 必須:複数の連続型フィールド。
	ヒストグラム フィールドの度数分布表を表示します。ヒストグラムは、分布の種類を決定したり、分布が歪んでいるかを確認するのに役立つ場合があります。ヒストグラム・グラフのノードを使用して、このグラフを作成することもできます。このノードには追加のオプションもあります。詳しくは、 257 ページの『ヒストグラムの「作図」タブ』を参照してください。 要件: 任意のタイプの単一フィールド。		正規分布のヒストグラム 正規分布の曲線を重ね合わせて、連続型フィールドの度数分布表を表示します。 必須:1 つの連続型フィールド。
	**3-D** ヒストグラム 1 組の連続型フィールドの度数分布表を表示します。 必須:連続型フィールドのペア。		**3-D** 密度 1 組の連続型フィールドの度数分布表を表示します。棒ではなく面を使用して分布を表示する以外は、3-D ヒストグラムに似ています。 必須:連続型フィールドのペア。

表 35. 利用可能なグラフ タイプ (続き) :

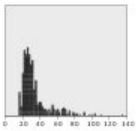
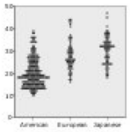
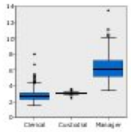
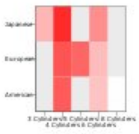
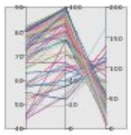
グラフのアイコン	説明	グラフのアイコン	説明
	ドット プロット ケースまたは行を表示し、*x* 軸上の異なるデータ点に積み上げます。このグラフは、データの分布を表している点で、ヒストグラムに似ていますが、特定のビン (値の範囲) の集計数ではなく、それぞれのケースまたは行を表示しています。 要件: 任意のタイプの単一フィールド。		**2-D** ドット・プロット 各ケースまたは各行を表示し、カテゴリー・フィールドのカテゴリーごとに *y* 軸上の異なるデータ点に積み上げます。 必須: カテゴリー・フィールドおよび連続型フィールド。
	**Boxplot** カテゴリー・フィールドのカテゴリーごとに連続型フィールドの 5 種類の統計量 (最小値、最初の 4 分位、中央値、3 番目の 4 分位、最大値) を計算します。結果は Boxplot またはスキーマの要素として表示されます。Boxplot は、連続データの分布がカテゴリー間でどのように変化するかを確認するのに役立つ場合があります。 必須: カテゴリー・フィールドおよび連続型フィールド。		ヒート マップ 2 つのカテゴリー・フィールドのカテゴリーの交差結果に関して、連続型フィールドの平均値を計算します。 必須:カテゴリー・フィールドおよび連続型フィールドのペア。
	並列 フィールドごとに平行軸を作成し、データ内にあるすべての行またはケースのフィールドの値に線を引きます。 必須:複数の連続型フィールド。		度数のコロプレス カテゴリー型フィールドの各カテゴリーの度数 (「データ キー」) を計算し、カテゴリーに対応するマップ・フィーチャーの度数を示すために彩度を使用するマップを描画します。 必須:1 つのカテゴリー・フィールド。キーが「データ キー」カテゴリーに一致するマップ・ファイル。

表 35. 利用可能なグラフ タイプ (続き) :

グラフのアイコン	説明	グラフのアイコン	説明
	平均値/中央値/合計のコロプレス カテゴリー型フィールド (「データキー」) の各カテゴリーに対する連続型フィールド(「色」) の平均値、中央値、連続フィールドの平均値、中央値を計算し、カテゴリーに対応するマップ・フィーチャーで計算された統計を示すために彩度を使用するマップを描画します。 必須: カテゴリー・フィールドおよび連続型フィールド。キーが「データ キー」カテゴリーに一致するマップ・ファイル。		値のコロプレス 別のカテゴリー型フィールド (「データ キー」) で分割した値に対応するマップ・フィーチャーのカテゴリー型フィールド (「色」) の値を示すために色を使用するマップを描画します。各フィーチャーの「色」フィールドに複数のカテゴリー値がある場合、中央値が使用されます。 必須: カテゴリー・フィールドのペア。キーが「データ キー」カテゴリーに一致するマップ・ファイル。
	度数のコロプレスの座標 コロプレス・マップの点を描画するための座標を指定する 2 つの連続型フィールド (経度と緯度) があるという点を除き、度数のコロプレスと類似しています。 必須:カテゴリー・フィールド、および連続型フィールドのペア。キーが「データ キー」カテゴリーに一致するマップ・ファイル。		平均値/中央値/合計のコロプレスの座標 コロプレス・マップの点を描画するために座標を指定する 2 つの連続型フィールド (経度と緯度) があるという点を除き、平均値/中央値/合計のコロプレスと類似しています。 必須 : 1 つのカテゴリー・フィールドおよび 3 つの連続型フィールド。キーが「データ キー」カテゴリーに一致するマップ・ファイル。
	値のコロプレスの座標 コロプレス・マップの点を描画するための座標を指定する 2 つの連続型フィールド (経度と緯度) があるという点を除き、値のコロプレスと類似しています。 必須:カテゴリー・フィールド、および連続型フィールドのペア。キーが「データ キー」カテゴリーに一致するマップ・ファイル。		マップの度数の棒グラフ 各マップ・フィーチャー (データキー) のカテゴリー・フィールド (カテゴリー) の各カテゴリーにおける行/ケースの割合を計算し、マップと、各マップ・フィーチャーの中央に棒グラフを描画します。 必須: カテゴリー・フィールドのペア。キーが「データ キー」カテゴリーに一致するマップ・ファイル。

表 35. 利用可能なグラフ タイプ (続き) :

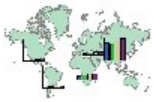
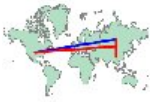
グラフのアイコン	説明	グラフのアイコン	説明
	マップの棒グラフ 連続型フィールド (値) の要約統計量を計算し、各マップ・フィーチャー (データ キー) のカテゴリー・フィールド (カテゴリー) の各カテゴリーに対する結果を各マップ・フィーチャーの中央に棒グラフとして表示します。 必須:カテゴリー・フィールドおよび連続型フィールドのペア。キーが「データ キー」カテゴリーに一致するマップ・ファイル。		マップの度数の円グラフ 各マップ・フィーチャー (データ キー) のカテゴリー・フィールド (カテゴリー) の各カテゴリーにおける行/ケースの割合を計算し、マップと、各マップ・フィーチャーの中央に円グラフのスライスとして割合を描画します。 必須: カテゴリー・フィールドのペア。キーが「データ キー」カテゴリーに一致するマップ・ファイル。
	マップの円グラフ 各マップ・フィーチャー (データ キー) のカテゴリー・フィールド (カテゴリー) の各カテゴリーにおける連続型フィールドの合計 (値) を計算し、マップと、各マップ・フィーチャーの中央に円グラフのスライスとして合計を描画します。 必須:カテゴリー・フィールドおよび連続型フィールドのペア。キーが「データ キー」カテゴリーに一致するマップ・ファイル。		マップの折れ線グラフ 各マップ・フィーチャー (データ キー) の別のフィールド (**X**) に対する連続型フィールドの要約統計量 (**Y**) を計算し、マップと、各マップ・フィーチャーの中央に値を繋ぐ折れ線グラフを描画します。 必須:カテゴリー・フィールドと任意の種類のフィールドのペア。キーが「データ キー」カテゴリーに一致するマップ・ファイル。
	参照マップの座標 点の座標を示す連続型フィールド (経度および緯度)を使用してマップおよび点を描画します。 必須:範囲フィールドのペア。マップ・ファイル。		参照マップの矢印 マップと、各矢印の始点 (始点の経度と始点の緯度) および終点 (終点の経度と終点の緯度)が描画されます。データのレコード/ケースはマップ内の矢印に対応します。 必須:4 つの連続型フィールド。マップ・ファイル。

表 35. 利用可能なグラフ タイプ (続き) :

グラフのアイコン	説明	グラフのアイコン	説明
	ポイント・オーバーレイ・マップ 参照マップを描画し、点フィーチャーを連続型フィールド (色) で色づけして別のポイント・マップに重ねます。 必須: カテゴリー・フィールドのペア。キーが「データ キー」カテゴリーに一致するポイント・マップ・ファイル。参照マップ・ファイル。		ポリゴン・オーバーレイ・マップ 参照マップを描画し、ポリゴン・フィーチャーを連続型フィールド (色) で色づけして別のポリゴン・マップに重ねます。 必須: カテゴリー・フィールドのペア。キーが「データ キー」カテゴリーに一致するポリゴン・マップ・ファイル。参照マップ・ファイル。
	ライン・オーバーレイ・マップ 参照マップを描画し、ライン・フィーチャーを連続型フィールド (色) で色づけして別のライン・マップに重ねます。 必須: カテゴリー・フィールドのペア。キーが「データ キー」カテゴリーに一致するライン・マップ・ファイル。参照マップ・ファイル。		

マップ視覚化の作成

多くの視覚化は、対象となるフィールド (変数) と、それらのフィールドを視覚化するためのテンプレートを選択するだけで、実行することができます。それ以上の選択や操作は必要ありません。マップの視覚化では、少なくともう 1 つのステップ (マップの視覚化用の地理情報を定義するマップ・ファイルの選択) が必要になります。

単純なマップを作成するための基本的な手順は以下のとおりです。

1. 「基本」タブで、対象となるフィールドを選択します。各種のマップ視覚化に必要なフィールドの種類と数については、 210 ページの『組み込まれている利用可能なグラフボード視覚化タイプ』を参照してください。
2. マップ・テンプレートを選択します。
3. 「詳細」タブをクリックします。
4. 「データ キー」などの必須のドロップダウン・リストが正しいフィールドに設定されていることを確認します。
5. 「マップ・ファイル」グループで「マップ・ファイルを選択」をクリックします。
6. 「マップの選択」ダイアログ・ボックスを使用して、マップ・ファイルとマップ キーを選択します。マップ キーの値は、「データ キー」で指定したフィールドの値と一致している必要があります。「比較」ボタンを使用すると、これらの値を比較することができます。オーバーレイ・マップ・テンプレートを選択する場合は、参照マップも選択する必要があります。参照マップはデータにはリンクされませ

ん。参照マップは、メイン・マップの背景として使用されます。「マップの選択」ダイアログ・ボックスについて詳しくは、 209 ページの『マップ視覚化のためのマップ・ファイルの選択』を参照してください。

7. 「**OK**」をクリックして「マップの選択」ダイアログ・ボックスを閉じます。
8. グラフボード・テンプレート選択で「実行」 をクリックしてマップ視覚化を作成します。

グラフボードの例

このセクションでは、使用可能なオプションについていくつかの例を挙げて説明します。これらの例では、結果として得られる視覚化の解釈についても説明します。

これらの例では、*graphboard.str* というストリームを使用します。このストリームは、*employee_data.sav*、*customer_subset.sav*、*worldsales.sav* というデータ・ファイルを参照します。これらのファイルは、いずれの IBM SPSS Modeler Client インストール済み環境でも、*Demos* フォルダーに格納されています。これは、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*graphboard.str* ファイルは、*streams* フォルダーにあります。

記載されている順に例を読み進めることをお勧めします。これは、後半の例は前半の例に基づいて作成されているためです。

例: 要約統計量を使用した棒グラフ

ここでは、セット型/カテゴリー変数のカテゴリーごとに、連続型数値フィールド/変数を要約した棒グラフを作成します。具体的には、男性と女性の平均給与を示す棒グラフを作成します。

この例と、これ以降のいくつかの例では、*Employee data* を使用します。これは、ある企業の従業員に関する情報が格納された仮定のデータ・セットです。

1. *employee_data.sav* を指す Statistics ファイル・ソース・ノードを追加します。
2. グラフボード・ノードを追加して、編集用にそれを開きます。
3. 「基本」タブで、 *Gender* と *Current Salary* を選択します (複数のフィールド/変数を選択するには、Ctrl とクリックを使用します)。
4. 「棒グラフ」を選択します。
5. 「要約」ドロップダウン・リストから「平均値」を選択します。
6. 「実行」をクリックします。
7. 次に表示される画面で、「フィールドと値ラベルを表示」ツールバー・ボタン (ツールバーの中央にある 2 個組のうちの 2 番目) をクリックします。

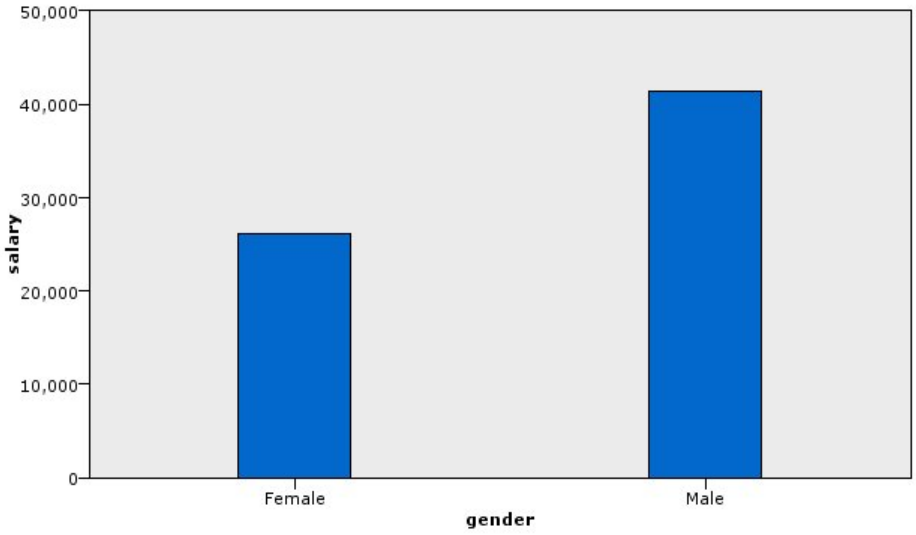


図 *8.* 要約統計量を使用した棒グラフ

以下のことがわかります。

- 棒の高さから、男性の平均給与が女性の平均給与より高いことが明確にわかります。

例: 要約統計量を使用した積み上げ棒グラフ

ここでは、積み上げ棒グラフを作成して、男性と女性の平均給与の差が職種の影響を受けているかどうかを調べます。特定の職種では、女性のほうが男性よりも平均的に給与が高いことが考えられます。

注: この例では *Employee data* を使用します。

1. グラフボード・ノードを追加して、編集用にそれを開きます。
2. 「基本」タブで *Employment Category* と *Current Salary* を選択します。(複数のフィールド/変数を選択するには、Ctrl とクリックを使用します。)
3. 「棒グラフ」を選択します。
4. 「要約」リストから「平均値」を選択します。
5. 「詳細」タブをクリックします。前のタブで選択した内容がこのタブに反映されることに注意してください。
6. 「オプションの外観」グループで、「色」ドロップダウン・リストから *gender* を選択します。
7. 「実行」をクリックします。

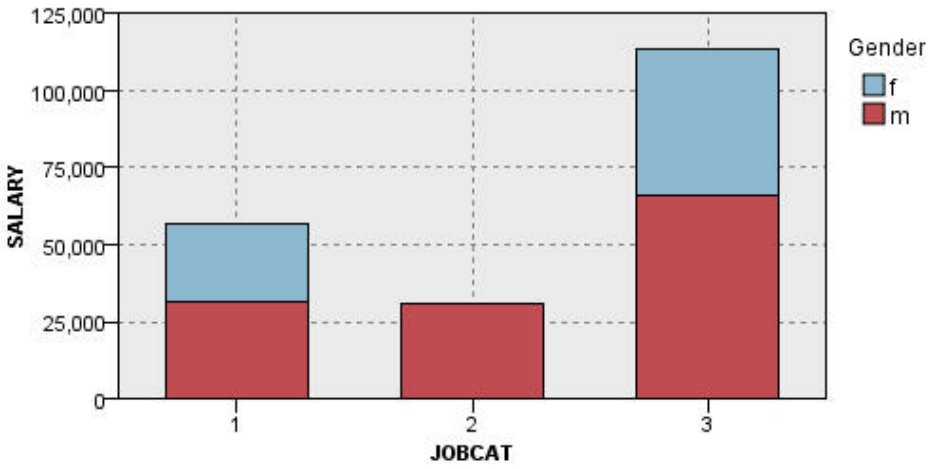


図 *9.* 積み上げ棒グラフ

以下のことがわかります。

- 各職種の平均給与の差は、全男性と全女性の平均給与を比較した棒グラフでの差と比べて大きいようには見えません。各グループに含まれる男性と女性の数が異なっている可能性があります。度数の棒グラフを作成すると、これを確認することができます。
- 職種にかかわらず、常に男性の平均給与が女性の平均給与よりも高くなっています。

例: パネル化されたヒストグラム

ここでは、性別ごとにパネル化されたヒストグラムを作成し、男性と女性の給与の度数分布を比較します。度数分布には、特定の給与範囲内に該当するケース/行の数が表示されます。パネル化されたヒストグラムにより、性別による給与の差を詳しく分析することができます。

注 : この例では *Employee data* を使用します。

1. グラフボード・ノードを追加して、編集用にそれを開きます。
2. 「基本」タブで *Current Salary* を選択します。
3. 「ヒストグラム」を選択します。
4. 「詳細」タブをクリックします。
5. 「パネルとアニメーション」グループで、「反対側にパネルを付ける」ドロップダウン・リストから「*gender*」を選択します。
6. 「実行」をクリックします。

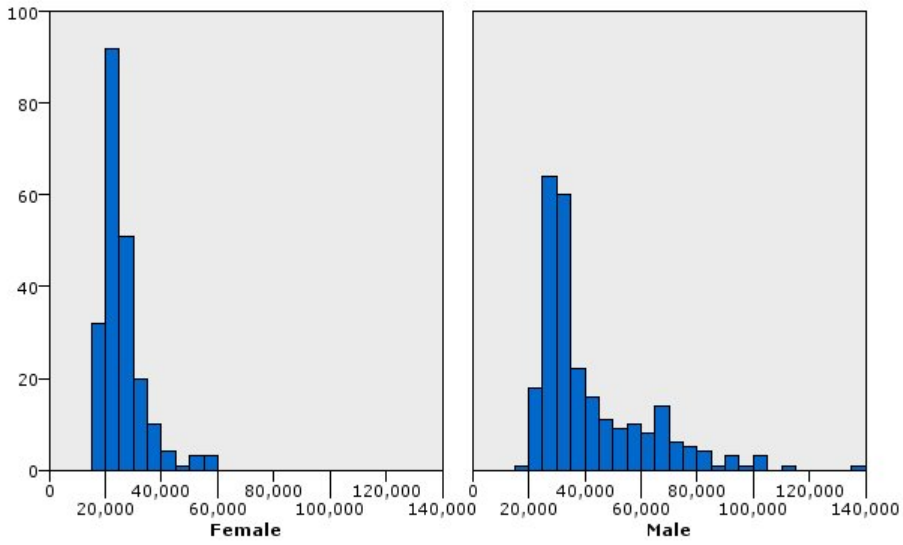


図 *10.* パネル化されたヒストグラム

以下のことがわかります。

- いずれの度数分布も正規分布ではありません。データが正規分布になる場合は釣鐘型の曲線になりますが、どちらのヒストグラムも釣鐘型の曲線のようにはなっていません。
- 高い棒ほど各グラフの左側に位置しています。したがって、男女とも、給与が低い人のほうが高い人より多くなっています。
- 給与の度数分布は、男性と女性で同じではありません。ヒストグラムの形状に注目してください。給与が高い男性が、給与が高い女性よりも多くなっています。

例: パネル化されたドット・プロット

ヒストグラムと同様に、ドット・プロットには連続した数値範囲の分布が表示されます。ただし、分割されたデータ範囲の度数を表示するヒストグラムとは異なり、ドット・プロットにはデータのすべての行/ケースが表示されます。そのため、ヒストグラムと比べてドット・プロットではより詳細に表示されます。実際に度数分布を分析する場合は、最初にドット・プロットを使用することをお勧めします。

注 : この例では *Employee data* を使用します。

1. グラフボード・ノードを追加して、編集用にそれを開きます。
2. 「基本」タブで *Current Salary* を選択します。
3. 「ドット プロット」を選択します。
4. 「詳細」タブをクリックします。
5. 「パネルとアニメーション」グループで、「反対側にパネルを付ける」ドロップダウン・リストから「*gender*」を選択します。
6. 「実行」をクリックします。
7. 結果出力ウィンドウを最大化すると、プロットが見やすくなります。

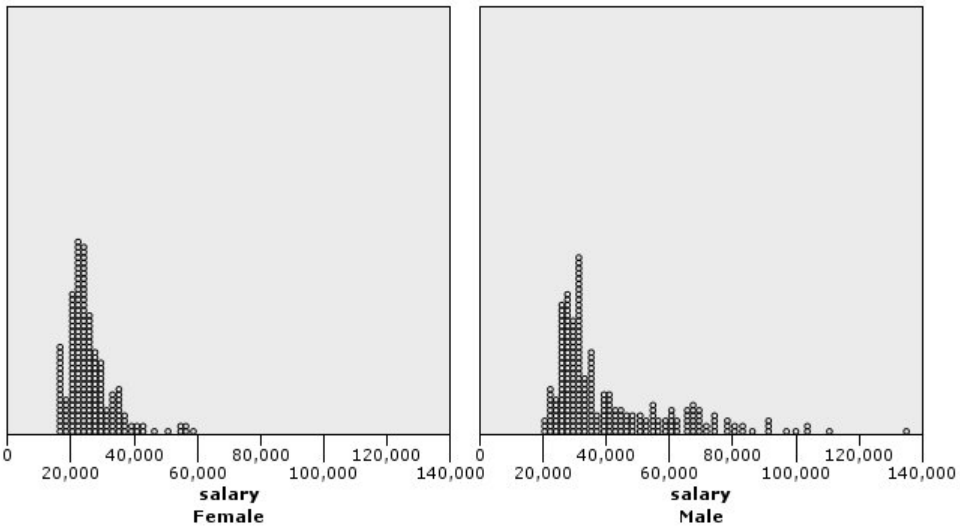


図 *11.* パネル化されたドット・プロット

ヒストグラム (220 ページの『例: パネル化されたヒストグラム』を参照) と比較すると、次のことが分かります。

- 女性のヒストグラムで 20,000 の位置に見られたピークは、ドット・プロットではそれほど顕著ではありません。その値の近くに多くのケース/行が集中していますが、それらの値のほとんどは 25,000 に近い値です。ヒストグラムでは、ここまで細かく分かりません。
- 男性のヒストグラムでは、男性の平均給与は 40,000 を超えると徐々に度数が減少していくように見えますが、ドット・プロットでは、40,000 を超えて 80,000 までは分布がほぼ一様になっています。その範囲の給与の値を任意に 1 つ選ぶと、その給与を得ている男性が 3 人以上存在します。

例: Boxplot

Boxplot も、データの分布を表示するために便利な視覚化です。Boxplot にはいくつかの統計指標が表示されます。これらの指標については、視覚化を作成した後に説明します。

注 : この例では *Employee data* を使用します。

1. グラフボード・ノードを追加して、編集用にそれを開きます。
2. 「基本」タブで、 *Gender* と *Current Salary* を選択します (複数のフィールド/変数を選択するには、Ctrl とクリックを使用します)。
3. 「**Boxplot**」を選択します。
4. 「実行」をクリックします。

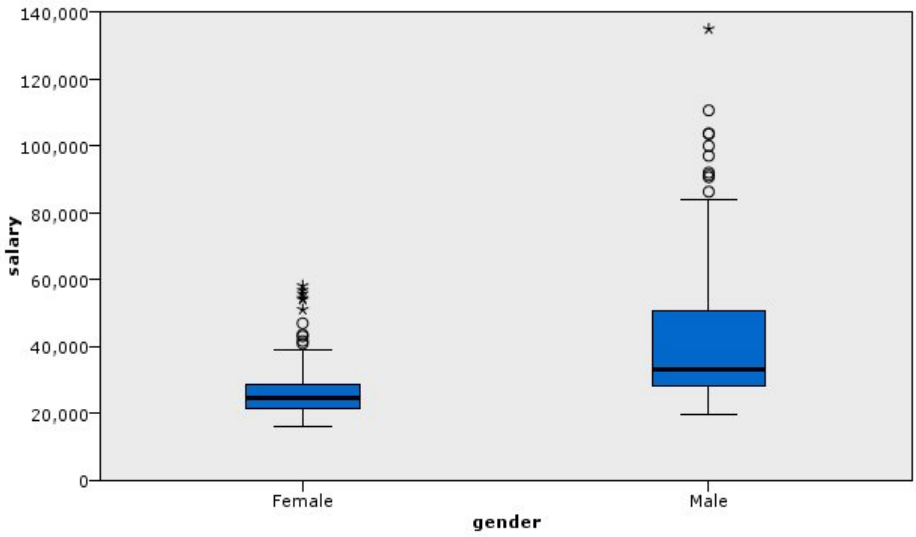


図 *12. Boxplot*

ここで、Boxplot の各部について説明します。

- 箱の中央にある濃い線は *salary* の中央値です。ケース/行の半数は中央値よりも大きな値を持ち、残りの半数は中央値よりも小さな値を持っています。平均値と同様に、中央値も中心傾向の指標です。平均値と異なるのは、極値を伴うケース/行による影響が小さいことです。この例では、中央値が平均値よりも低くなっています (218 ページの『例: 要約統計量を使用した棒グラフ』と比較してください)。平均値と中央値にこのような差が見られる場合は、平均値を押し上げる極値を持つケース/行が少ないことを示しています。つまり、給与が高い従業員の数が少ないということです。
- 箱の最下部は 25 パーセンタイルを示しています。25% のケース/行が、この 25 パーセンタイルよりも低い値を持っています。箱の最上部は 75 パーセンタイルを表しています。25% のケース/行が、この 75 パーセンタイルよりも高い値を持っています。つまり、50% のケース/行が箱の内側に入っています。女性の箱は、男性の箱よりもずっと短くなっています。このことから、女性の *salary* のばらつきが男性よりも小さいことが分かります。多くの場合、箱の上端と下端をヒンジと呼びます。
- 箱から延びている T 形の棒を内堀またはひげと呼びます。これらは、箱の高さの 1.5 倍 (その範囲内の値を持つケース/行がない場合は、最小値または最大値) まで延びています。データが正規分布に従う場合は、データの約 95% が内堀の間に入ると予測されます。この例では、女性の内堀の範囲が男性と比べて狭くなっており、このことからも女性の *salary* のばらつきが男性よりも小さいことが分かります。
- 図中の各点は外れ値です。この値は、内堀の間に入らない値として定義されます。外れ値とは、極端な値のことです。アスタリスク (星形) は、外れ値の中でも極端な値です。これは、箱の高さの 3 倍を超える値を持つケース/行を表しています。女性と男性の両方に、いくつかの外れ値があります。ここで、平均値が中央値よりも大きかったことを思い出してください。平均値が中央値よりも大きくなったのは、これらの外れ値が原因です。

例: 円グラフ

ここでは、別のデータ・セットを使用して他の種類の視覚化を説明します。使用するデータ・セットは *customer_subset* です。このデータ・セットは、顧客に関する情報を持つ仮定のデータ・ファイルです。

最初に、円グラフを作成して各地域の顧客の比率を調べます。

1. *customer_subset.sav* を指す Statistics ファイル・ソース・ノードを追加します。

2. グラフボード・ノードを追加して、編集用にそれを開きます。
3. 「基本」タブで *Geographic indicator* を選択します。
4. 「度数の円」を選択します。
5. 「実行」をクリックします。

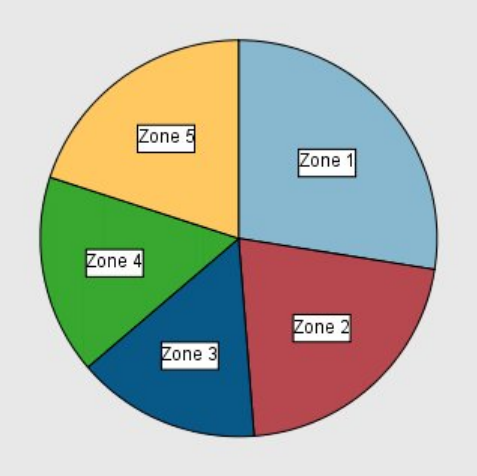


図 *13.* 円グラフ

以下のことがわかります。

- ゾーン 1 には、他のいずれのゾーンよりも多くの顧客が存在します。
- 他のゾーンでは顧客が均等に分布しています。

例: ヒート・マップ

ここでは、カテゴリー別のヒート・マップを作成して、それぞれの地域グループと年齢グループの顧客の平均所得を調べます。

注 : この例では *customer_subset* を使用します。

1. グラフボード・ノードを追加して、編集用にそれを開きます。
2. 「基本」タブで *Geographic indicator*、*Age category*、および *Household income in thousands* をこの順序で選択します (複数のフィールド/変数を選択するには、Ctrl とクリックを使用します)。
3. 「ヒート マップ」を選択します。
4. 「実行」をクリックします。
5. 結果出力ウィンドウで、「フィールドと値ラベルを表示」ツールバー・ボタン (ツールバーの中央にある 2 個のうちの右側) をクリックします。

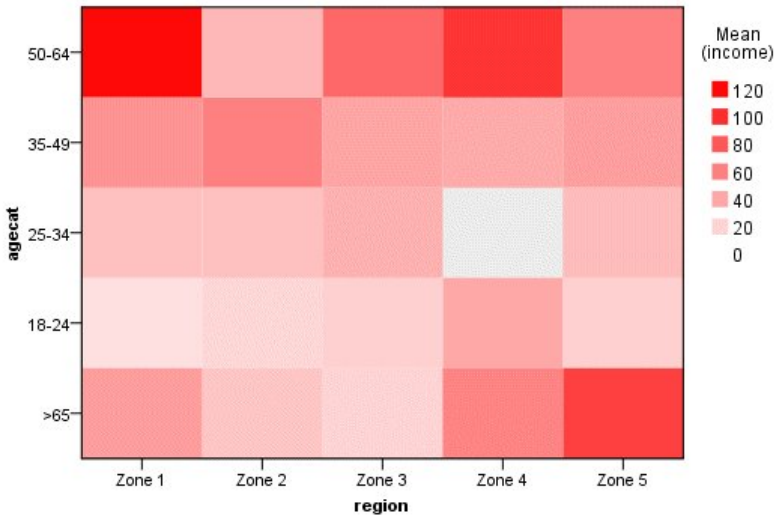


図 *14.* カテゴリー別のヒート・マップ

以下のことがわかります。

- ヒート・マップは表に似ており、数値の代わりに色を使用して各セルの値を表します。明るく濃い赤が最も高い値を示し、灰色が低い値を示します。各セルの値は、カテゴリーの各ペアの連続型フィールド/変数の平均値です。
- ゾーン 2 とゾーン 5 を除き、年齢が 50 歳から 64 歳までの顧客のグループは、平均世帯所得が他のグループよりも高くなっています。
- 年齢が 25 歳から 34 歳までの顧客は、ゾーン 4 には存在しません。

例: 散布図行列 (SPLOM)

ここでは、いくつかの異なる変数の散布図行列を作成して、データ・セットの変数間に関係があるかどうかを調べてみます。

注 : この例では *customer_subset* を使用します。

1. グラフボード・ノードを追加して、編集用にそれを開きます。
2. 「基本」タブで、 *Age in years*、*Household income in thousands*、および *Credit card debt in thousands* を選択します (複数のフィールド/変数を選択するには、Ctrl とクリックを使用します)。
3. 「**SPLOM**」を選択します。
4. 「実行」をクリックします。
5. 出力ウィンドウを最大化すると、行列が見やすくなります。

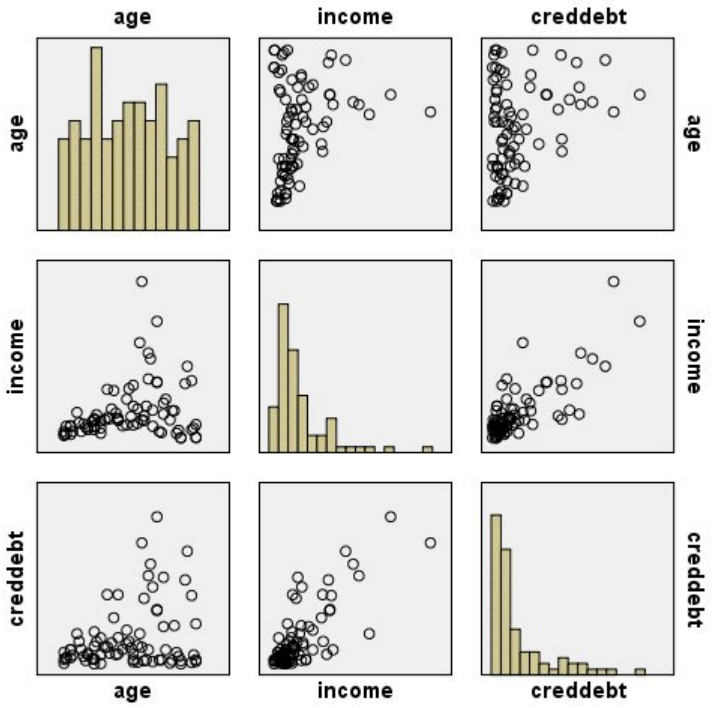


図 *15.* 散布図行列 *(SPLOM)*

以下のことがわかります。

- 対角線上に表示されているヒストグラムは、SPLOM の各変数の分布を示しています。*age* のヒストグラムは左上のセルに、*income* のヒストグラムは中央のセルに、*creddebt* のヒストグラムは右下のセルに表示されています。いずれの変数も正規分布に従っているようには見えません。つまり、いずれのヒストグラムも釣鐘型の曲線のようにはなっていません。また、*income* と *creddebt* のヒストグラムが正の方向にゆがんでいることにも注意してください。
- *age* と他の変数の間には何の関係も見られません。
- *income* と *creddebt* の間には線型の関係があります。つまり、*income* が増加するに従って *creddebt* も増加します。これらの変数と、関連する他の変数の個別の散布図を作成すると、関係を詳しく検討することができます。

例: 合計のコロプレス (カラー・マップ)

ここでは、マップの視覚化を作成します。次に、この視覚化のバリエーションの作成例を見てみます。使用するデータ・セットは *worldsales* です。このデータ・セットは、大陸と製品ごとの販売収入データを持つ仮定のデータ・ファイルです。

1. グラフボード・ノードを追加して、編集用にそれを開きます。
2. 「基本」タブで *Continent* と *Revenue* を選択します (複数のフィールド/変数を選択するには、Ctrl とクリックを使用します)。
3. 「合計のコロプレス **(Choropleth of Sums)**」を選択します。
4. 「詳細」タブをクリックします。
5. 「オプションの外観」グループで、「データ・ラベル」ドロップダウン・リストから「*Continent*」を選択します。
6. 「マップ・ファイル」グループで「マップ・ファイルを選択」をクリックします。

7. 「マップの選択」ダイアログ・ボックスで、「マップ」が「*Continents*」に設定され、「マップ キー」が「*CONTINENT*」に設定されていることを確認します。
8. 「マップ値とデータ値を比較」グループで「比較」をクリックし、マップ キーをデータ キーに一致させます。この例では、すべてのデータ キーの値が、対応するマップ キーとフィーチャーを持っています。オセアニアについてはデータがないことも分かります。
9. 「マップの選択」ダイアログ・ボックスで「**OK**」をクリックします。
10. 「実行」をクリックします。

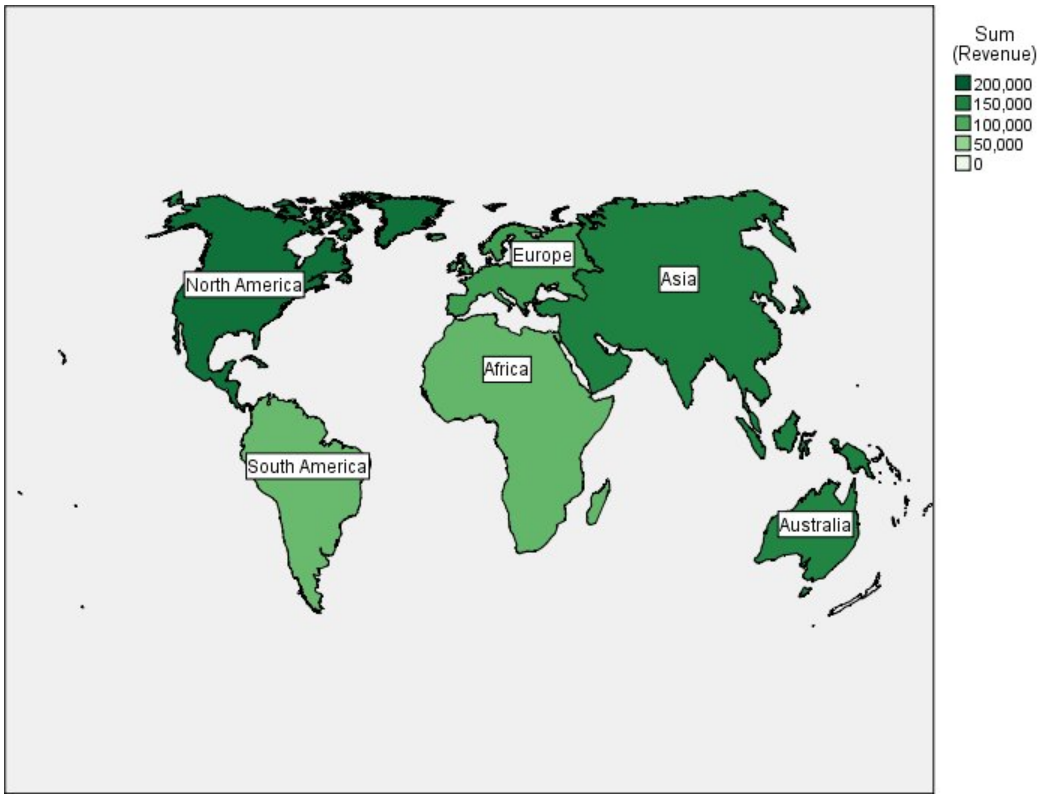


図 *16.* 合計のコロプレス

このマップ視覚化から、収入は北アメリカで最も高く、南アメリカとアフリカで最も低いことが容易に分かります。データ・ラベルの外観に「*Continent*」を使用したため、各大陸にラベルが付いています。

例: マップ上の棒グラフ

この例では、各大陸における製品別の収入の内訳を表示します。

注 : この例では *worldsales* を使用します。

1. グラフボード・ノードを追加して、編集用にそれを開きます。
2. 「基本」タブで、「*Continent*」、「*Product*」、および「*Revenue*」を選択します (複数のフィールド/変数を選択するには、Ctrl とクリックを使用します)。
3. 「マップの棒グラフ」を選択します。
4. 「詳細」タブをクリックします。

   特定のタイプのフィールドを複数使用する場合は、各フィールドが正しいスロットに割り当てられていることを確認してください。

5. 「カテゴリー」ドロップダウン・リストから *Product* を選択します。
6. 「値」ドロップダウン・リストから「*Revenue*」を選択します。
7. 「データ キー」ドロップダウン・リストから「*Continent*」を選択します。
8. 「要約」ドロップダウン・リストから「合計」を選択します。
9. 「マップ・ファイル」グループで「マップ・ファイルを選択」をクリックします。
10. 「マップの選択」ダイアログ・ボックスで、「マップ」が「*Continents*」に設定され、「マップ キー」が「*CONTINENT*」に設定されていることを確認します。
11. 「マップ値とデータ値を比較」グループで「比較」をクリックし、マップ キーをデータ キーに一致させます。この例では、すべてのデータ キーの値が、対応するマップ キーとフィーチャーを持っています。オセアニアについてはデータがないことも分かります。
12. 「マップの選択」ダイアログ・ボックスで「**OK**」をクリックします。
13. 「実行」をクリックします。
14. 結果出力ウィンドウを最大化すると、表示が見やすくなります。

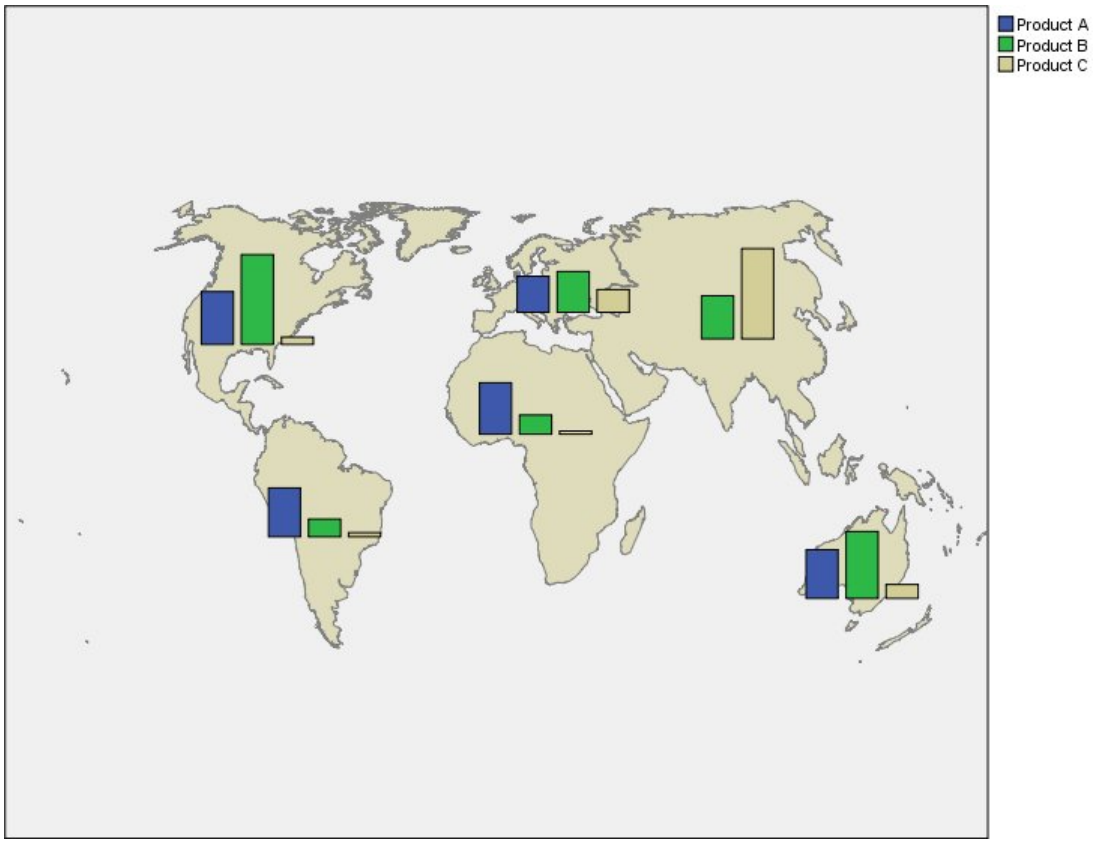


図 *17.* マップ上の棒グラフ

以下のことがわかります。

- 南アメリカとアフリカでは、製品全体にわたる総収入の分布が非常に類似しています。
- *Product C* から得られた収入は、アジア以外ではどの地域でも最も低くなっています。
- *Product A* から得られた収入は、アジアではまったくないかごくわずかな収入になっています。

グラフボードの「外観」タブ

グラフ作成前に外観オプションを指定できます。

一般的な外観オプション

表題: グラフのタイトルに使用するテキストを入力します。

サブタイトル: グラフのサブタイトルに使用するテキストを入力します。

キャプション: グラフのキャプションに使用するテキストを入力します。

「サンプリング」。大規模データ・セット用の手法を指定します。最大データ・セット・サイズを使用するか、またはデフォルトのレコード数を使用することができます。「サンプル」 を選択すると、大きいデータ・セットに対するパフォーマンスが向上します。代わりに、「すべてのデータを使用」 を選択して、すべてのデータ・ポイントを作図することもできます。ただし、この場合ソフトウェアのパフォーマンスが大幅に低下する可能性があります。

スタイル・シートの外観オプション

どの視覚化テンプレート (およびスタイル・シートとマップ) を使用するかを制御するための 2 つのボタンが用意されています。

「管理」。コンピューターで視覚化テンプレート、スタイル・シート、およびマップを管理します。視覚化テンプレート、スタイル・シート、およびマップをローカル・マシンでインポート、エクスポート、名前変更、および削除できます。詳しくは、トピック 231 ページの『テンプレート、スタイル・シート、マップ・ファイルの管理』を参照してください。

「場所」。視覚化テンプレート、スタイル・シート、およびマップが保管されている場所を変更します。現在の場所は、ボタンの右側に表示されます。 詳しくは、トピック 230 ページの『テンプレート、スタイル・シート、マップの位置の設定』を参照してください。

次の例では、表示オプションがグラフ内のどこに表示されるかを示します (注 : すべてのグラフですべてのオプションが使用されるわけではありません)。

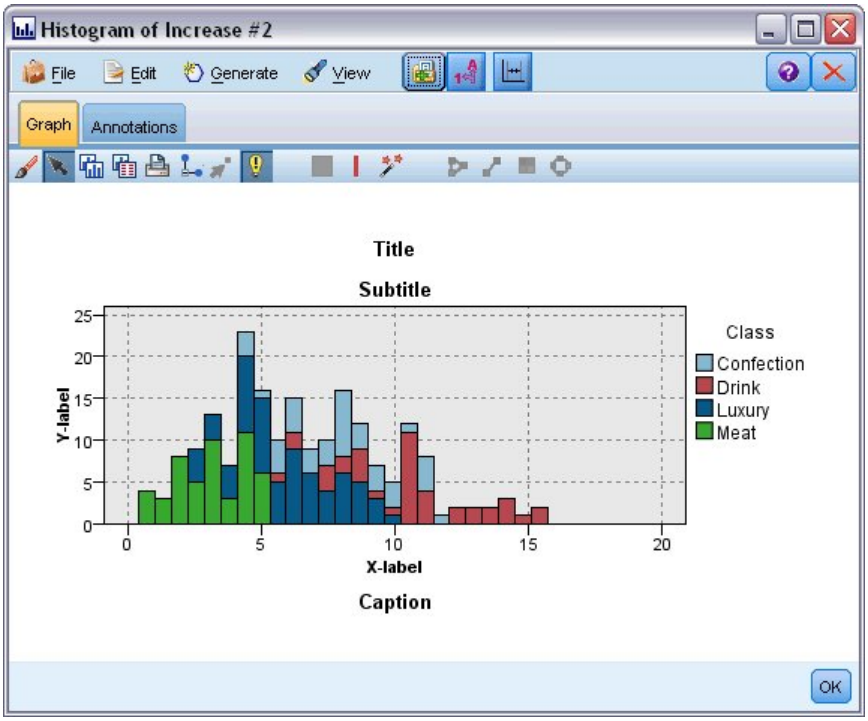


図 *18.* さまざまなグラフ表示オプションの場所

テンプレート、スタイル・シート、マップの位置の設定

視覚化テンプレート、視覚化タイルシート、マップ・ファイルは、特定のローカル・フォルダーまたは IBM SPSS Collaboration and Deployment Services Repository に保存されています。テンプレート、スタイル・シート、マップを選択する場合、この場所に組み込まれたものだけが表示されます。すべてのテンプレート、スタイル・シート、マップ・ファイルを 1 つの場所に保存することによって、IBM SPSS アプリケーションはそれらに容易にアクセスできるようになります。この場所にテンプレート、スタイル・シート、マップ・ファイルをさらに追加する方法については、 231 ページの『テンプレート、スタイル・シート、マップ・ファイルの管理』を参照してください。

テンプレート、スタイル・シート、マップ・ファイルの位置の設定するには

1. テンプレート、スタイル・シート、マップ・ファイルのダイアログボックスで、「場所...」 をクリックして「テンプレートとスタイル・シート」ダイアログ・ボックスが表示されます。
2. テンプレート、スタイル・シート、マップ・ファイルのデフォルトの場所について、オプションを選択します。

   「ローカル・コンピューター」。テンプレート、スタイル・シート、およびマップ・ファイルは、ローカル・コンピューター上の特定のフォルダーに配置されます。 Windows XP の場合、このフォルダーは *C:¥Documents and Settings¥<user>¥Application Data¥SPSSInc¥Graphboard* となります。フォルダーは変更できません。

   IBM SPSS Collaboration and Deployment Services Repository. テンプレート、スタイル・シート、マップ・ファイルは、IBM SPSS Collaboration and Deployment Services Repository のユーザー指定のフォルダーにあります。特定のフォルダーを指定するには、「フォルダー」 を選択します。

詳しくは、『テンプレート、スタイル・シート、マップ・ファイルの場所として IBM SPSS Collaboration and Deployment Services Repository を使用』を参照してください。

3. 「**OK**」をクリックします。

テンプレート、スタイル・シート、マップ・ファイルの場所として IBM SPSS Collaboration and Deployment Services Repository を使用

視覚化テンプレートおよびスタイル・シートを IBM SPSS Collaboration and Deployment Services Repository に保存できます。この場所は、IBM SPSS Collaboration and Deployment Services Repository の固有のフォルダーです。これがデフォルトの場所として設定されている場合、この場所のテンプレート、スタイル・シート、マップ・ファイルを選択できます。

IBM SPSS Collaboration and Deployment Services Repository のフォルダーをテンプレート、スタイル・シート、マップ・ファイルとして設定するには

1. 「場所」ボタンのあるダイアログ・ボックスで、「場所**...**」 をクリックします。
2. IBM SPSS Collaboration and Deployment Services Repository を選択します。
3. 「フォルダー」をクリックします。

   注：IBM SPSS Collaboration and Deployment Services Repository に接続していない場合、接続情報を求めるメッセージが表示されます。
4. 「フォルダーの選択」ダイアログ・ボックスで、テンプレート、スタイル・シート、マップ・ファイルが保存されるフォルダーを選択します。
5. 必要に応じて、「ラベルの取得」でラベルを選択することができます。ラベルの付いたテンプレート、スタイル・シート、マップ・ファイルのみが表示されます。
6. 特定のテンプレート、スタイル・シート、マップ・ファイルを含むフォルダーを検索する場合、「検索」タブでテンプレート、スタイル・シート、マップ・ファイルを検索できます。 [フォルダーの選択」ダイアログ・ボックスでは、検索されたテンプレート、スタイル・シート、マップ・ファイルが保存されているフォルダーを自動的に選択します。
7. 「フォルダーの選択」 をクリックします。

テンプレート、スタイル・シート、マップ・ファイルの管理

使用しているコンピューターにローカルに保存されているテンプレート、スタイル・シート、マップ・ファイルを管理するには、「テンプレート、スタイル・シート、およびマップの管理 (Manage Templates, Stylesheets, and Maps)」ダイアログ・ボックスを使用します。このダイアログ・ボックスでは、使用しているコンピューターにローカルに保存されている視覚化テンプレート、スタイル・シート、マップ・ファイルについて、インポート、エクスポート、名前の変更、削除を行うことができます。

テンプレート、スタイル・シート、またはマップを選択するためのいずれかのダイアログ・ボックスで「管理**...**」をクリックします。

「テンプレート、スタイル・シート、およびマップの管理 (Manage Templates, Stylesheets, and Maps)」ダイアログ・ボックス

「テンプレート」タブには、すべてのローカル・テンプレートがリストされます。「スタイル・シート」タブには、すべてのローカル・スタイル・シートがリストされ、サンプル・データによる視覚化の例が表示されます。いずれかのスタイル・シートを選択して、そのスタイルを視覚化の例に適用することができます。詳しくは、トピック 315 ページの『スタイル・シートの適用』を参照してください。「マップ」タブに

は、すべてのローカル・マップ・ファイルがリストされます。このタブには、サンプルの値を持つマップキー、コメント (マップの作成時にコメントを指定した場合)、マップのプレビューも表示されます。

以下のボタンは、現在選択されているタブ上で機能します。

インポート: 視覚化テンプレート、スタイル・シート、またはマップ・ファイルをファイル・システムからインポートします。テンプレート、スタイル・シート、またはマップ・ファイルをインポートすると、IBM SPSS アプリケーションで使用できるようになります。テンプレート、スタイル・シート、またはマップ・ファイルが別のユーザーから送信された場合は、アプリケーションで使用する前にそのファイルをインポートしてください。

エクスポート: 視覚化テンプレート、スタイル・シート、またはマップ・ファイルをファイル・システムにエクスポートします。テンプレート、スタイル・シート、またはマップ・ファイルを別のユーザーに送信する場合は、それをエクスポートする必要があります。

名前変更: 選択された視覚化テンプレート、スタイル・シート、またはマップ・ファイルの名前を変更します。既に使用されている名前に変更することはできません。

マップ キーのエクスポート **(Export Map Key)**: マップ キーをカンマ区切り値 (CSV) ファイルとしてエクスポートします。このボタンは「マップ」タブのみで使用することができます。

削除: 選択された 1 つ以上の視覚化テンプレート、スタイル・シート、またはマップ・ファイルを削除します。複数のテンプレート、スタイル・シート、またはマップ・ファイルを選択するには、Ctrl キーを押しながらクリックします。削除操作は元に戻すことができないため、慎重に行ってください。

マップ・シェープファイルの変換と配布

グラフボード・テンプレート選択により、視覚化テンプレートと SMZ ファイルの組み合わせからマップ視覚化を作成することができます。SMZ ファイルは、マップを描画するための地理情報 (国境線など) を持つという点において ESRI シェープファイル (SHP ファイル形式) と似ていますが、マップ視覚化用に最適化されているという点が異なっています。グラフボード・テンプレート選択は、厳選された SMZ ファイルとともに事前にインストールされています。マップ視覚化で既存の ESRI シェープファイルを使用する場合は、最初にマップ変換ユーティリティーを使用してそのシェープファイルを SMZ ファイルに変換する必要があります。マップ変換ユーティリティーは、単一のレイヤーを含むポイント、折れ線、またはポリゴン (形状タイプ 1、3、5) の ESRI シェープファイルをサポートしています。

マップ変換ユーティリティーを使用すると、ESRI シェープファイルを変換するだけでなく、マップの詳細度の変更、フィーチャー・ラベルの変更、フィーチャーの結合、フィーチャーの移動など、多くの変更操作を必要に応じて実行することができます。また、マップ変換ユーティリティーを使用して、既存の SMZ ファイル (事前にインストールされているファイルを含む) を変更することもできます。

インストール済み SMZ ファイルの編集

1. 管理システムから SMZ ファイルをエクスポートします。詳しくは、トピック 231 ページの『テンプレート、スタイル・シート、マップ・ファイルの管理』を参照してください。
2. マップ変換ユーティリティーを使用して、エクスポートした SMZ ファイルを開いて編集します。このファイルは、別の名前を付けて保存することをお勧めします。詳しくは、トピック 233 ページの『マップ変換ユーティリティーの使用』を参照してください。
3. 変更した SMZ ファイルを管理システムにインポートします。詳しくは、トピック 231 ページの『テンプレート、スタイル・シート、マップ・ファイルの管理』を参照してください。

マップ・ファイルの追加リソース

マップ処理のニーズをサポートする SHP ファイル形式の地理空間データは、さまざまな非公開ソースや公開ソースから入手することができます。無料のデータを探す場合は、自治体の Web サイトを確認してください。本製品の多くのテンプレートは、GeoCommons ( ) とアメリカ・センサス局 (http://www.census.gov) から入手した、一般に公開されているデータに基づいています。

重要: IBM 以外の製品に関する情報は、その製品の供給者、出版物、もしくはその他の公に利用可能なソースから入手したものです。IBM は、それらの製品のテストは行っておりません。したがって、他社製品に関する実行性、互換性、またはその他の要求については確証できません。 IBM 以外の製品の性能に関する質問は、それらの製品の供給者にお願いします。本書において IBM 以外の Web サイトに言及している場合がありますが、便宜のため記載しただけであり、決してそれらの Web サイトを推奨するものではありません。それらの Web サイトにある資料は、この IBM プログラムに付属する特記事項ファイルで明記されていない限り、この IBM プログラムの資料の一部ではありません。それらの資料サイトは、お客様の責任でご使用ください。

マップの主要な概念

シェープファイルに関する主要な概念を理解すると、マップ変換ユーティリティーを効果的に使用するのに役立ちます。

シェープファイルは、マップを描画するための地理情報を提供します。マップ変換ユーティリティーがサポートするシェープファイルには、以下の 3 種類があります。

- ポイント: このシェープファイルは、地点 (都市など) の位置を示します。
- 折れ線: このシェープファイルは、経路とその位置 (川など) を示します。
- ポリゴン: このシェープファイルは、境界で囲まれた領域とその位置 (国など) を示します。

最も頻繁に使用するのは、ポリゴン・シェープファイルです。コロプレス・マップは、ポリゴン・シェープファイルから作成されます。コロプレス・マップは、色を使用して個々の多角形 (領域) 内の値を表します。ポイントのシェープファイルと折れ線のシェープファイルは、通常、ポリゴン・シェープファイルの上に重ねられます。例えば、米国の都市のポイント・シェープファイルは、米国の州のポリゴン・シェープファイルの上に重ねられます。

シェープファイルはフィーチャーから構成されます。フィーチャーは、個々の地理的エンティティーです。例えば、国、都道府県、市区町村などがフィーチャーに該当します。シェープファイルには、フィーチャーに関するデータも格納されます。これらのデータは属性に格納されます。属性は、データ・ファイルのフィールドや変数に類似しています。フィーチャーには、マップ キーである属性が 1 つ以上存在します。マップ キーは、ラベル (国や都道府県の名前など) にすることができます。マップ キーをデータ・ファイルの変数/フィールドにリンクすると、マップ視覚化が作成されます。

SMZ ファイルに保存できるのはキー属性だけであることに注意してください。マップ変換ユーティリティーは、これ以外の属性の保存をサポートしていません。そのため、異なるレベルで集計する場合は、複数の SMZ ファイルを作成する必要があります。例えば、米国の州と地域を集計する場合は、個別の SMZ ファイル (州を識別するキーを持つファイルと、地域を識別するキーを持つファイル) が必要になります。

マップ変換ユーティリティーの使用

マップ変換ユーティリティーの開始方法

メニューから次の項目を選択します。

「ツール」 > 「マップ変換ユーティリティー」

マップ変換ユーティリティーには 4 つの主要な画面 (ステップ) があります。そのうちの 1 ステップは複数のサブステップから構成されており、マップ・ファイルの編集方法を詳細に制御することができます。

ステップ 1 - 宛先ファイルとソース・ファイルの選択

最初に、ソース・マップ・ファイルと、変換後のマップ・ファイルの宛先を選択する必要があります。シェープファイルの場合は、*.shp* ファイルと *.dbf* ファイルの両方が必要です。

変換する **.shp (ESRI)** ファイルまたは **.smz** ファイルを選択する **(Select a .shp (ESRI) or .smz file for conversion):** 使用しているコンピューター上の既存のマップ・ファイルを参照します。これは、SMZ ファイルに変換して保存するファイルです。シェープファイルの *.dbf* ファイルは、*.shp* ファイルと同じベース・ファイル名で同じ場所に格納する必要があります。*.dbf* ファイルには *.shp* ファイルの属性情報が含まれているため、このファイルは必須です。

変換されるマップ・ファイルの保存先およびファイル名を設定する **(Set a destination and file name for the converted map file):** 元のマップ・ソースから作成される SMZ ファイルのパスとファイル名を入力します。

- テンプレート選択にインポートする **(Import to the Template Chooser):** ファイル・システム上のファイルに保存するだけでなく、必要に応じてテンプレート選択の管理リストにマップを追加することもできます。このオプションを選択すると、使用しているコンピューターにインストールされている IBM SPSS 製品のテンプレート選択で、マップが自動的に使用可能になります。ここでテンプレート選択にインポートしなかった場合は、後から手動でインポートする必要があります。テンプレート選択の管理システムにマップをインポートする方法については、 231 ページの『テンプレート、スタイル・シート、マップ・ファイルの管理』を参照してください。

ステップ 2 - マップ キーの選択

ここでは、SMZ ファイルに含めるマップ キーを選択します。マップ キーを選択したら、マップの表示に影響する一部のオプションを変更することができます。マップ変換ユーティリティーの後続のステップでは、マップのプレビューが表示されます。選択された表示オプションを使用して、マップのプレビューが生成されます。

プライマリー・マップ キーを選択する **(Choose primary map key):** マップ上のフィーチャーの識別とラベル付けを行うためのプライマリー・キーとなる属性を選択します。例えば、世界地図のプライマリー・キーは、国名を識別する属性にすることができます。プライマリー・キーは、データをマップ上のフィーチャーにリンクする役割も果たすため、選択する属性の値 (ラベル) は、データの値に一致している必要があります。属性を選択すると、サンプルのラベルが表示されます。これらのラベルを変更する必要がある場合は、これ以降のステップで変更することができます。

追加のキーを選択する **(Choose additional keys to include):** プライマリー・マップ キーのほかに、生成する SMZ ファイルに追加したい他のキー属性があれば、そのキー属性にチェック・マークを付けてください。例えば、一部の属性については、ラベルが翻訳されていることがあります。データを他の言語でコード化する場合は、それらの属性を保存することをお勧めします。ただし、追加のキーとして選択できるのは、プライマリー・キーと同じフィーチャーを表すキーだけです。例えば、プライマリー・キーが米国の州の省略されていない名前である場合は、米国の州を表す代替キー (州の省略名など) のみ選択することができます。

自動的にマップを平滑化する **(Automatically smooth the map):** 多角形を含むシェープファイルは、通常、統計分野でのマップ視覚化で使用するにはデータ・ポイントが多すぎ、情報も詳しすぎます。過剰な詳

細情報は混乱のもとであり、パフォーマンスに悪影響が及ぶ可能性もあります。平滑化によって詳細度を低下させ、マップのおおまかな特徴だけを残すことができます。その結果、マップの外観が整理され、表示が高速になります。マップを自動的に平滑化する場合の最大角度は 15 度で、保持率は 99 パーセントです。これらの設定について詳しくは、『マップの平滑化』を参照してください。別のステップで、後から追加の平滑化を適用することができます。

同じフィーチャーの接触しているポリゴン間の境界を削除する **(Remove boundaries between touching polygons in the same feature):** 一部のフィーチャーには、そのメイン・フィーチャーの内部に境界線を持つサブフィーチャーが含まれている場合があります。例えば、大陸の世界地図の内側に、各大陸内に存在する国の境界線が含まれている場合があります。このオプションを選択すると、内側の境界線がマップに表示されなくなります。大陸の世界地図を例にとると、このオプションを指定した場合、国境線が削除され、大陸の境界線は残ったままになります。

ステップ 3 - マップの編集

ここまでの操作で、マップの基本的なオプションが指定されました。ここからは、さらに詳細なオプションを編集することができます。ここで説明する変更操作は、オプションです。マップ変換ユーティリティーのこのステップでは、関連するタスクをガイドに従って実行し、変更内容を確認できるようマップのプレビューを表示します。シェープファイルの種類 (ポイント、折れ線、またはポリゴン) や座標系によっては、一部のタスクを実行できない場合があります。

すべてのタスクについて、マップ変換ユーティリティーの左側に、以下に示す共通のコントロールが用意されています。

マップにラベルを表示する **(Show the labels on the map):** デフォルトでは、フィーチャー・ラベルはプレビューには表示されません。ラベルを表示するかどうかを選択することができます。ラベルはフィーチャーを識別する場合に便利ですが、プレビュー・マップで直接選択する場合の妨げになることがあります。このオプションは、必要に応じてオンにしてください (フィーチャー・ラベルを編集する場合など)。

マップ・プレビューの色を指定する **(Color the map preview):** デフォルトでは、プレビュー・マップの領域は濃淡のない色で表示されます。すべてのフィーチャーが同じ色で表示されます。個々のマップ・フィーチャーに対して、各種の色を割り当てることができます。このオプションにより、マップ上で異なるフィーチャーを区別しやすくなります。フィーチャーを結合するときに、新しいフィーチャーがプレビューにどのように表示されるかを確認したい場合に特に便利です。

すべてのタスクについて、マップ変換ユーティリティーの右側に、以下に示す共通のコントロールが用意されています。

元に戻す**:** 「元に戻す」をクリックすると、直前の状態に戻ります。最大 100 件の変更を元に戻すことができます。

マップの平滑化**:** 多角形を含むシェープファイルは、通常、統計分野でのマップ視覚化で使用するにはデータ・ポイントが多すぎ、情報も詳しすぎます。過剰な詳細情報は混乱のもとであり、パフォーマンスに悪影響が及ぶ可能性もあります。平滑化によって詳細度を低下させ、マップのおおまかな特徴だけを残すことができます。その結果、マップの外観が整理され、表示が高速になります。ポイント・マップと折れ線マップの場合は、このオプションを使用することはできません。

最大角 **(Max. angle):** 最大角は、ほぼ一直線に並ぶ一連の点を平滑化する場合の許容度を指定します。値は 1 から 20 までの範囲で指定する必要があります。この値が大きいほど直線平滑化の許容度が高くなり、多くの点が破棄されて、おおまかな特徴だけがマップに残るようになります。直線平滑化を適用するために、マップ変換ユーティリティーは、マップ上の点を 3 つずつ選択し、それらの点によって形成される

内角を検査します。180 からこの角度を引いた値が指定値よりも小さい場合、マップ変換ユーティリティーは中間の点を破棄します。つまり、マップ変換ユーティリティーは、3 つの点が形成する線がほぼ一直線に並んでいるかどうかを検査します。ほぼ一直線に並んでいる場合、マップ変換ユーティリティーはその線を、両端の点を結んで中間の点を破棄した直線として扱います。

保持率 **(Percent to keep):** 保持率により、マップを平滑化する際に保持する陸地面積の量を決定します。値は 90 から 100 までの範囲で指定する必要があります。このオプションは、フィーチャーに複数の島が含まれている場合など、複数の多角形を持つフィーチャーにのみ影響します。フィーチャーから多角形を引いた総面積が元の面積の指定された割合よりも大きい場合、マップ変換ユーティリティーはその多角形をマップから破棄します。マップ変換ユーティリティーによってフィーチャーのすべての多角形が削除されることはありません。つまり、適用する平滑化の量にかかわらず、フィーチャーの多角形は必ず 1 つ以上存在することになります。

最大角と保持率を選択したら、「適用」をクリックします。この操作によってプレビューが更新され、平滑化による変更が反映されます。再度マップを平滑化する必要がある場合は、必要なレベルの平滑化になるまでこの操作を繰り返します。ただし、平滑化には限界があります。平滑化を繰り返すと、マップをそれ以上平滑化できない状態になります。

フィーチャー・ラベルの編集**:** 必要に応じてフィーチャー・ラベルを編集するだけでなく (必要なデータに一致させる場合など)、マップでのラベルの位置を変更することもできます。ラベルを変更する必要はないと考えられる場合でも、マップから視覚化を作成する前に必ずラベルを確認してください。デフォルトでは、ラベルはプレビューには表示されないため、「マップにラベルを表示する **(Show the labels on the map)**」を選択してラベルを表示してください。

キー**:** 確認や編集を行うフィーチャー・ラベルを持つキーを選択します。

フィールド: このリストには、選択したキーに含まれているフィーチャー・ラベルが表示されます。ラベルを編集するには、リスト内のラベルをダブルクリックします。ラベルがマップ上に表示されている場合は、マップ・プレビューで直接フィーチャー・ラベルをダブルクリックすることもできます。ラベルを実際のデータ・ファイルと比較する場合は、「比較」をクリックします。

「**X**」**/**「**Y**」**:** これらのテキスト・ボックスには、マップ上で選択されたフィーチャー・ラベルの現在の中心点がリストされます。単位はマップの座標で表示されます。これらの単位は、ローカルのデカルト座標 (State Plane Coordinate System など) や地理座標 (**X** が経度で **Y** が緯度) になる場合があります。ラベルの新しい位置の座標を入力してください。ラベルが表示されている場合は、マップ上のラベルをクリックしてドラッグすることもできます。テキスト・ボックスが新しい位置に更新されます。

比較**:** 特定のキーのフィーチャー・ラベルと比較されるデータ値を持つデータ・ファイルがある場合は、「比較」をクリックすると、「外部データ・ソースとの比較 (Compare to an External Data Source)」ダイアログ・ボックスが表示されます。このダイアログ・ボックスでデータ・ファイルを開き、そのファイルの値をマップ キーのフィーチャー・ラベルと直接比較することができます。

「外部データ・ソースとの比較 ***(Compare to an External Data Source)***」ダイアログ・ボックス*:* 「外部データ・ソースとの比較 (Compare to an External Data Source)」ダイアログ・ボックスでは、タブ区切り値ファイル (拡張子は *.txt*)、カンマ区切り値ファイル (拡張子は *.csv*)、IBM SPSS Statistics 用に書式設定されたデータ・ファイル (拡張子は *.sav*) を開くことができます。ファイルを開いたら、データ・ファイルのフィールドを選択して、特定のマップ キーのフィーチャー・ラベルと比較することができます。この状態で、マップ・ファイルでの不一致を訂正することができます。

データ・ファイルのフィールド **(Fields in the data file):** フィーチャー・ラベルと比較したい値を持つフィールドを選択します。*.txt* ファイルまたは *.csv* ファイルの最初の行に各フィールドの説明用ラベルが含

まれている場合は、「最初の行を列ラベルとして使用 **(Use first row as column labels)**」にチェック・マークを付けてください。それ以外の場合は、各フィールドがデータ・ファイル内の位置によって識別されます (「Column 1」や「Column 2」など)。

比較するキー **(Key to compare):** データ・ファイルのフィールド値と比較したいフィーチャー・ラベルを持つマップ キーを選択します。

比較**:** 値を比較する準備が整ったら、このオプションをクリックします。

比較の結果 **(Comparison Results):** デフォルトでは、「比較の結果 (Comparison Results)」テーブルには、データ・ファイル内の一致しなかったフィールド値だけがリストされます。アプリケーションは、通常はスペースの有無を確認することにより、関連するフィーチャー・ラベルを探します。「マップ・ラベル *(Map Label)*」列のドロップダウン・リストをクリックして、マップ・ファイルのフィーチャー・ラベルを、表示されているフィールド値と突き合せます。対応するフィーチャー・ラベルがマップ・ファイルに存在しない場合は、「不一致のままにする *(Leave Unmatched)*」を選択してください。既にフィーチャー・ラベルに一致しているフィールド値を含め、すべてのフィールド値を表示する場合は、「不一致のケースのみを表示 **(Display only unmatched cases)**」をクリアしてください。この操作は、1 つ以上の一致項目をオーバーライドする場合に行います。

フィールド値にフィーチャーを一致させる場合、各フィーチャーを使用できるのは 1 回だけです。複数のフィーチャーを 1 つのフィールド値に一致させる場合は、それらのフィーチャーを結合してから、対象のフィールド値に一致させることができます。フィーチャーの結合について詳しくは、『フィーチャーの結合』を参照してください。

フィーチャーの結合**:** フィーチャーを結合すると、マップ上で大きな地域を作成する場合に便利です。例えば州のマップを作成する場合に、複数の州 (この例におけるフィーチャー) を、より大きな北部、南部、東部、西部の各地域に結合することができます。

キー**:** 結合するフィーチャーを識別するためのフィーチャー・ラベルを持つマップ キーを選択します。

フィールド: 結合する最初のフィーチャーをクリックします。次に、Ctrl キーを押しながら、結合したい他のフィーチャーをクリックします。フィーチャーはマップ・プレビューでも選択されることに注意してください。リストから選択するだけでなく、マップ・プレビューで直接フィーチャーをクリックすることも、Ctrl キーを押しながらクリックすることもできます。

結合するフィーチャーを選択してから「結合」をクリックすると、「結合したフィーチャー名の指定 (Name the Merged Feature)」ダイアログ・ボックスが表示され、新しいフィーチャーにラベルを適用することができます。フィーチャーを結合したら、「マップ・プレビューの色を指定する **(Color the map preview)**」にチェック・マークを付けて、予期したとおりの結果になっていることを確認してください。

フィーチャーを結合したら、新しいフィーチャーのラベルを移動することができます。この操作は、「フィーチャー・ラベルの編集 *(Edit the feature labels)*」タスクで実行することができます。詳しくは、トピック 236 ページの『フィーチャー・ラベルの編集』を参照してください。

「結合したフィーチャー名の指定 ***(Name the Merged Feature)***」ダイアログ・ボックス***:*** 「結合したフィーチャー名の指定 (Name the Merged Feature)」ダイアログ・ボックスでは、結合した新しいフィーチャーにラベルを割り当てることができます。

「ラベル」テーブルにはマップ・ファイルの各キーの情報が表示され、各キーにラベルを割り当てることができます。

新規ラベル**:** 特定のマップ キーに割り当てる、結合後のフィーチャーの新しいラベルを入力します。

キー**:** 新しいラベルの割り当て先となるマップ キー。

古いラベル **(Old Labels):** 新しいフィーチャーに結合されるフィーチャーのラベル。

接触するポリゴン間の境界を削除 **(Remove boundaries between touching polygons):** 結合されたフィーチャーから境界線を削除する場合は、このオプションにチェック・マークを付けます。例えば、複数の州をいくつかの地域に結合した場合、このオプションを選択すると個々の州を囲む境界線が削除されます。

フィーチャーの移動**:** マップ内のフィーチャーは移動することができます。この機能は、本土や離島などのフィーチャーを一緒に配置する場合に便利です。

キー**:** 移動するフィーチャーを識別するためのフィーチャー・ラベルを持つマップ キーを選択します。

フィールド: 移動するフィーチャーをクリックします。フィーチャーはマップ・プレビューでも選択されることに注意してください。マップ・プレビューで直接フィーチャーをクリックすることもできます。

「**X**」**/**「**Y**」**:** これらのテキスト・ボックスには、マップ上のフィーチャーの現在の中心点がリストされます。単位はマップの座標で表示されます。これらの単位は、ローカルのデカルト座標 (State Plane Coordinate System など) や地理座標 (**X** が経度で **Y** が緯度) になる場合があります。フィーチャーの新しい位置の座標を入力してください。マップ上のフィーチャーをクリックしてドラッグすることもできます。テキスト・ボックスが新しい位置に更新されます。

フィーチャーの削除**:** 不要なフィーチャーをマップから削除することができます。マップ視覚化で重要ではないフィーチャーを削除することで、表示を見やすくすることができます。

キー**:** 削除するフィーチャーを識別するためのフィーチャー・ラベルを持つマップ キーを選択します。

フィールド: 削除したいフィーチャーをクリックします。複数のフィーチャーを同時に削除する場合は、Ctrl キーを押しながらフィーチャーをクリックします。フィーチャーはマップ・プレビューでも選択されることに注意してください。リストから選択するだけでなく、マップ・プレビューで直接フィーチャーをクリックすることも、Ctrl キーを押しながらクリックすることもできます。

個々の要素の削除**:** フィーチャー全体を削除するだけでなく、湖や小島など、フィーチャーを構成する個々の要素の一部だけを削除することもできます。ポイント・マップの場合、このオプションは使用できません。

要素**:** 削除したい要素をクリックします。複数の要素を同時に削除する場合は、Ctrl キーを押しながら要素をクリックします。要素はマップ・プレビューでも選択されることに注意してください。リストから選択するだけでなく、マップ・プレビューで直接要素をクリックすることも、Ctrl キーを押しながらクリックすることもできます。要素名のリストには説明が記載されないため (フィーチャー内の各要素には番号が割り当てられます)、マップ・プレビューで選択内容を表示して、必要な要素が選択されていることを確認してください。

投影法の設定**:**

マップの投影法により、3 次元の地球を 2 次元で表現するための方法を指定します。いずれの投影法でもゆがみが発生します。ただし、世界地図を表示するのに適した投影法もあれば、より狭い範囲のマップを表示するのに適した投影法もあります。また、元のフィーチャーの形状が保持される投影法もあります。形状が保持される投影法は、正角図法です。このオプションは、地理座標 (経度と緯度) を使用するマップの場合のみ使用することができます。

マップ変換ユーティリティーの他のオプションとは異なり、投影法はマップ視覚化の作成後に変更することができます。

投影法**:** マップの投影法を選択します。世界地図または半球地図を作成する場合は、ローカル 投影法、メルカトル 図法、またはヴィンケル 図法を使用してください。狭い領域の場合は、ローカル 投影法、ランベルト正角円錐 図法、または横メルカトル 図法を使用してください。いずれの投影法も、基準面に WGS83 楕円体が使用されます。

- マップがローカル座標系 (State Plane Coordinate System など) で作成された場合、常にローカル投影法が使用されます。この座標系は、地理座標 (経度と緯度) ではなく、デカルト座標によって定義されます。ローカル投影法では、水平線と垂直線がデカルト座標系に等間隔に配置されます。ローカル投影法は正角ではありません。
- メルカトル図法は、世界地図を描画するための正角図法です。水平線と垂直線は直線で、常に直交します。メルカトル図法は北極と南極に近づくに従って無限大に拡大するため、使用するマップに北極や南極が含まれている場合は使用できません。北極や南極に極限に近づくと、ゆがみが最大限になります。
- ヴィンケル図法は、世界地図を描画するための非正角図法です。この図法は正角ではありませんが、形状と大きさのバランスが適度に取れています。赤道とグリニッジ子午線を除き、すべての線が曲線になります。使用する世界地図に北極や南極が含まれている場合は、この投影法が適しています。
- ランベルト正角円錐図法は、その名前から分かるように正角図法です。南北に比べて東西の距離が長い、大陸以下の規模の土地の地図に使用されます。
- 横メルカトル図法も、正角図法の 1 つです。大陸以下の規模の土地の地図に使用されます。この投影法は、東西に比べて南北の距離が長い土地の地図に使用されます。

ステップ 4 - 完了

この時点で、マップ・ファイルについて説明するコメントを追加し、マップ キーからサンプル・データ・ファイルを作成することができます。

マップ キー **(Map Keys):** マップ・ファイルに複数のキーがある場合は、プレビューに表示したいフィーチャー・ラベル持つマップを選択します。マップからデータ・ファイルを作成すると、それらのラベルがデータ値で使用されます。

コメント**:** マップについて説明したりユーザーに関連する追加情報を提供したりするコメント (元のシェープファイルのソースなど) を入力します。コメントは、グラフボード・テンプレート選択の管理システムに表示されます。

フィーチャー・ラベルからデータ・セットを作成 **(Create a data set from the feature labels):** 表示されているフィーチャー・ラベルからデータ・ファイルを作成する場合は、このオプションにチェック・マークを付けます。「参照**...**」をクリックすると、場所とファイル名を指定することができます。拡張子 *.txt* を追加すると、ファイルがタブ区切り値ファイルとして保存されます。拡張子 *.csv* を追加すると、ファイルがカンマ区切り値ファイルとして保存されます。拡張子 *.sav* を追加すると、ファイルが IBM SPSS Statistics 形式で保存されます。拡張子を指定しなかった場合のデフォルトは SAV です。

マップ・ファイルの配布

マップ変換ユーティリティーの最初のステップでは、変換後の SMZ ファイルの保存先を選択します。グラフボード・テンプレート選択の管理システムにマップを追加することが既に選択されている場合もあります。管理システムに保存することを選択した場合は、同じコンピューター上で稼働するすべての IBM SPSS 製品でマップを使用できるようになります。

マップを他のユーザーに配布するには、そのユーザーに SMZ を送信する必要があります。これにより、そのユーザーが管理システムを使用してマップをインポートできるようになります。単に、ステップ 1 で場所を指定したファイルを送信するだけです。管理システムに存在するファイルを送信する場合は、以下の手順に従い、最初にそのファイルをエクスポートする必要があります。

1. 「テンプレートの選択」で「管理...」をクリックします。
2. 「マップ」タブをクリックします。
3. 配布するマップを選択します。
4. 「エクスポート...」をクリックしてファイルの保存先を選択します。

これで、物理マップ・ファイルを他のユーザーに送信できるようになりました。ユーザーは、この処理を逆の順序で行ってマップを管理システムにインポートする必要があります。

散布図ノード

散布図ノードでは、数値フィールド間の相関が示されます。散布図と呼ばれる点を使用した作図を作成したり、折れ線を使用することができます。ダイアログ・ボックスで「X モード」を指定すると、3 種類の折れ線を作成することができます。

X モード = ソート

「X モード」を 「ソート」 に設定すると、*x* 軸に作図するフィールドの値に基づいてデータがソートされます。この場合、グラフ上で左から右へ進む 1 つの線が生成されます。名義型をオーバーレイとして使用すると、グラフの左から右へと進む異なる色の複数の線が生成されます。

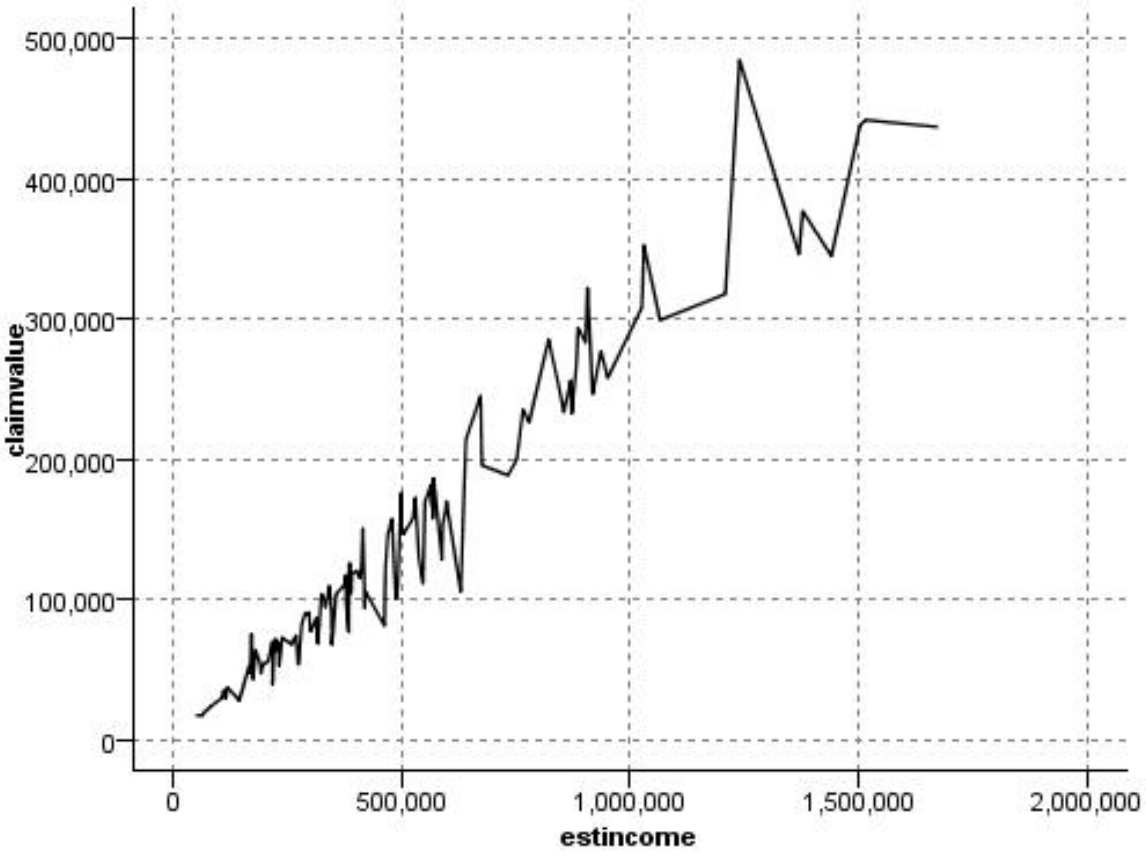


図 *19.* 「*X* モード」を「ソート」に設定した折れ線

X モード = オーバーレイ

「X モード」を 「オーバーレイ」 に設定すると、同じグラフ上で複数の折れ線が作成されます。オーバーレイ・プロットの場合、データをソートすることはできません。*x* 軸の値が増え続ける限り、データは 1 本の線上に作図されます。値が減少すると新しい線が作成されます。例えば、*x* 値が 0 から 100 に変化する場合、*y* 値は 1 つの線に作図されます。しかし、*x* 値が 100 を下回ると、最初の線のほかに新しい線が作図されます。このため、最終的にはグラフに多数の作図が描画されることもあります。これは、連続する複数の *y* 値を比較する場合に便利です。このタイプの作図は、連続する 24 時間単位の電力需要など、定期的な時間コンポーネントを持つデータに適しています。

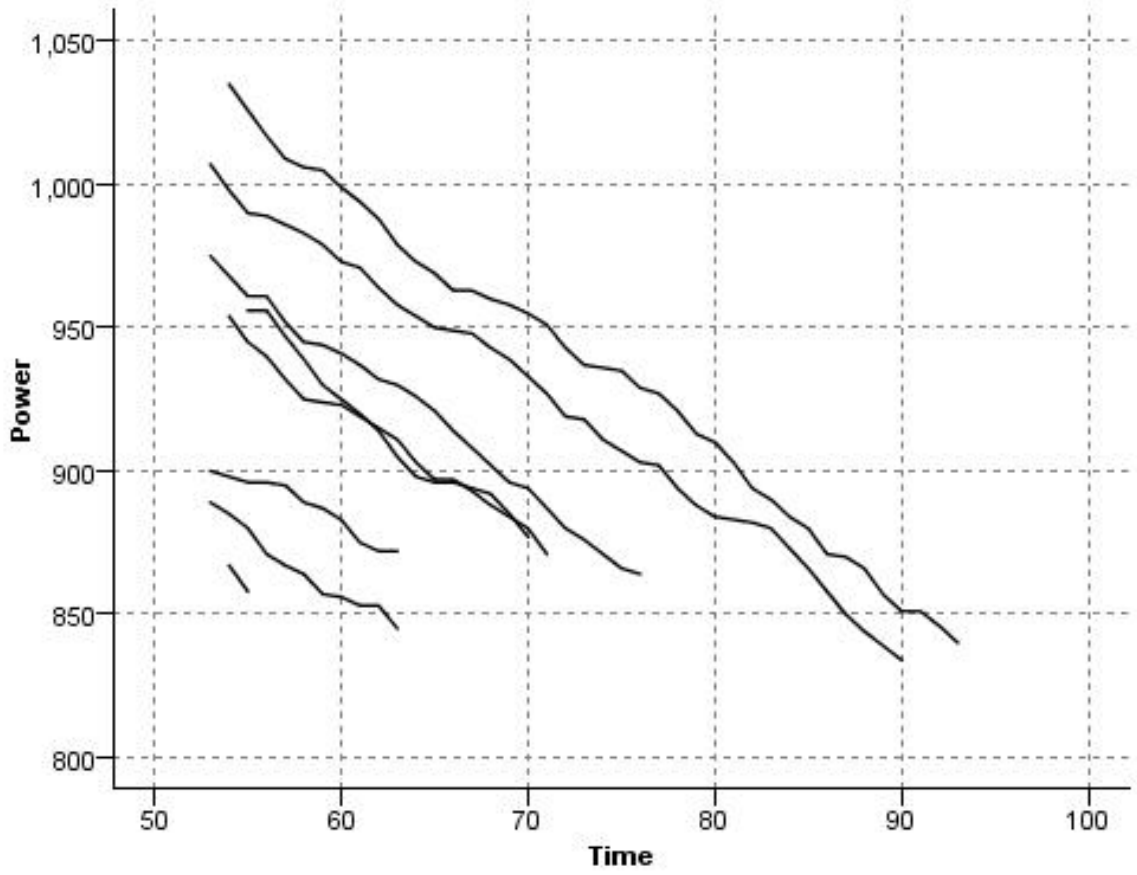


図 *20.* 「*X* モード」を「オーバーレイ」に設定した折れ線

X モード = 読み込み順

「X モード」を 「読み込み順」 に設定すると、*x* 値と *y* 値はデータ・ソースから読み込まれた順に作図されます。このオプションは、傾向、つまりデータの順序に依存するパターンを調べるときに使用する、時系列コンポーネントを持つデータに適しています。このタイプの散布図を作成する前に、データをソートする必要がある場合があります。また、「X モード」を 「ソート」 と 「読み込み順」 に設定した場合の 2 つの類似プロットを比較して、パターンがどの程度ソートに依存しているかを調べるときにも役立ちます。

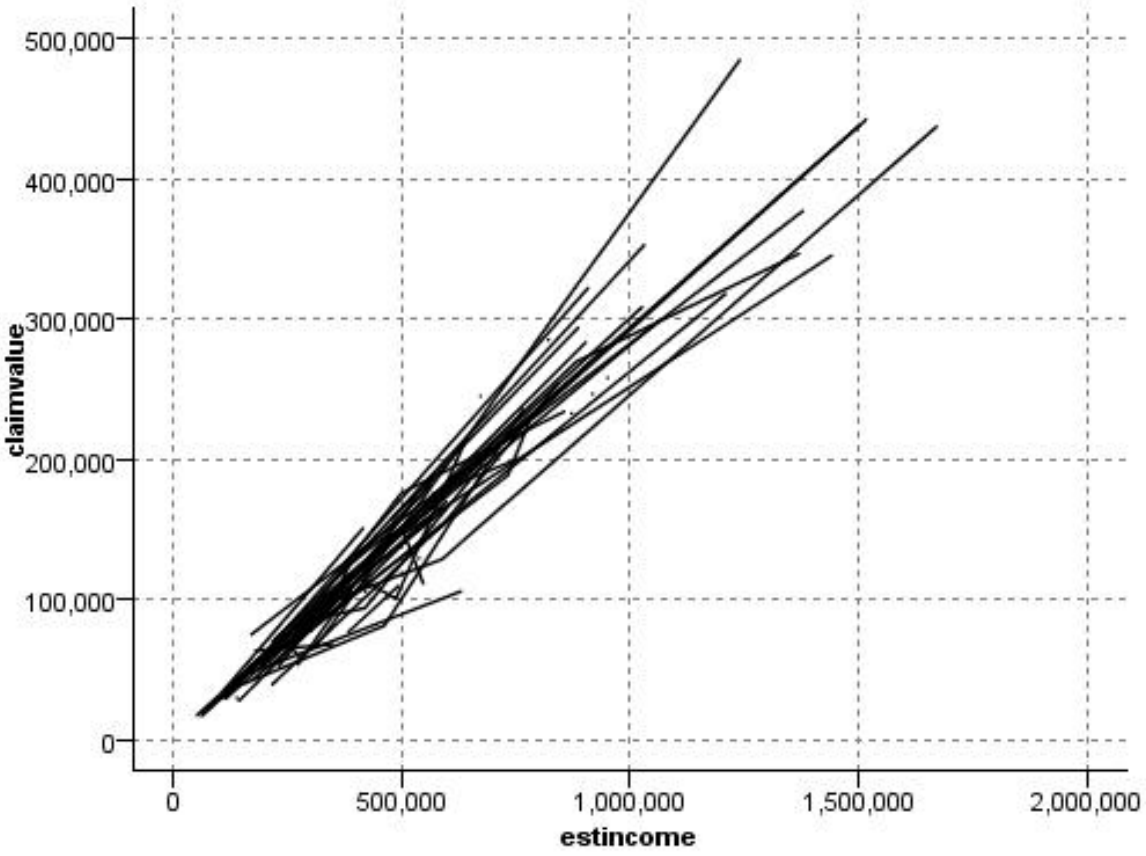


図 *21.* 最初「*X* モード］を「ソート」で、次に「読み込み順」で再実行した折れ線

グラフボード・ノードを使用して散布図や折れ線を生成することもできます。ただし、このノードでは、選択肢のオプション数が多くなります。 詳しくは、トピック 210 ページの『組み込まれている利用可能なグラフボード視覚化タイプ』を参照してください。

散布図ノードのタブ

散布図は、*X* フィールドの値に対する *Y* フィールドの値を表しています。多くの場合、これらのフィールドはそれぞれ従属変数と独立変数に対応しています。

X フィールド。リストから、横の *x* 軸に表示するフィールドを選択します。

Y フィールド**:**リストから、縦の *y* 軸に表示するフィールドを選択します。

Z フィールド**:**「3 次元グラフ」ボタンをクリックすると、*z* 軸に表示するフィールドをリストから選択できます。

オーバーレイ**:** データ値のカテゴリーを描くにはさまざまな方法があります。例えば、*maincrop* (主作物) を色のオーバーレイとして使用して、申請者による主作物の成長に応じた *estincome* (推定所得) と *claimvalue* (申請値) の値を示すことができます。詳しくは、トピック 201 ページの『外観、オーバーレイ、パネル、およびアニメーション』を参照してください。

オーバーレイ タイプ **:**オーバーレイ関数が表示されるか、平滑化が表示されるかを指定します。平準化とオーバーレイ 関数は、常に y の関数として計算されます。

- なし**:** オーバーレイを表示しません。
- 平滑化**:** 局部的に重みを付けたインタラクティブな強力な最小 2 乗法 (LOESS) を使用して計算された平準化フィット・ラインを表示します。この方法は散布図内の狭い領域に焦点をあてて、一連の回帰を効果的に計算します。これで、滑らかな曲線を作成するために後に結合される、一連の「局所的な」回帰線が作成されます。

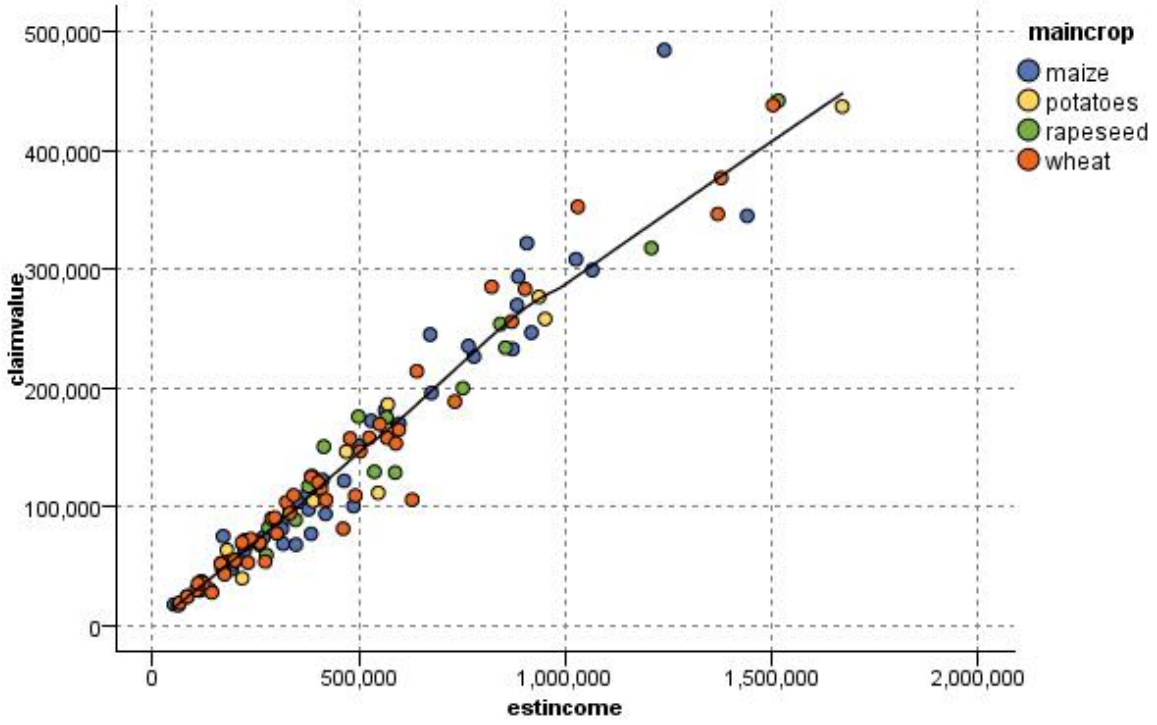


図 *22. LOESS* 平準化オーバーレイによる作図

- 関数**:** 実際の値と比較する既知の関数を指定する場合に選択します。例えば、実際の値と予測値を比較する場合は、y = x という関数をオーバーレイとして作図できます。「**y =**」 テキスト・ボックスで関数を指定してください。デフォルトの関数は y = x ですが、x の代わりに、2 次関数や任意の式などのあらゆる種類の関数を指定することもできます。

  注 : パネルまたはアニメーション・グラフでオーバーレイ関数を使用することはできません。

作図のオプションを設定したら、「実行」 をクリックして、ダイアログ・ボックスから直接プロットを実行できます。「オプション」タブを使用して、区分け、X モード、およびスタイルなどを指定することもできます。

散布図の「オプション」タブ

スタイル。作図のスタイルとして 「ポイント」 または 「線」 のどちらかを選択します。「線」 を選択すると、「**X** モード」 コントロールが有効になります。「ポイント」 を選択すると、プラス記号 (+) をデフォルトのポイント形状として使用します。いったんグラフを作成したら、ポイントの形状およびサイズを変更することができます。

X モード。折れ線グラフの場合は、「X モード」フィールドを選択して、折れ線のスタイルを定義する必要があります。「ソート」、「オーバーレイ」、または「読み込み順」を選択します。「オーバーレイ」または 「読み込み順」 を選択した場合、最初の *n* レコードのサンプリングに使用する最大データ・セット・サイズを指定する必要があります。それ以外の場合は、デフォルトの 2,000 レコードが使用されます。

自動 **X** 範囲。この軸に沿ったデータ中の値の範囲全体を使用します。指定した 「最小」 および 「最大」 に基づいて値の一部を明示的に使用する場合は、選択を解除してください。この範囲は、値を入力するか矢印を使用して指定します。デフォルトでは、グラフの構築を高速化するために、自動範囲のオプションが選択されています。

自動 **Y** 範囲。この軸に沿ったデータ中の値の範囲全体を使用します。指定した 「最小」 および 「最大」 に基づいて値の一部を明示的に使用する場合は、選択を解除してください。この範囲は、値を入力するか矢印を使用して指定します。デフォルトでは、グラフの構築を高速化するために、自動範囲のオプションが選択されています。

自動 **Z** 範囲。「作図」タブで 3 次元グラフが指定されている場合のみ。 この軸に沿ったデータ中の値の範囲全体を使用します。指定した 「最小」 および 「最大」 に基づいて値の一部を明示的に使用する場合は、選択を解除してください。この範囲は、値を入力するか矢印を使用して指定します。デフォルトでは、グラフの構築を高速化するために、自動範囲のオプションが選択されています。

ジッター。拡散とも呼ばれます。ジッターは、多くの値が繰り返されるデータ・セットの点プロットの場合に便利です。値の分布を明確にするため、ジッタを使用して実際の値の周囲に無作為 (ランダム) にポイントを分散できます。

前のバージョンの *IBM SPSS Modeler* ユーザーに対する注意 :散布図に使用するジッタ値は、本リリースの IBM SPSS Modeler では異なるメトリックを使用しています。前のバージョンでは、実際の数字が値になりましたが、今回はフレーム・サイズの比率に変更されています。つまり、古いストリームで使われている拡散値は大きすぎる可能性があります。このリリースでは、ゼロ以外の拡散値は 0.2 に変換されます。

プロットするレコードの最大数。大きいデータ・セットの作図手法を指定します。最大データ・セット・サイズを使用するか、またはデフォルト値 (2,000 レコード) を使用することができます。「ビン」 または 「サンプル」 を選択すると、大きいデータ・セットに対するパフォーマンスが向上します。代わりに、「すべてのデータを使用」 を選択して、すべてのデータ・ポイントを作図することもできます。ただし、この場合ソフトウェアのパフォーマンスが大幅に低下する可能性があります。

注 :「X モード」を 「オーバーレイ」 または 「読み込み順」 に設定した場合、これらのオプションは無効になり、最初の *n* レコードだけが使用されます。

- ビン。データ・セットに格納されているレコード数が、指定した数より大きい場合に、分割を有効にします。分割を行うと、グラフが細かいグリッドに分割されてから、作図や各グリッド・セルに現れるポイント数のカウントが実際に行われます。最終的なグラフでは、ビン重心 (ビン中のすべてのポイントの位置の平均) でセルごとに 1 つのポイントが作図されます。作図されたシンボルの大きさは、その領域中にあるポイント数を示しています (サイズをオーバーレイとして使用しない場合)。重心とサイズでポイント数を表すことにより、密集領域への過度の作図 (画一的な色の集合) やシンボルの羅列 (人工的

な重心パターン) を避けることができます。そのため、分割された作図は大きいデータ・セットを表すための最適な方法となっています。このようなシンボルの羅列は、特定のシンボル (特にプラス記号 [+]) が、生データ中に存在しない密集領域を生成するような競合がある場合に発生します。

- サンプル。ここに入力した数のレコードまで、無作為にデータのサンプリングを行います。デフォルトは 2,000 です。

散布図の「外観」タブ

グラフ作成前に外観オプションを指定できます。

表題: グラフのタイトルに使用するテキストを入力します。

サブタイトル: グラフのサブタイトルに使用するテキストを入力します。

キャプション: グラフのキャプションに使用するテキストを入力します。

X ラベル: 自動的に生成された x 座標 (水平) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

Y ラベル: 自動的に生成された y 座標 (垂直) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

Z ラベル: 3 次元のグラフのみで利用可能で、自動的に生成された z 座標ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

グリッドの表示: デフォルトで選択されているこのオプションは、散布図またはグラフの背景にグリッドを表示する場合に使用します。グリッドを使用すると、領域やバンドの分割点を簡単に決めることができます。グリッド線は、グラフの背景が白でない限り、常に白で表示されます。背景が白の場合は、灰色で表示されます。

散布図グラフの使用方法

散布図および線グラフは、基本的に Y に対する X の作図です。例えば、農業助成金申請における不法行為を調べる場合、申請者の申告している所得とニューラル・ネットワークによる推定所得を作図することができます。主要作物の種類などをオーバーレイすることにより、申請内容 (値または数字) と作物の種類の関係があるかどうかを描き出すことができます。

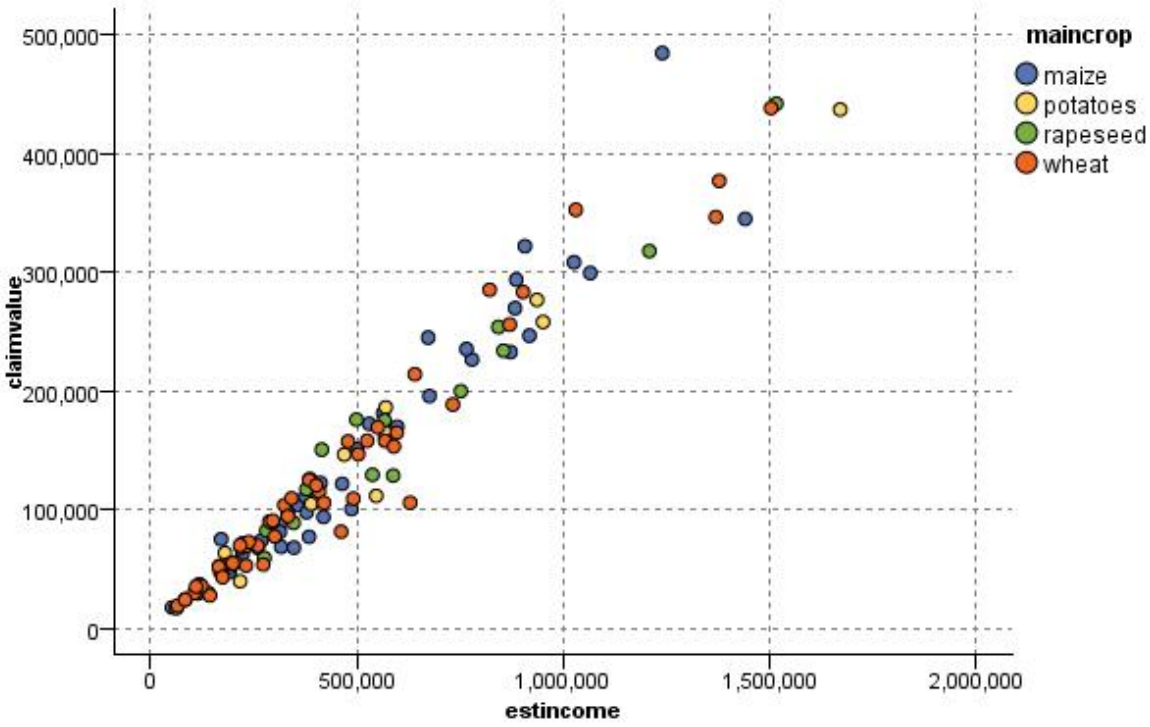


図 23. 推定所得と申請値間の関係を、主要作物の種類をオーバーレイとして描画

散布図グラフ、線グラフ、および評価グラフは、*X* に対する *Y* の関係を 2 次元で表すため、領域を定義、要素をマーク、またはバンドを描画することによってグラフを簡単に操作することができます。これらの領域、バンド、または要素で表示されるデータのノードを生成することもできます。 詳しくは、トピック 293 ページの『グラフの検証』を参照してください。

線グラフ・ノード

線グラフは、1 つの *X* フィールドに対して複数の *Y* フィールドを表示する特殊な種類の作図です。*Y* フィールドは色線で作図され、それぞれ「スタイル」フィールドを 「ライン」 に、「X モード」フィールドを 「ソート」 に設定した散布図ノードと等しくなります。時系列データがあり、時間の経過に伴うさまざまな変数の変動を調査するような場合に、線グラフが役立ちます。

線グラフの「作図」タブ

X フィールド。リストから、横の *x* 軸に表示するフィールドを選択します。

Y フィールド。*X* フィールドの値の範囲にわたって表示する 1 つ以上のフィールドをリストから選択します。複数のフィールドを選択するには、フィールド・ピッカー・ボタンを使用してください。リストからフィールドを削除する場合は、削除ボタンをクリックします。

オーバーレイ。データ値のカテゴリーを描くにはさまざまな方法があります。例えば、アニメーション・オーバーレイを使用して、データ中の各値を示す複数の散布図を表示することができます。これは、カテゴリ

ーが 10 個以上ある場合に役立ちます。カテゴリー数が 15 を超えると、パフォーマンスが低下する可能性があります。 詳しくは、トピック 201 ページの『外観、オーバーレイ、パネル、およびアニメーション』を参照してください。

正規化。すべての Y 値をグラフの 0 から 1 の範囲に対応させて表示する場合に選択します。正規化を使用することで、各系列の値の範囲内での差異が原因で不明確になりかねない各線の間の関係が明らかになります。また、同じグラフ上に複数の線を作図する場合や、隣り合ったパネル内で作図を比較する場合に、正規化をお勧めします。(正規化は、すべてのデータ値が似たような範囲内に収まる場合は不要です。)

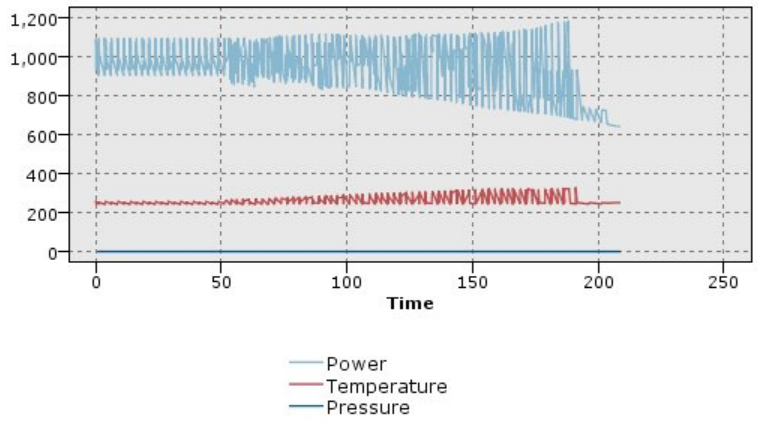


図 24. 時間の経過による発電装置の変動を示す標準の線グラフ (正規化を行わないと、圧力の作図は参照できないことに注意してください)

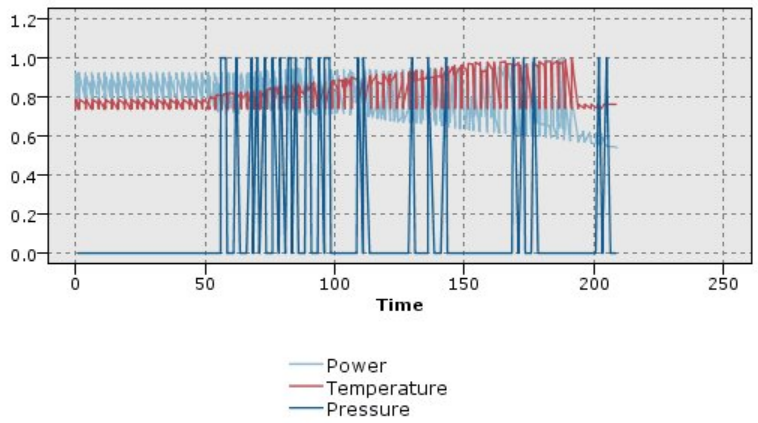


図 25. 圧力の作図を表示する正規化された線グラフ

オーバーレイ機能。実際の値と比較する既知の関数を指定する場合に選択します。例えば、実際の値と予測値を比較する場合は、y = x という関数をオーバーレイとして作図できます。「**y =**」 テキスト・ボックスで関数を指定してください。デフォルトの関数は y = x ですが、x の代わりに、2 次関数や任意の式などのあらゆる種類の関数を指定することもできます。

注 : パネルまたはアニメーション・グラフでオーバーレイ関数を使用することはできません。

レコード数が次の値より大きい場合**:** 大規模データ・セットの作図の手法を指定します。最大データ・セット・サイズを使用するか、またはデフォルト値 (2,000 ポイント) を使用することができます。「ビン」 または 「サンプル」 を選択すると、大きいデータ・セットに対するパフォーマンスが向上します。代わりに、「すべてのデータを使用」 を選択して、すべてのデータ・ポイントを作図することもできます。ただし、この場合ソフトウェアのパフォーマンスが大幅に低下する可能性があります。

注 : 「X モード」を 「オーバーレイ」 または 「読み込み順」 に設定した場合、これらのオプションは無効になり、最初の *n* レコードだけが使用されます。

- ビン。データ・セットに格納されているレコード数が、指定した数より大きい場合に、分割を有効にします。分割を行うと、グラフが細かいグリッドに分割されてから、作図や各グリッド・セルに現れる接続数のカウントが実際に行われます。最終的なグラフでは、ビン重心 (ビン中のすべての接続の位置の平均) でセルごとに 1 つの接続が作図されます。
- サンプル。ここに指定した数のレコードまで、無作為にデータのサンプリングを行います。

線グラフの「外観」タブ

グラフ作成前に外観オプションを指定できます。

表題**:** グラフのタイトルに使用するテキストを入力します。

サブタイトル**:** グラフのサブタイトルに使用するテキストを入力します。

キャプション**:** グラフのキャプションに使用するテキストを入力します。

X ラベル**:** 自動的に生成された *x* 座標 (水平) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

Y ラベル**:** 自動的に生成された *y* 座標 (垂直) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

グリッドの表示**:** デフォルトで選択されているこのオプションは、散布図またはグラフの背景にグリッドを表示する場合に使用します。グリッドを使用すると、領域やバンドの分割点を簡単に決めることができます。グリッド線は、グラフの背景が白でない限り、常に白で表示されます。背景が白の場合は、灰色で表示されます。

線グラフの使用方法

散布図および線グラフは、基本的に *Y* に対する *X* の作図です。例えば、農業助成金申請における不法行為を調べる場合、申請者の申告している所得とニューラル・ネットワークによる推定所得を作図することができます。主要作物の種類などをオーバーレイすることにより、申請内容 (値または数字) と作物の種類の関係があるかどうかを描き出すことができます。

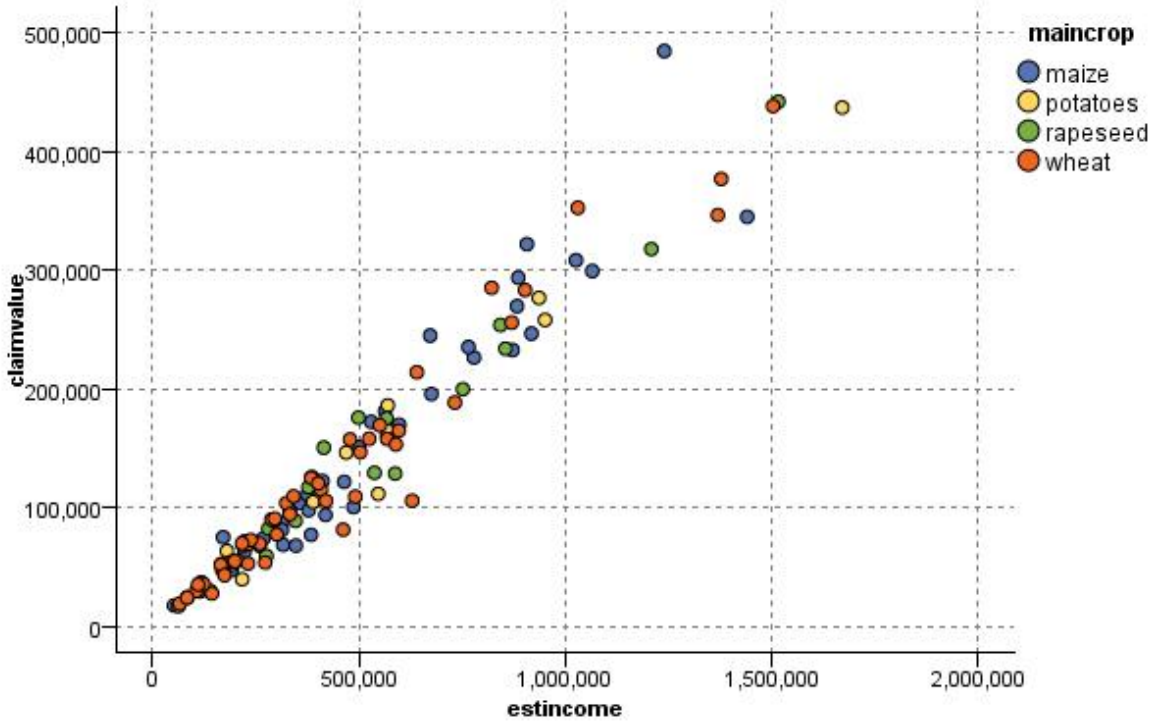


図 *26.* 推定所得と申請値間の関係を、主要作物の種類をオーバーレイとして描画

散布図グラフ、線グラフ、および評価グラフは、*X* に対する *Y* の関係を 2 次元で表すため、領域を定義、要素をマーク、またはバンドを描画することによってグラフを簡単に操作することができます。これらの領域、バンド、または要素で表示されるデータのノードを生成することもできます。 詳しくは、トピック 293 ページの『グラフの検証』を参照してください。

時系列ノード

時系列ノードを使用すると、1 つ以上の時系列を時間の経過に従ってプロットして表示することができます。一連の作図には数値が含まれていなければならず、周期が一定の時間の領域に作図することを前提とします。

SPSS Modeler バージョン 17.1 以前の場合は、通常、時系列ノードの前に時間区分ノードを使用して *TimeLabel* フィールドを生成します。このフィールドはデフォルトでグラフに *x* 軸のラベルを指定するのに使用されます。

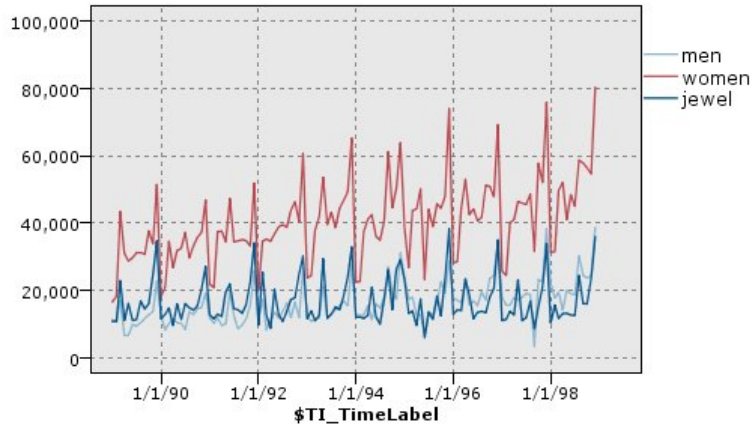


図 27. 男女の衣類や宝石の売り上げを長期間プロット

干渉およびイベントの作成

コンテキスト・メニューからフィールド作成 (フラグまたは名義型) ノードを生成し、時系列からイベントおよび干渉フィールドを作成することができます。例えば、鉄道会社のストライキが発生した場合イベント・フィールドを作成することができ、そこでイベントが発生した場合はドライブ ステートを真 (true)、発生しない場合は偽 (false) となります。干渉フィールドの場合、例えば価格の値上げに対して、フィールド作成 (カウント型) を使用し、古い価格に対しては 0 、新しい価格に対しては 1 を指定して、値上げの日付を識別することができます。 詳しくは、トピック 162 ページの『フィールド作成ノード』を参照してください。

時系列の「作図」タブ

作図。時系列データの作図方法の選択を提供します。

- 選択された系列。選択された時系列に対する作図値。信頼区間の作図時このオプションを選択する場合、「正規化」 チェック・ボックスを解除します。
- 選択された時系列モデル。時系列モデルとともに使用し、このオプションはすべての関連フィールド (信頼区間同様、実際のおよび予測された値) を 1 つ以上の選択された時系列に対して作図します。このオプションは、ダイアログ・ボックスのその他のオプションを無効化します。信頼区間を作図する場合、このオプションが優先されます。

系列。プロット対象の時系列データを含む 1 つ以上のフィールドを選択します。このデータは数値でなければなりません。

X 軸ラベル。デフォルトのラベルを選択するか、単一のフィールドを選択して、作図で *x* 軸のラベルとして使用します。「デフォルト」を選択すると、(SPSS Modeler バージョン 17.1 以前で作成されたストリームの) 上流の時間区分ノード、または上流の時間区分ノードが存在しない場合は連続する整数から生成された TimeLabel フィールドが使用されます。

別のパネルに時系列を表示。別のパネルにそれぞれの時系列を表示するかどうかを指定します。または、パネルを選択しない場合は、すべての時系列が同じグラフ上に作図され、平準化は利用できません。すべての時系列を同一グラフに作図する場合、各時系列は別々の色で表されます。

正規化。すべての *Y* 値をグラフの 0 から 1 の範囲に対応させて表示する場合に選択します。正規化を使用することで、各系列の値の範囲内での差異が原因で不明確になりかねない各線の間の関係が明らかになり

ます。また、同じグラフ上に複数の線を作図する場合や、隣り合ったパネル内で作図を比較する場合に、正規化をお勧めします。(正規化は、すべてのデータ値が似たような範囲内に収まる場合は不要です。)

表示: グラフに表示する要素を 1 つ以上選択します。「ライン」、「ポイント」、「(LOESS) 平準化」から選択できます。平準化は、時系列を別のパネルに表示している場合に使用できます。デフォルトでは、ライン要素が選択されています。グラフ・ノードを実行する前に、1 つ以上の作図要素を必ず選択してください。何も選択しないと、作図対象が選択されていないことを示すエラーが返されます。

レコードを制限。作図されるレコードを制限したい場合は、このオプションを選択します。データ・ファイルの先頭部分から読み取った、作図されるレコードの数を、「プロットするレコードの最大数」 オプションで指定します。デフォルトでは、2,000 に設定されています。データ・ファイルの最後の n 個のレコードを作図したい場合は、このノードを実行する前にソートノードを使用してレコードを時間の降順に並べ替えることができます。

時系列の「外観」タブ

グラフ作成前に外観オプションを指定できます。

表題: グラフのタイトルに使用するテキストを入力します。

サブタイトル: グラフのサブタイトルに使用するテキストを入力します。

キャプション: グラフのキャプションに使用するテキストを入力します。

X ラベル: 自動的に生成された x 座標 (水平) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

Y ラベル: 自動的に生成された y 座標 (垂直) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

グリッドの表示: デフォルトで選択されているこのオプションは、散布図またはグラフの背景にグリッドを表示する場合に使用します。グリッドを使用すると、領域やバンドの分割点を簡単に決めることができます。グリッド線は、グラフの背景が白でない限り、常に白で表示されます。背景が白の場合は、灰色で表示されます。

レイアウト: 時系列についてのみ、時間値を横軸に表示するか縦軸に表示するかを指定することができます。

時系列グラフの使用方法

時系列グラフを生成すると、さまざまなオプションを使用してグラフの表示を調整したり、詳細な分析のためのノードを生成したりできるようになります。 詳しくは、トピック 293 ページの『グラフの検証』を参照してください。

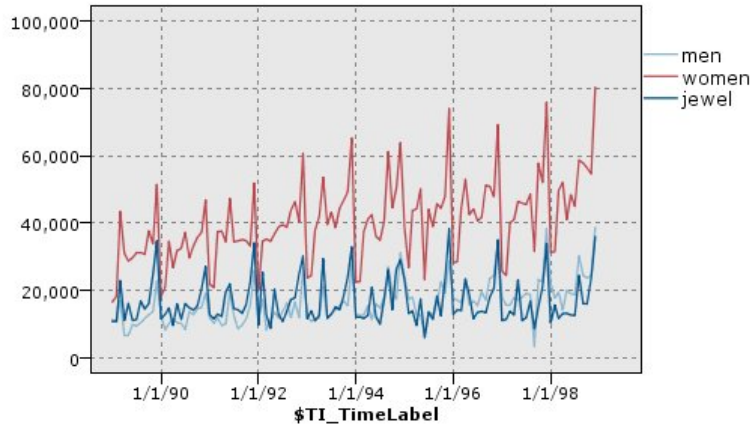


図 *28.* 男女の衣類や宝石の売り上げを長期間プロット

時系列を作成し、バンドを定義して結果を調べると、「ノードの生成」メニューとコンテキスト・メニューのオプションを使用して、バランス・ノード、条件抽出ノード、またはフィールド作成ノードを生成できるようになります。 詳しくは、トピック 300 ページの『グラフからのノードの生成』を参照してください。

棒グラフ・ノード

棒グラフまたはテーブルは、ローンの種類や性別など、データ・セット内のシンボル値 (非数値) の出現頻度を示します。一般に棒グラフ・ノードは、モデルの作成前にバランス・ノードを使用して修正できる、データの不均衡を表す場合に使用されます。バランス・ノードは、棒グラフ・ウィンドウの「ノードの生成」メニューを使用して自動的に生成することができます。

グラフボード・ノードを使用して棒グラフを作成することもできます。ただし、このノードでは、選択肢のオプション数が多くなります。 詳しくは、トピック 210 ページの『組み込まれている利用可能なグラフボード視覚化タイプ』を参照してください。

注：数値の出現頻度を表すには、ヒストグラム・ノードを使います。

棒グラフの「作図」タブ

作図。棒グラフの種類を選択してください。選択したフィールドの棒グラフを表示する場合は、「選択したフィールド」 を選択します。データ・セットのフラグ型フィールドの真 (true) の値の棒グラフを表示するには、「すべてのフラグ **(**真の値**)**」 を選択します。

フィールド。値の分布を表示する名義型またはフラグ型のフィールドを選択します。数値として明示的に設定されていないフィールドだけがリストに表示されます。

オーバーレイ。色のオーバーレイとして使用する名義型またはフラグ型フィールドを選択し、指定したフィールドの各値内の値の棒グラフを表示します。例えば、マーケティング・キャンペーンの回答者 (*pep*) を子供の数 (*children*) のオーバーレイとして使用し、家族サイズの回答を描画することができます。 詳しくは、トピック 201 ページの『外観、オーバーレイ、パネル、およびアニメーション』を参照してください。

色で正規化。グラフの幅全体を占めるようにすべてのバーを表示する場合に選択します。オーバーレイした値は各バーの比率と等しくなるため、カテゴリーを簡単に比較できます。

ソート。棒グラフに値を表示する方法を選択します。アルファベット順に並べる場合は 「アルファベット順」 を選択します。また、出現頻度の降順に並べる場合は 「出現頻度順」 を選択します。

プロポーショナル・スケール。最大カウントを持つ値が作図の幅全体を占めるように値の棒グラフを表示する場合に選択します。他のすべての棒はこの値を基準に表示されます。このオプションを選択しないと、バーは各値の合計カウントに従って表示されます。

棒グラフの「外観」タブ

グラフ作成前に外観オプションを指定できます。

表題: グラフのタイトルに使用するテキストを入力します。

サブタイトル: グラフのサブタイトルに使用するテキストを入力します。

キャプション: グラフのキャプションに使用するテキストを入力します。

X ラベル: 自動的に生成された x 座標 (水平) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

Y ラベル: 自動的に生成された y 座標 (垂直) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

グリッドの表示: デフォルトで選択されているこのオプションは、散布図またはグラフの背景にグリッドを表示する場合に使用します。グリッドを使用すると、領域やバンドの分割点を簡単に決めることができます。グリッド線は、グラフの背景が白でない限り、常に白で表示されます。背景が白の場合は、灰色で表示されます。

棒グラフ・ノードの使用方法

棒グラフ・ノードは、データ・セット中のシンボル値の分布を表すために使用されます。棒グラフは、操作ノードの前段階で、データの調査と不均衡の修正を行うためにしばしば利用されます。例えば、子供がいない回答者のインスタンスが、他の種類の回答者よりも頻繁に発生しているような場合、後のデータ・マイニング操作でより有益なルールを作成するために、これらのインスタンスを減らすことができます。このような不均衡を調査、修正するために、棒グラフ・ノードが役立ちます。

棒グラフ・ノードは、データを分析するグラフやテーブルを両方作成するという点で通常と異なります。

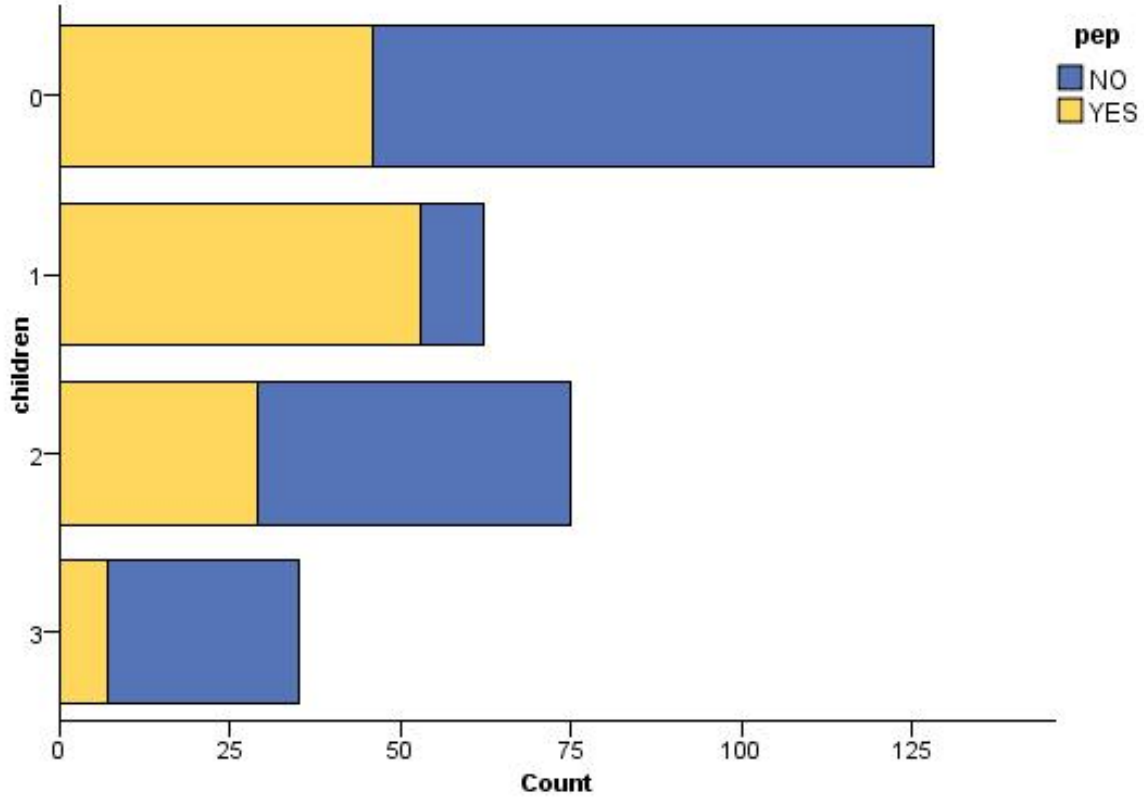


図 *29.* マーケティング・キャンペーンの回答者で子供がいる人、いない人の数を表す棒グラフ

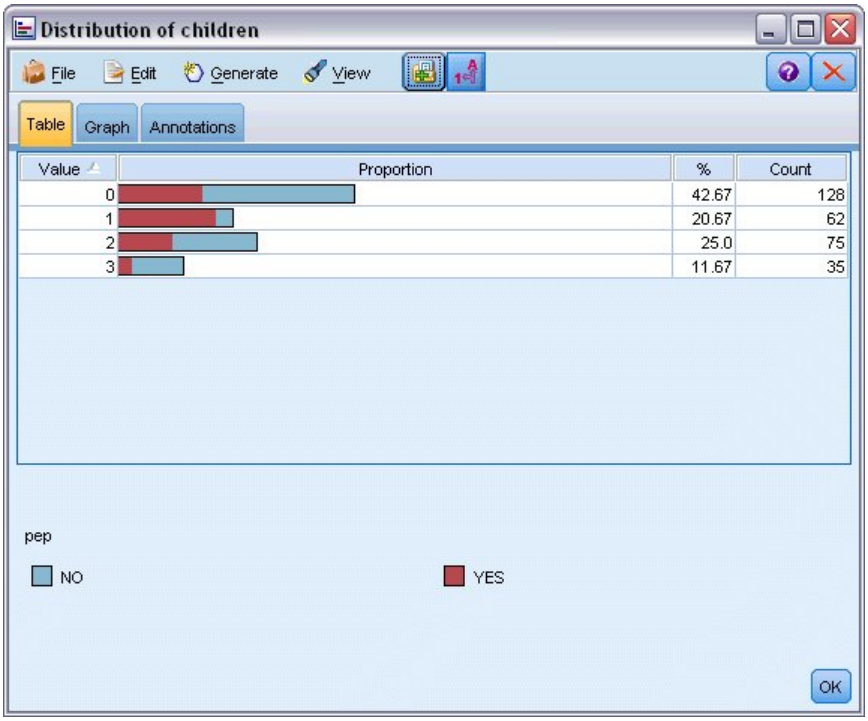


図 *30.* マーケティング・キャンペーンの回答者で子供がいる人、いない人の割合を表す棒グラフ テーブル

棒グラフ テーブルや棒グラフを作成して結果を調査したら、メニュー・オプションを使用して値のグループ化、値のコピー、およびノードの生成などを行って、データを準備することができます。さらに、グラフやテーブルの情報を、MS Word や MS PowerPoint など他のアプリケーションで使用するためにコピーまたはエクスポートすることができます。 詳しくは、トピック 316 ページの『グラフの印刷、保存、コピー、およびエクスポート』を参照してください。

棒グラフ テーブルから値を選択してコピーするには

1. 値のセットを選択するには、行の上でマウス・ボタンを押したままドラッグします。「編集」メニューから 「すべて選択」 を選択して、すべての値を選択することもできます。
2. 「編集」メニューから 「テーブルのコピー」 または 「テーブルのコピー **(**フィールド名を含む**)**」 を選択します。
3. クリップボードまたは目的のアプリケーションに貼り付けます。

   注 : バーが直接コピーされることはありません。代わりにテーブルの値がコピーされます。つまり、オーバーレイされた値はコピー先のテーブルに表示されません。

棒グラフ テーブルから値をグループ化するには

1. Ctrl キーを押しながら、グループ化する値を選択します。
2. 「編集」メニューから、「グループ化」 を選択します。

注 : 値をグループ化またはグループ化を解除する場合、「グラフ」タブのグラフは、自動的に再描画されて変更を表示します。

次の作業を行うこともできます。

- 棒グラフリストのグループ名を選択し、「編集」メニューから 「グループ解除」 を選択して値のグループ化を解除する。

- 棒グラフリストのグループ名を選択し、「編集」メニューから 「グループの編集」 を選択してグループを編集する。このオプションを選択すると、値をグループに追加、削除できるダイアログ・ボックスが表示されます。

「ノードの生成」メニューのオプション

「ノードの生成」メニューのオプションを使用して、データのサブセットの選択、フラグ型フィールドの作成、値の再グループ化、値の再分類、グラフまたはテーブルのデータの平均化などの作業を行うことができます。これらの操作により、データの準備ノードが生成され、それがストリーム領域に配置されます。生成されたノードを使用するには、これを既存のストリームと接続します。 詳しくは、トピック 300 ページの『グラフからのノードの生成』を参照してください。

ヒストグラム・ノード

ヒストグラム・ノードでは、数値フィールドの値の出現頻度が示されます。多くの場合、ヒストグラム・ノードは、操作やモデルの構築前にデータを調べる場合に使用されます。棒グラフ・ノードと同様、ヒストグラム・ノードはデータの不均衡を調べる場合にもよく使用されます。グラフボード・ノードを使用してヒストグラムを生成することもできますが、このノードでは、選択肢のオプションが多くなります。 詳しくは、トピック 210 ページの『組み込まれている利用可能なグラフボード視覚化タイプ』を参照してください。

注 : シンボル値フィールドの値の出現頻度を表すには、棒グラフ・ノードを使用してください。

ヒストグラムの「作図」タブ

フィールド**:** 値の分布を表示する数値フィールドを選択します。シンボル値 (カテゴリー) として明示的に定義されていないフィールドだけがリストに表示されます。

オーバーレイ**:** 指定されたフィールドの値のカテゴリーを示すシンボル値フィールドを選択します。「オーバーレイ フィールド」でフィールドを選択すると、ヒストグラムは選択したフィールドの各カテゴリーを色で表す積み重ねグラフに変換されます。ヒストグラム・ノードを使用すると、色、パネル、アニメーションの 3 つのオーバーレイがあります。 詳しくは、トピック 201 ページの『外観、オーバーレイ、パネル、およびアニメーション』を参照してください。

ヒストグラムの「オプション」タブ

自動 **X** 範囲**:** この軸に沿ったデータ中の値の範囲全体を使用します。指定した 「最小」 および 「最大」 に基づいて値の一部を明示的に使用する場合は、選択を解除してください。この範囲は、値を入力するか矢印を使用して指定します。デフォルトでは、グラフの構築を高速化するために、自動範囲のオプションが選択されています。

ビン**:** 「数を固定」または「幅を固定」のいずれかを選択します。

- 表示するには、「数を固定」 を選択します。バーの幅は、指定する範囲とビン数によって決まります。「ビン数」 オプションで、グラフで使用するビン数を指示します。矢印を使用して、数を調節してください。
- また、固定幅のバーを持つグラフを作成するには、「幅を固定」 を選択します。ビンの数は、指定した幅と値の範囲によって決まります。「ビン幅」 オプションでバーの幅を指示します。

色で正規化**:** すべてのバーを同じ高さに揃え、オーバーレイした値を各バーの全ケースに対する割合 (パーセント) として表示する場合に選択します。

正規曲線の表示: データの平均や変数を表示するグラフに、正規曲線を追加します。

各色ごとに個別のバンド: オーバーレイした各値をグラフ上に個別のバンドとして表示する場合に選択します。

ヒストグラムの「外観」タブ

グラフ作成前に外観オプションを指定できます。

表題: グラフのタイトルに使用するテキストを入力します。

サブタイトル: グラフのサブタイトルに使用するテキストを入力します。

キャプション: グラフのキャプションに使用するテキストを入力します。

X ラベル: 自動的に生成された x 座標 (水平) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

Y ラベル: 自動的に生成された y 座標 (垂直) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

グリッドの表示: デフォルトで選択されているこのオプションは、散布図またはグラフの背景にグリッドを表示する場合に使用します。グリッドを使用すると、領域やバンドの分割点を簡単に決めることができます。グリッド線は、グラフの背景が白でない限り、常に白で表示されます。背景が白の場合は、灰色で表示されます。

ヒストグラムの使用方法

ヒストグラムには、x 軸の範囲の値を取る数値フィールドの値の分布が表示されます。ヒストグラムは、集計棒グラフと同様に動作します。集計棒グラフは、単一のフィールドについての値の頻度ではなく、別のフィールドの値に関連する 1 つの数値フィールドの値の分布を表します。

グラフを作成したら、結果を調べてバンドを定義し、x 軸に沿って値を分割したり、領域を定義したりできます。また、グラフ内で要素をマークすることもできます。 詳しくは、トピック 293 ページの『グラフの検証』を参照してください。

「ノードの生成」メニューのオプションを使用すると、グラフ内、具体的にはバンド、領域、マークされた要素内のデータを使用して、バランス・ノード、条件抽出ノード、フィールド作成ノードを作成することができます。この種のグラフは、操作ノードの前段階において、ストリームで使用するグラフからバランス・ノードを生成して、データの調査と不均衡の修正を行うために頻繁に使用されます。また、フィールド作成ノード (フラグ型) を生成して各レコードがどのバンドに該当するかを表すフィールドを追加したり、条件抽出ノードを生成して特定のセットまたは値の範囲内のすべてのレコードを選択することもできます。特定のデータのサブセットをさらに詳細に調査するような場合に、このような操作が役立ちます。 詳しくは、トピック 300 ページの『グラフからのノードの生成』を参照してください。

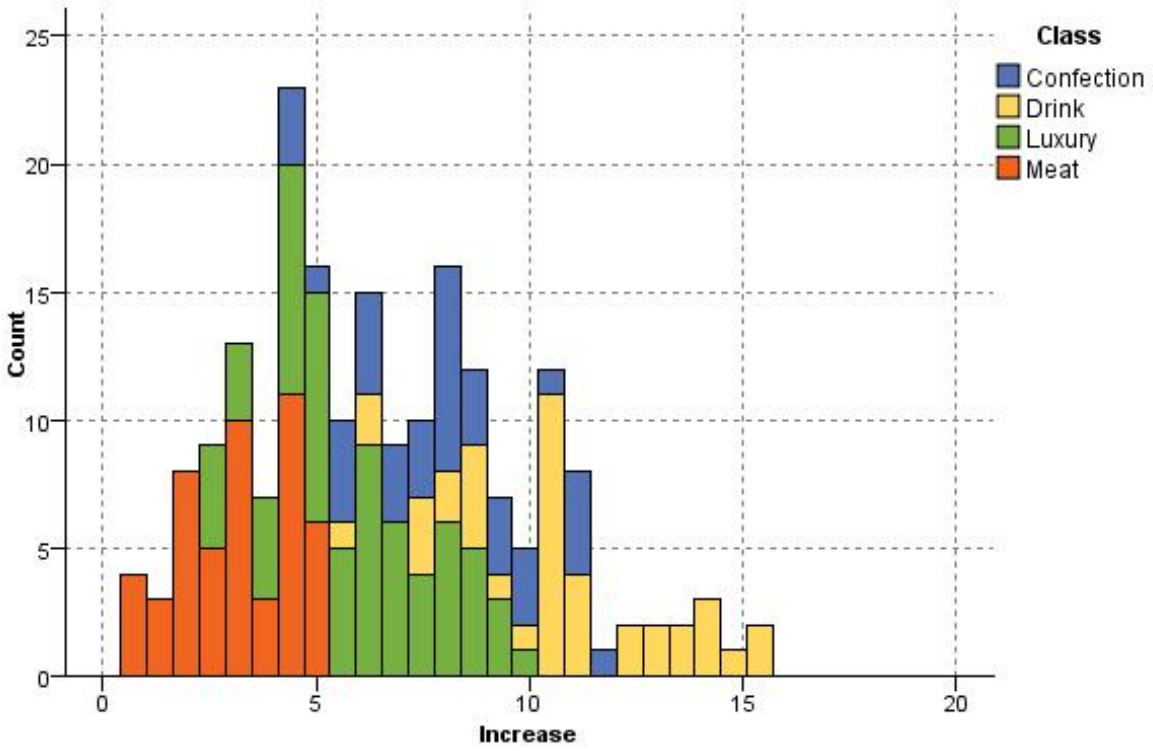


図 *31.* プロモーションによる購入の増加の分布をカテゴリー別に表したヒストグラム

集計棒グラフ・ノード

集計棒グラフはヒストグラムと似ていますが、1 つのフィールドの値の出現頻度ではなく、別のフィールドの値と連関がある 1 つの数値フィールドの値の棒グラフが示される点が異なります。集計棒グラフは、値が時間の経過とともに変化する変数やフィールドを表示する場合に役立ちます。3 次元グラフを使用して、分布をカテゴリー別に表示するシンボル値軸を追加することもできます。2 次元の集計グラフが積み上げ棒グラフで、使用されている場合はオーバーレイで表示されます。 詳しくは、トピック 201 ページの『外観、オーバーレイ、パネル、およびアニメーション』を参照してください。

集計棒グラフの「作図」タブ

集計棒グラフ **:**表示するフィールドを選択します。このフィールドの値は、「対象フィールド」 で指定するフィールドの値の範囲にわたって収集されます。シンボル値として定義されていないフィールドだけがリストに表示されます。

対象**:** 「初期データの収集」で指定されたフィールドの表示に使用する値を持つフィールドを選択します。

フィールド**:** 3 次元グラフを作成する際に有効になります。このオプションにより、カテゴリー別の集計フィールドを表示するための名義型フィールドまたはフラグ型フィールドを選択することができます。

演算**:** 集計棒グラフの各バーによって表される対象を選択します。オプションには、「合計」、「平均値」、「最大値」、「最小値」、および 「標準偏差」 が含まれています。

オーバーレイ**:** 選択したフィールドの値のカテゴリーを示すシンボル値フィールドを選択します。オーバーレイ フィールドを選択すると、集計棒グラフが変換され、複数のバーがカテゴリー別に色分けされて表示されます。このノードには、色、パネル、アニメーションの 3 つのオーバーレイがあります。 詳しくは、トピック 201 ページの『外観、オーバーレイ、パネル、およびアニメーション』を参照してください。

集計棒グラフの「オプション」タブ

自動 **X** 範囲**:** この軸に沿ったデータ中の値の範囲全体を使用します。指定した 「最小」 および 「最大」 に基づいて値の一部を明示的に使用する場合は、選択を解除してください。この範囲は、値を入力するか矢印を使用して指定します。デフォルトでは、グラフの構築を高速化するために、自動範囲のオプションが選択されています。

ビン**:** 「数を固定」または「幅を固定」のいずれかを選択します。

- 表示するには、「数を固定」 を選択します。バーの幅は、指定する範囲とビン数によって決まります。「ビン数」 オプションで、グラフで使用するビン数を指示します。矢印を使用して、数を調節してください。
- また、固定幅のバーを持つグラフを作成するには、「幅を固定」 を選択します。ビンの数は、指定した幅と値の範囲によって決まります。「ビン幅」 オプションでバーの幅を指示します。

集計棒グラフの「外観」タブ

グラフ作成前に外観オプションを指定できます。

表題**:** グラフのタイトルに使用するテキストを入力します。

サブタイトル**:** グラフのサブタイトルに使用するテキストを入力します。

キャプション**:** グラフのキャプションに使用するテキストを入力します。

対象フィールド・ラベル **:**自動生成されたラベルを承認するか、「ユーザー設定」 を選択してラベルを指定します。

初期データの収集ラベル **:**自動生成されたラベルを承認するか、「ユーザー設定」 を選択してラベルを指定します。

フィールド・ラベル **:**自動生成されたラベルを承認するか、「ユーザー設定」 を選択してラベルを指定します。

グリッドの表示**:** デフォルトで選択されているこのオプションは、散布図またはグラフの背景にグリッドを表示する場合に使用します。グリッドを使用すると、領域やバンドの分割点を簡単に決めることができます。グリッド線は、グラフの背景が白でない限り、常に白で表示されます。背景が白の場合は、灰色で表示されます。

次の例では、表示オプションが 3-D バージョンのグラフ内のどこに表示されるかを示します

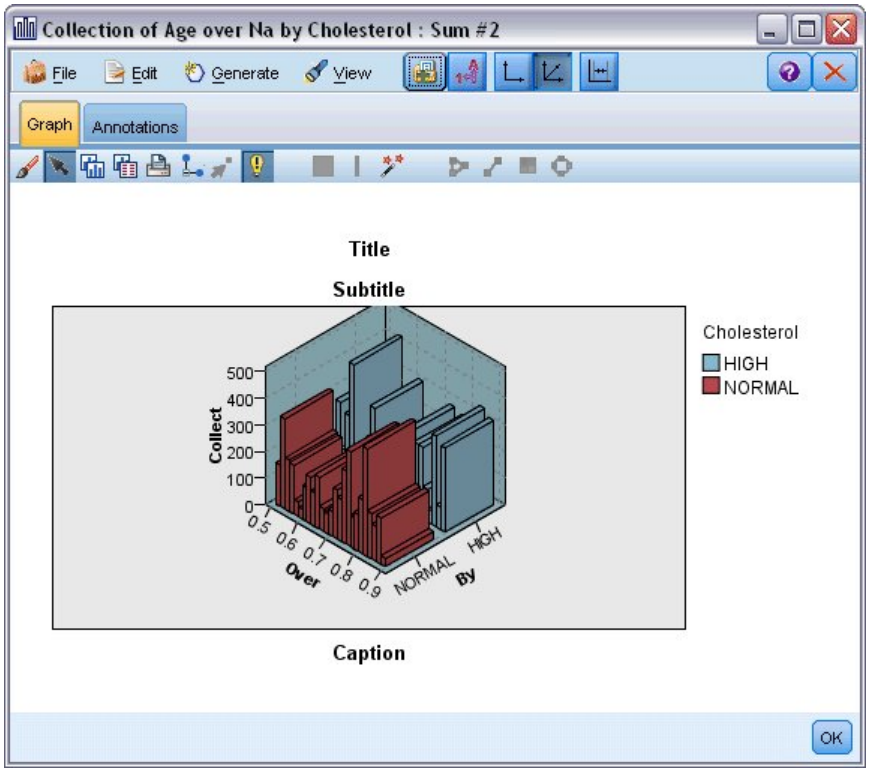


図 *32. 3-D* 集計棒グラフのグラフ外観オプションの位置

集計棒グラフの使用方法

集計棒グラフは、単一のフィールドについての値の頻度ではなく、別のフィールドの値に関連する 1 つの数値フィールドの値の分布を表します。ヒストグラムは、集計棒グラフと同様に動作します。ヒストグラムには、*x* 軸の範囲の値を取る数値フィールドの値の分布が表示されます。

グラフを作成したら、結果を調べてバンドを定義し、*x* 軸に沿って値を分割したり、領域を定義したりできます。また、グラフ内で要素をマークすることもできます。 詳しくは、トピック 293 ページの『グラフの検証』を参照してください。

「ノードの生成」メニューのオプションを使用すると、グラフ内、具体的にはバンド、領域、マークされた要素内のデータを使用して、バランス・ノード、条件抽出ノード、フィールド作成ノードを作成することができます。この種のグラフは、操作ノードの前段階において、ストリームで使用するグラフからバランス・ノードを生成して、データの調査と不均衡の修正を行うために頻繁に使用されます。また、フィールド作成ノード (フラグ型) を生成して各レコードがどのバンドに該当するかを表すフィールドを追加したり、条件抽出ノードを生成して特定のセットまたは値の範囲内のすべてのレコードを選択することもできます。特定のデータのサブセットをさらに詳細に調査するような場合に、このような操作が役立ちます。 詳しくは、トピック 300 ページの『グラフからのノードの生成』を参照してください。

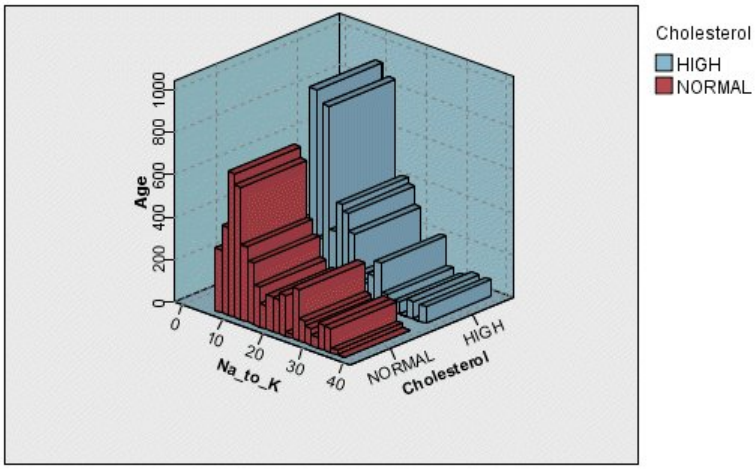


図 *33.* ナトリウム値/カリウム値の合計と年齢、およびコレステロール値の上限と正常値を表す *3* 次元集計棒グラフ

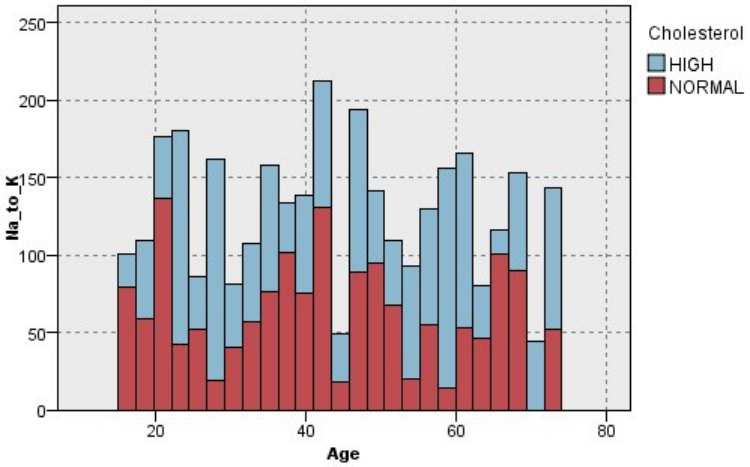


図 *34. z* 軸は表示せずにコレステロールを色のオーバーレイとして使用した集計棒グラフ

Web グラフ・ノード

Web グラフ・ノードでは、2 つ以上のシンボル値フィールドの値の相関の強さが示されます。このグラフでは相関の強さがさまざまな線の種類で示されます。Web グラフ・ノードを使用して、E コマース・サイトや従来の小売店で購入されたさまざまな商品の関係を調査することができます。

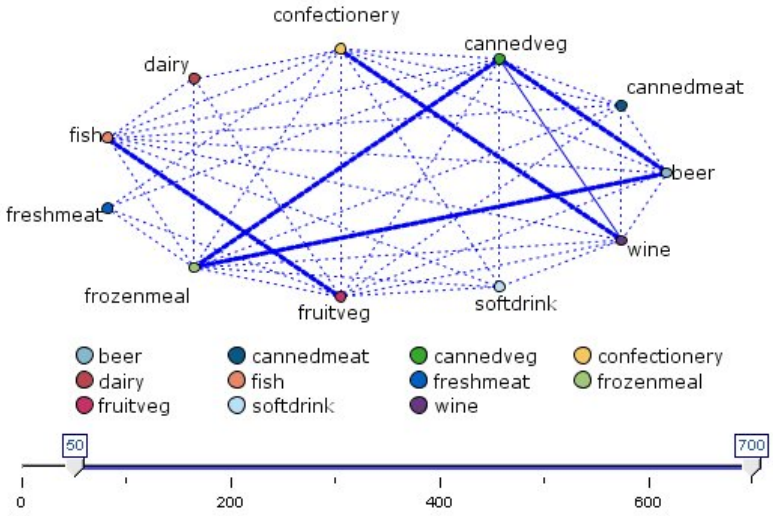


図 *35.* 食料雑貨の購入の関係を表した *Web* グラフ

Web グラフ

Web グラフ・ノードは、シンボル値フィールド間の相関の強さが示される点で MultiWeb グラフ・ノードと似ています。ただし、Web グラフ・ノードには、1 つの終点フィールドに対する 1 つ以上の始点フィールドからの相関だけが表示されます。これらの相関は一方方向です。

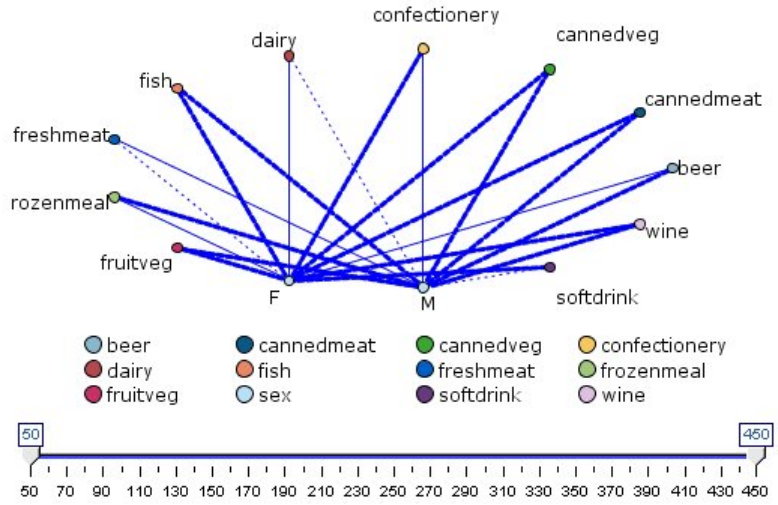


図 *36.* 購入された食料雑貨品と性別の関係を表示した *Web* グラフ

MultiWeb グラフ・ノードと同様、グラフではさまざまな線の種類を使用して相関の強さが示されます。Web グラフ・ノードを使用して、性別と特定商品の購入傾向との関連などを調査することができます。

Web グラフの「作図」タブ

MultiWeb グラフ**:** 指定したすべてのフィールド間の相関の強さを表す Web グラフを作成する場合に選択します。

Web グラフ**:** 複数のフィールドとある 1 つのフィールドの値 (例：性別や宗教など) の相関の強さを表す Web グラフを作成する場合に選択します。このオプションを選択すると、「終点フィールド」が有効になります。また、下にある「フィールド」コントロール名が、「始点フィールド」に変わります。

終点フィールド **(Web** グラフの場合**):** Web グラフで使用するフラグ型または名義型フィールドを選択します。数値として明示的に設定されていないフィールドだけがリストに表示されます。

フィールド**/**始点フィールド**:** Web グラフを作成するためのフィールドを選択します。数値として明示的に設定されていないフィールドだけがリストに表示されます。フィールド・ピッカー・ボタンを使用して複数のフィールドを選択するか、またはフィールドの種類を選択します。

注：Web グラフの場合、このコントロールは始点フィールドを選択するために使用します。

真 **(true)** のフラグだけを表示**:** フラグ型フィールドが真 (true) のフラグだけを表示する場合に選択します。このオプションを選択すると、Web グラフの表示が単純化されます。またこのオプションは、正の値の出現が特に重要となるデータによく使用されます。

線の値**:** ドロップダウン・リストからしきい値のタイプを選択します。

- 「絶対値」 を選択すると、しきい値はそれぞれ一対の値を持つレコードの数に基づいて設定されます。
- 「パーセント **(**全体**)**」には、Web グラフで示された値の各ペアの頻度の割合として、リンクで表されているケースの絶対数が表示されます。
- また、「パーセンテージ **(**小さいフィールド**/**値**)**」 と 「パーセンテージ **(**大きいフィールド**/**値**)**」 には、割合の評価に使うフィールドと値が示されます。例えば、「薬品」 フィールドに薬品 *Y* の値を持つレコードが 100 個あり、「血圧」 フィールドに低の値を持つレコードが 10 個しかない場合を考えてみましょう。このとき、薬品 *Y* と低の両方の値を持つレコードが 7 個あると、このレコードの割合は小さいフィールド (「血圧」 フィールド) または大きいフィールド (「薬品」 フィールド) のどちらを参照するかに応じて、それぞれ 70% または 7% となります。

注：Web グラフの場合、上記の 3 番目のオプションは利用できません。代わりに、「パーセンテージ **("**終点**"**フィールド**/**値**)**」 と 「パーセンテージ **("**始点**"**フィールド**/**値**)**」 を選択することができます。

強いリンクほど太い**:** これはデフォルトで選択されています。これが、フィールド間のリンクを表示する標準的な方法です。

弱い相関ほど太い**:** 太線で表示されているリンクの意味を逆にする場合に選択します。このオプションは、不正行為の検出や外れ値の調査などで頻繁に使用されます。

Web グラフの「オプション」タブ

Web グラフ・ノードの「オプション」タブには、出力グラフをカスタマイズするためのさまざまなオプションが用意されています。

リンク数**:**出力グラフに表示されるリンク数を制御するために使われるオプションを次に示します。「弱い関係ほど太い」 や 「強い **(**右の値超過**)**」 など、一部のオプションは、出力グラフ・ウィンドウでも使用することができます。最終的なグラフでスライダを使用して、表示するリンク数を調整することもできます。

- 表示するリンクの最大値**:**出力グラフに表示する最大リンク数を指定します。矢印を使用して、値を調節してください。
- 次の値より大きいリンクを表示**:**Web グラフにリンクを表示する場合の最小値を指定します。矢印を使用して、値を調節してください。

- すべてのリンクを表示:最小値や最大値の設定に関係なく、すべてのリンクを表示します。フィールド数が多い場合にこのオプションを選択すると、処理時間が増加する可能性があります。

少数レコードを破棄:リンクをサポートするレコード数が少ない場合、リンクを無視するときに選択します。「最小レコード/行」 フィールドに数値を入力して、このオプションのしきい値を設定します。

大量レコードを破棄:強くサポートされているリンクを無視するときに選択します。「最大レコード/行」フィールドに数値を入力します。

弱い (右の値未満):弱いリンク (点線) と通常のリンク (実線) のしきい値を示す数値を指定します。これより小さな値のリンクはすべて弱いリンクと見なされます。

強いリンク (以上): 強いリンク (太線) と通常のリンク (実線) のしきい値を示す数値を指定します。これより大きな値のリンクはすべて強いリンクと見なされます。

リンク・サイズ:リンクのサイズを制御するオプションを指定します。

- リンク・サイズは可変:このオプションを選択すると、実際のデータ値に基づいて接続強度の変動を反映する、リンク・サイズの範囲が表示されます。
- リンク・サイズを強い/通常/弱いで表示: このオプションを選択すると、3 種類の接続強度 (強い、中間、弱い) が表示されます。これらのカテゴリーのしきい値は、上記のほかにも最終的なグラフで指定することができます。

Web グラフ表示:Web グラフの表示の種類を選択します。

- サークル・レイアウト: 標準の Web グラフ表示を選択します。
- ネットワーク・レイアウト: もっとも強いリンクをグループ化するアルゴリズムの使用を選択します。このオプションは、空間的な差異や重み付けられた線を使用して、強いリンクを強調することを目的にしています。
- 方向付きレイアウト:「分布図」タブから 「終点フィールド」 選択をその方向に対するフォーカスとして使用する、Web グラフ表示を作成します。
- グリッドのレイアウト: 規則的なスペースを含むグリッド・パターンにレイアウトされた Web グラフ表示を作成します。

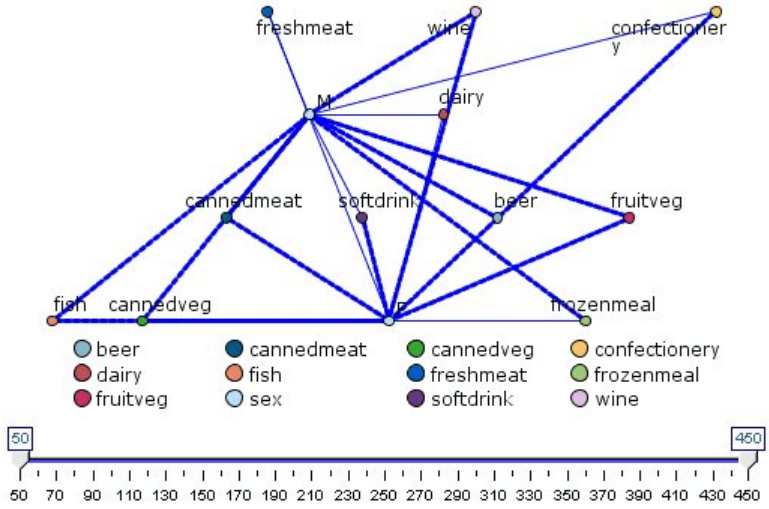


図 37. 冷凍食品と缶詰野菜から他の食料雑貨への強いリンクを表した *Web* グラフ

注: (Web グラフのスライダーか、Web グラフ・ノードの「オプション」タブの「次の値より大きいリンクを表示」コントロールのいずれかを使用して) 表示されているリンクをフィルタリングすると、表示されたままであるリンクすべてが単一の値であるという状態になる場合があります (つまり、それらのリンクは、Web グラフ・ノードの「オプション」タブの「弱い (右の値未満)」コントロールと「強い (右の値超過)」コントロールによって定義したすべて弱いリンク、すべて中間のリンク、またはすべて強いリンクのいずれかです)。これが発生した場合、すべてのリンクは Web グラフ出力に中程度の幅の行ですべて表示されます。

Web グラフの「外観」タブ

グラフ作成前に外観オプションを指定できます。

表題: グラフのタイトルに使用するテキストを入力します。

サブタイトル: グラフのサブタイトルに使用するテキストを入力します。

キャプション: グラフのキャプションに使用するテキストを入力します。

凡例を表示: 凡例の表示を指定することができます。多数のフィールドのある作図の場合、凡例を非表示にすると作図の外観を改善できます。

ラベルをノードとして使用: ラベルを隣接表示するのではなく、ラベル テキストを各ノード内に表示します。作図するフィールド数が少ない場合、グラフが読みやすくなります。

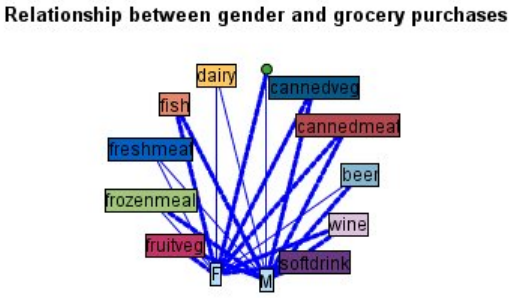


図 *38.* ラベルをノードとして表示する *Web* グラフ

Web グラフの使用方法

Web グラフ・ノードは、2 つ以上のシンボル値フィールドの値の相関の強さを表示するために使用されます。接続は、接続の強度を示すさまざまな種類の線としてグラフに表示されます。例えば、Web グラフ・ノードを使用して、コレステロール レベル、血圧、および患者の治療に効果的な薬品の相関を調べることができます。

- 強い接続は太い線で表されます。これは、2 つの値が強く関係しており、より詳細に調査する必要があることを示しています。
- 中程度の接続は、中間の線で表示されます。
- 弱い接続は点線で表されます。

- 2 つの値間に線がない場合は、これら 2 つの値が同じレコード内には完全に存在していないか、またはこれらの組み合わせの発生回数が、「Web グラフ・ノード」ダイアログ・ボックスで指定されたしきい値を下回っていることを意味しています。

Web グラフ・ノードを生成すると、さまざまなオプションを使用してグラフの表示を調整したり、詳細な分析のためのノードを生成したりできるようになります。

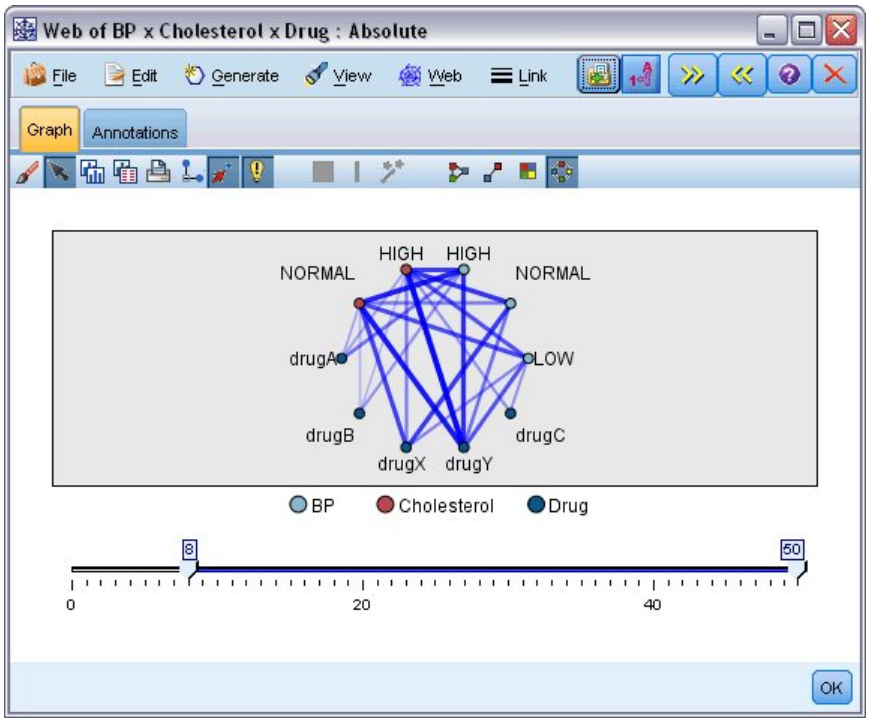


図 *39.* 正常な血圧と薬品 *X*、高コレステロールと薬品 *Y* などの強い相関を表す *Web* グラフ

MultiWeb グラフ・ノードと Web グラフ・ノードでは、次の操作ができます。

- Web グラフの表示レイアウトを変更する。
- ポイントを隠して、表示を単純化する。
- 線の種類を制御するしきい値を変更する。
- 「選択された」相関を示すために、値間の線を強調表示する。
- 1 つ以上の選択されたレコードに条件抽出ノードを生成するか、または Web グラフ内の 1 つ以上の相関に関連付けられたフラグ型フィールド作成ノードを生成する。

ポイントを調整するには

- 移動: ポイントをマウスでクリックし、新しい場所にドラッグして移動します。これによって、新しい場所を反映して Web グラフが再描画されます。
- 非表示: Web グラフのポイント上を右クリックし、コンテキスト・メニューから 「非表示」 または 「非表示にして再計算」 を選択して、ポイントを隠します。「非表示」 を選択すると、選択したポイントとこのポイントに関連付けられた線が隠されます。「非表示にして再計算」 を選択すると、変更を反映して Web グラフが再描画されます。このとき、手動による移動は取り消されます。
- 表示: グラフ・ウィンドウの「Web グラフ」メニューから、「すべて表示」 または 「すべて表示して再計算」 を選択して、隠したすべてのポイントを表示します。「すべて表示して再計算」 を選択すると、前に隠したすべてのポイントとその相関を含めて Web グラフが再描画されます。

線を選択または「強調表示」するには

選択した行は赤で強調表示されます。

1. 1 行を選択して、その行を左クリックします。
2. 複数の行を選択するには、次のいずれかを実行します。

- カーソルを使用して、選択したい行のポイントの周りで円を描きます。
- Ctrl キーを押しながら、選択する各行を左クリックします。

グラフの背景をクリックするか、グラフ・ウィンドウの Web メニューから 「選択解除」 を選択して、選択するすべての行の選択を解除します。

異なるレイアウトを使用して Web グラフを表示するには

「Web グラフ」メニューから、「サークル レイアウト」、「ネットワーク レイアウト」、「方向付きレイアウト」、または「グリッドのレイアウト」 を選択し、グラフのレイアウトを変更します。

リンク スライダをオンまたはオフにするには

「表示」メニューから、「リンク スライダ」を選択します。

単一の相関に対してレコードを選択したり、フラグを設定するには

1. 対象となる相関を表す線を右クリックします。
2. コンテキスト・メニューから 「リンクの条件抽出ノード生成」 または 「リンクのフィールド作成ノード生成」 を選択します。

適切なオプションと指定した条件を備えた条件抽出ノードまたはフィールド作成ノードが、自動的にストリーム領域に追加されます。

- 条件抽出ノードでは、指定された相関に該当するレコードがすべて選択されます。
- フィールド作成ノードでは、データ・セット全体のレコードにおいて、選択された相関が真 (true) であるかどうかを示すフラグが生成されます。フラグ・フィールドには、低_薬品 C や 薬品 C_低 など、相関関係にある 2 つの値を下線で連結した名前が付けられます。

相関グループのレコードを選択したり、フラグを設定するには

1. Web グラフで、対象となる相関を表す線を選択します。
2. グラフ・ウィンドウの「ノードの生成」メニューで、「条件抽出ノード **(AND)**」、「条件抽出ノード **(OR)**」、「フィールド作成ノード **(AND)**」、または 「フィールド作成ノード **(OR)**」 を選択します。

- "OR" ノードでは条件が分岐されます。つまり、このノードは選択したいずれかの相関に該当するレコードに適用されます。
- "AND" ノードでは条件が結合されます。つまり、このノードは選択したすべての相関に該当するレコードだけに適用されます。選択した相関が相互に排他的な場合は、エラーが発生します。

選択が完了すると、適切なオプションと指定した条件を備えた条件抽出ノードまたはフィールド作成ノードが、自動的にストリーム領域に追加されます。

注**:** (Web グラフのスライダーか、Web グラフ・ノードの「オプション」タブの「次の値より大きいリンクを表示」コントロールのいずれかを使用して) 表示されているリンクをフィルタリングすると、表示されたままであるリンクすべてが単一の値であるという状態になる場合があります (つまり、それらのリンクは、Web グラフ・ノードの「オプション」タブの「弱い **(**右の値未満**)**」コントロールと「強い **(**右の値超

過)」コントロールによって定義したすべて弱いリンク、すべて中間のリンク、またはすべて強いリンクのいずれかです)。これが発生した場合、すべてのリンクは Web グラフ出力に中程度の幅の行ですべて表示されます。

Web グラフのしきい値の調整

Web グラフを作成したら、最低限表示する線を変更するために、ツールバーやスライダを使用して、線のスタイルを制御するしきい値を調整することができます。また、ツールバーの黄色い二重矢印ボタンをクリックして「Web グラフ」ウィンドウを展開し、しきい値に関する付加オプションを表示することもできます。次に、「コントロール」 タブをクリックして、付加オプションを表示します。

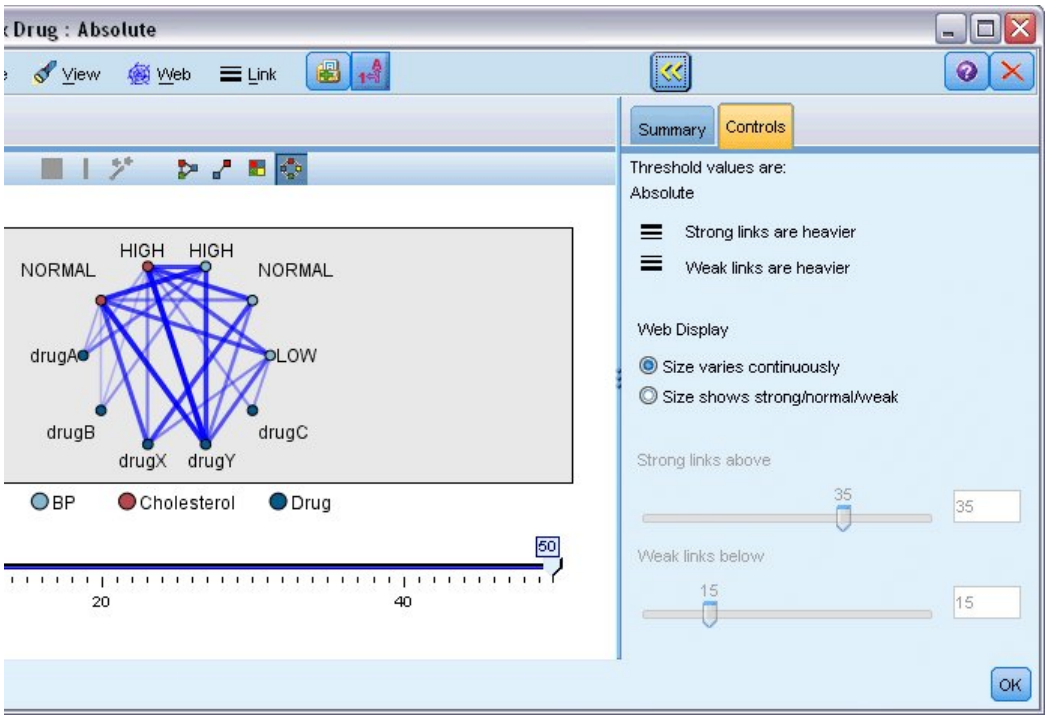


図 *40.* ウィンドウを展開して表示およびしきい値オプションを表示

しきい値**:**「Web グラフ・ノード」ダイアログ・ボックスでの作成中に選択したしきい値のタイプが表示されます。

強いリンクほど太い**:** これはデフォルトで選択されています。これが、フィールド間のリンクを表示する標準的な方法です。

弱い相関ほど太い**:** 太線で表示されているリンクの意味を逆にする場合に選択します。このオプションは、不正行為の検出や外れ値の調査などで頻繁に使用されます。

Web グラフ表示**:** 出力グラフ中のリンクのサイズを制御するオプションを指定します。

- サイズは可変**:** このオプションを選択すると、実際のデータ値に基づいて接続強度の変動を反映する、リンク・サイズの範囲が表示されます。
- サイズを強い/通常/弱いで表示**:** このオプションを選択すると、3 種類の接続強度 (強い、中間、弱い) が表示されます。これらのカテゴリーのしきい値は、上記のほかにも最終的なグラフで指定することができます。

強いリンク **(**以上**):** 強いリンク (太線) と通常のリンク (実線) のしきい値を示す数値を指定します。これより大きな値のリンクはすべて強いリンクと見なされます。スライダを使用して値を調整するか、またはフィールドに値を入力します。

弱い **(**右の値未満**):**弱いリンク (点線) と通常のリンク (実線) のしきい値を示す数値を指定します。これより小さな値のリンクはすべて弱いリンクと見なされます。スライダを使用して値を調整するか、またはフィールドに値を入力します。

Web グラフのしきい値を調整したら、Web グラフ ツールバーにある Web グラフ メニューを使用して、新しいしきい値で Web グラフを再計算または再描画することができます。もっともよく意味のあるパターンが現れる設定を見つけたら、グラフ・ウィンドウの「Web グラフ」メニューから 「親ノードの更新」 を選択して、Web グラフ・ノード (親ノードとも呼ばれます) の元の設定を更新することができます。

Web グラフの概要の作成

強い、中間、および弱いリンクを記載した Web グラフの概要ドキュメントを作成するには、ツールバーの黄色い二重矢印ボタンをクリックして「Web グラフ」ウィンドウを展開します。次に、「要約」 タブをクリックして、各種類のリンクのテーブルを表示します。テーブルは、トグル ボタンを使用して展開したり、省略することができます。

要約を印刷するには、Web グラフ・ウィンドウのメニューから次を選択します。

「ファイル」 > 「要約の印刷」

評価ノード

評価ノードを使用すると、簡単に予測モデルを評価および比較して、アプリケーションに最適なモデルを選択できます。評価グラフでは、各モデルの特定の結果の予測方法が示されます。また評価グラフでは、レコードは予測フィールドと予測の確信度に基づいてソートされて等しいサイズのグループ (分位) に分割され、分位ごとにビジネスに関する基準の値が高い方から順番に作図されます。作図には、複数のモデルが異なる線で示されます。

結果は、特定の値または値の範囲をヒットとして定義することで処理されます。通常、ヒットはある種の成功 (顧客への販売など) や対象となるイベント (特定の医療診断など) を示します。ダイアログ・ボックスの「オプション」タブでヒット基準を定義したり、次のようなデフォルトのヒット基準を使用することができます。

- フラグ型出力フィールドの場合、ヒットはそのまま真の値に対応しています。
- 一方名義型出力フィールドの場合は、セットの最初の値がヒットを表します
- 連続型出力フィールドの場合、フィールドの範囲の中間より大きい値がヒットになります。

評価グラフには 6 種類あり、それぞれ強調される評価基準は異なります。

ゲイン・グラフ

ゲインは、各分位で発生した総ヒット数の割合として定義されます。ゲインは、(分位内のヒット数 / 総ヒット数) × 100% として計算されます。

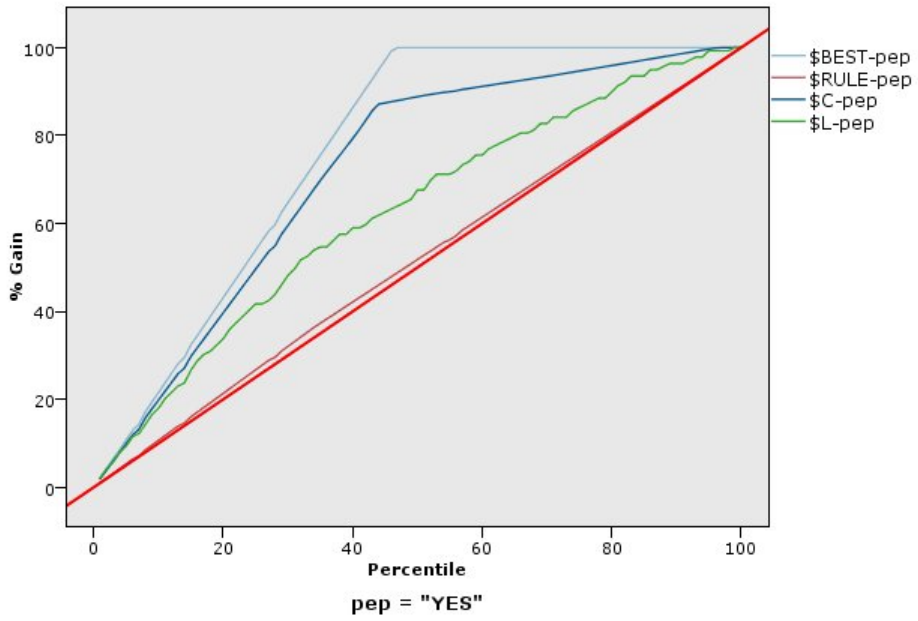


図 *41.* 基準、ベスト・ライン、およびビジネス・ルールが表示されたゲイン・グラフ (累積)

リフト・グラフ

リフトでは、各分位でヒットしたレコードの割合 (パーセント) が、トレーニング・データ内の全ヒットの割合と比較されます。これは、(分位内のヒット数 / 分位内のレコード数) / (総ヒット数 / 総レコード数) で算出されます。

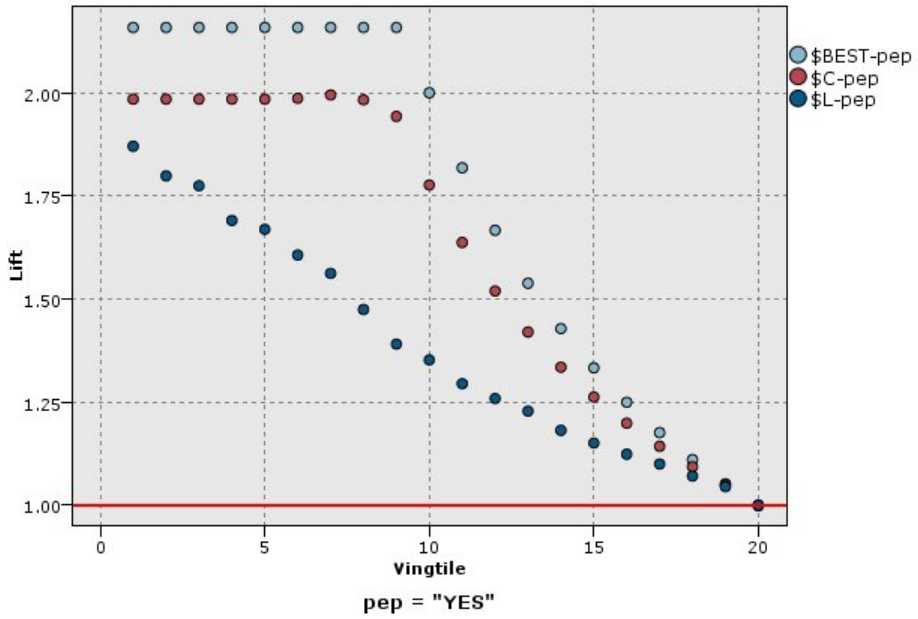


図 *42.* ポイントおよびベスト・ラインを使用するリフト・グラフ (累積)

回答グラフ

回答は、分位内のヒットしたレコードの単純な割合です。回答は、(分位内のヒット数 / 分位内のレコード数) × 100% として計算されます。

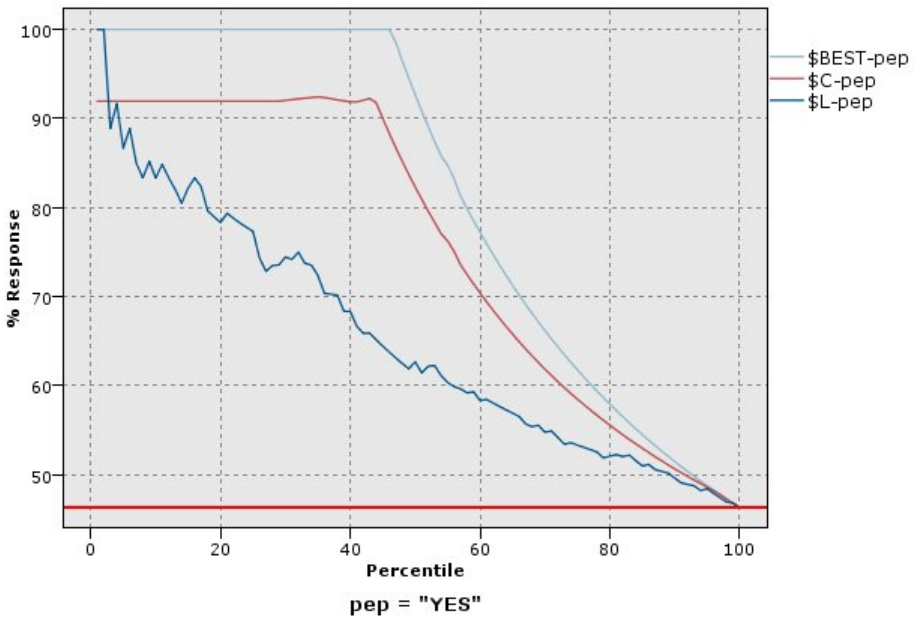


図 *43.* ベスト・ラインが表示された回答グラフ *(*累積*)*

利益グラフ

プロフィットは、各レコードの収益から、そのレコードのコストを引いた値と等しくなります。分位のプロフィットは、その分位の全レコードのプロフィットを合計したものです。収益はヒットだけに適用されることを前提としますが、コストはすべてのレコードに適用されます。また、プロフィットとコストは固定にすることも、データのフィールドで定義することもできます。プロフィットは、(分位内のレコードの合計収入 – 分位内のレコードの合計コスト) として計算されます。

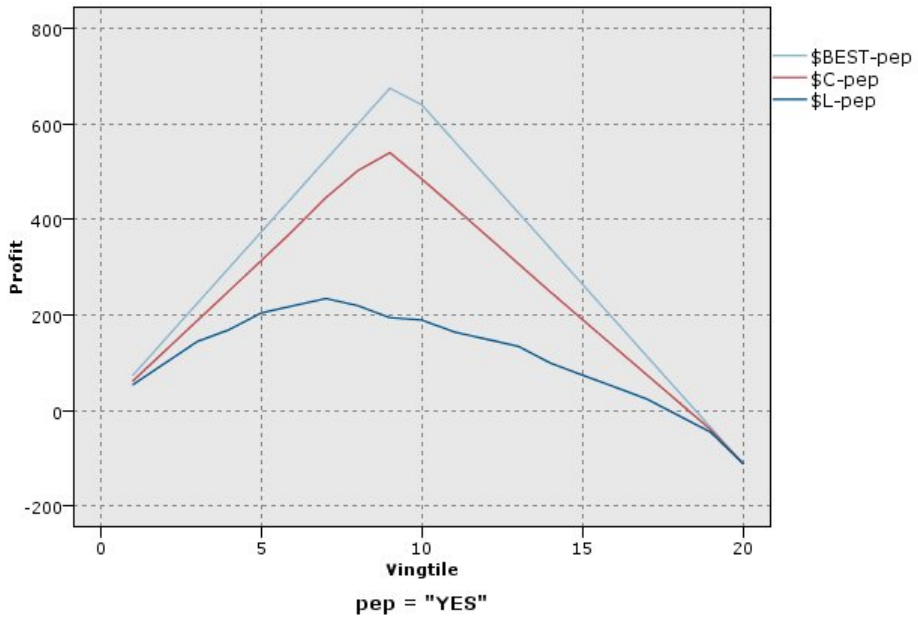


図 *44.* ベスト・ラインが表示された利益グラフ (累積)

ROI グラフ

ROI (投資収益率) は、収益とコストを定義するという点でプロフィットに似ています。ROI とは、分位のプロフィットとコストの比較です。ROI は (分位のプロフィット / 分位のコスト) × 100% として計算されます。

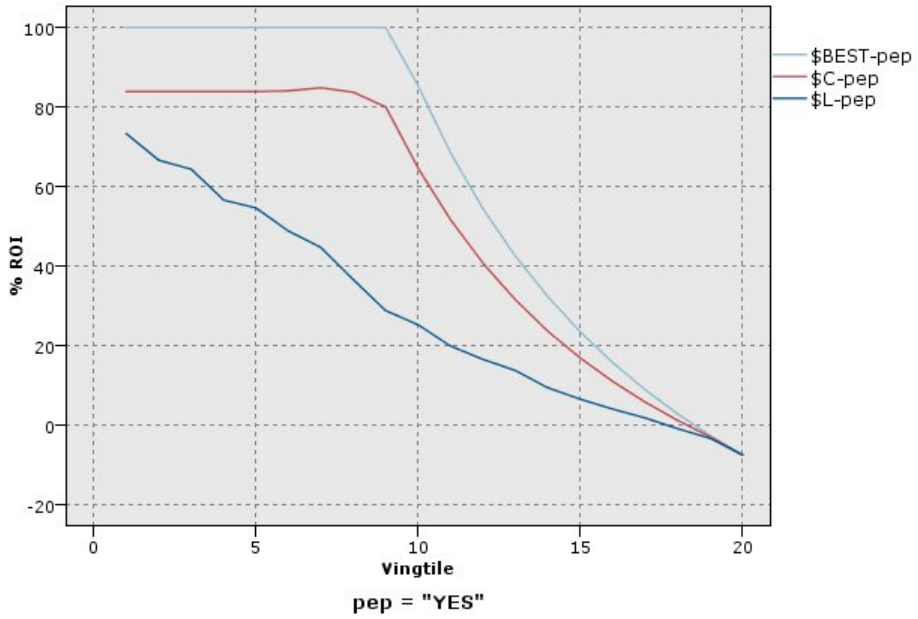


図 *45.* ベスト・ラインが表示された *ROI* グラフ (累積)

ROC グラフ

ROC (受信者操作特性) はバイナリ分類子と併用する必要があります。ROC を使用すると、分類子の視覚化、編成、パフォーマンスに基づく選択を行うことができます。ROC グラフは、分類子の真陽性率 (感度) を偽陽性率に対してプロットします。ROC グラフは、利得 (真陽性) とコスト (偽陽性) の相対的なトレードオフを図示します。真陽性は、ヒットでありかつヒットと分類されたインスタンスです。したがって、真陽性率は「真陽性の数/実際にヒットであったインスタンスの数」で計算されます。偽陽性は、外れであるがヒットと分類されたインスタンスです。したがって、偽陽性率は「偽陽性の数/実際に外れであったインスタンスの数」で計算されます。

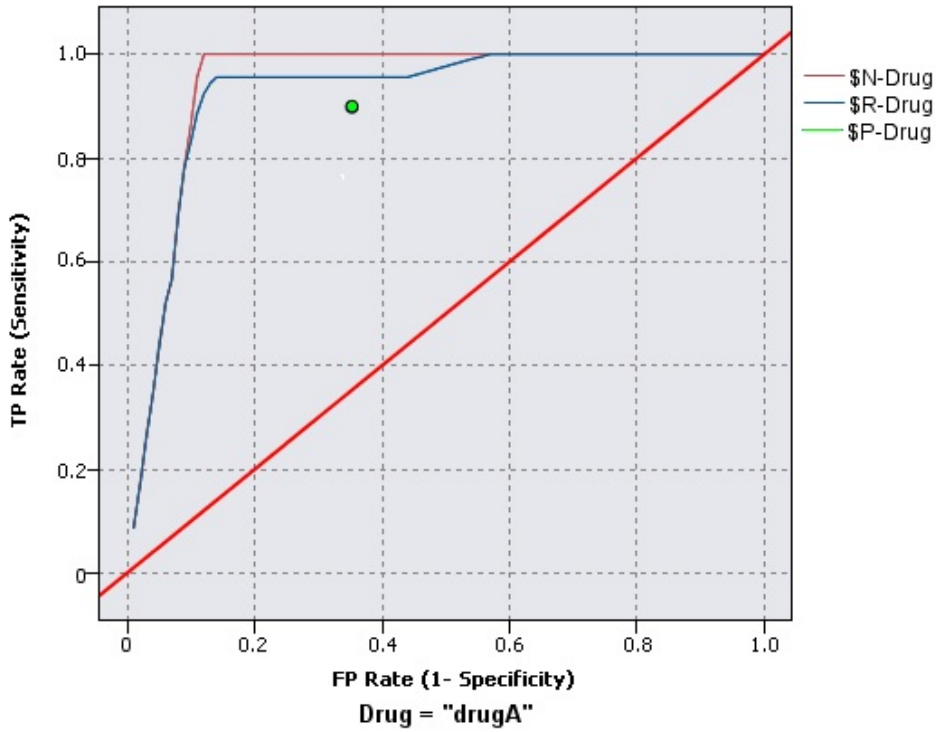


図 *46.* ベスト・ラインが表示された *ROC* グラフ

評価グラフは、各ポイントが対応する分位とそれより上位にあるすべての分位の値と等しくなるように、累積で表すこともできます。累積グラフは、通常モデルの全体的な性能を表す場合場合に役に立ちます。一方、非累積グラフは、そのモデルにおける特定の問題領域を表す場合に役に立ちます。

評価の「作図」タブ

グラフの種類。次に示すタイプの 1 つを選択します。「ゲイン」、「回答」、「リフト」、「プロフィット」、「**ROI**」(投資収益率)、「**ROC**」(受信者操作特性) のうち、いずれかのタイプを選択します。

累積作図。累積グラフを作成する場合に選択します。累積グラフの値は、各分位とそれより上位にあるすべての分位に対して作図されます。(ROC グラフの場合、「累積プロット」は使用できません。)

ベースラインを含める。プロットにベースラインを含める場合に選択します。ベースラインは、ヒットが完全にランダムに分布し、確信度と無関係になる状態を示します。(プロフィット・グラフおよび ROI グラフの場合、「ベースラインを含める」は使用できません。)

ベスト・ラインを含める。作図にベスト・ラインを含めて、完全な確信度 (ヒットがケースの 100%) を示す場合に選択します。(ROC グラフの場合、「ベストラインを含める」は使用できません。)

すべてのグラフ タイプに利益基準を使用する。評価測定の計算時に、通常のヒット数ではなく、利益基準 (コスト、収益、および重み) を使用する場合に選択します。特定の数値目標があるモデル (例えば、オファーに対応して顧客から得られる利益を予測するモデルなど) の場合、この目標フィールドの値により、ヒット数より正確にモデルのパフォーマンスが測定できます。ゲイン・グラフ、回答グラフ、およびリフト・グラフの場合は、このオプションを選択すると、「コスト」、「収益」、および「重み」のフィールドが使用可能になります。これら 3 つのグラフ・タイプで利益基準を使用するためには、「収益」を目標フィールドに設定して「コスト」を 0.0 に設定し、利益が収益と同じになるようにし、ヒット条件として「真」が定義されたユーザーを指定して、すべてのレコードがヒットとしてカウントされるようにすることをお勧めします。(ROC グラフの場合、「すべてのグラフ タイプに利益基準を使用する」は使用できません。)

次を使用する予測済み/予測フィールドの検出。「モデル出力フィールドのメタデータ」を選択してそれらのメタデータを使用するグラフで予測済みフィールドを検索するか、「フィールド名の形式」を選択して名前によってフィールドを検索します。

散布図スコア・フィールド。このチェック・ボックスを選択すると、スコア・フィールド・ピッカーが有効になります。次に、1 つ以上の範囲、または連続したスコア フィールド (厳密には予測モデルではないが、ヒットになる傾向の点でレコードをランク付けするのに役立つ可能性のあるフィールド) を選択します。評価ノードでは、1 つ以上のスコア フィールドの任意の組み合わせを 1 つ以上の予測モデルと比較できます。典型的な例は、複数の RFM フィールドを最適な予測モデルと比較することです。

目標: フィールド・ピッカーを使用して対象フィールドを選択します。2 つ以上の値を持つインスタンス化されたフラグまたは名義型フィールドを選択してください。

注 : この対象フィールドは、スコア フィールド (予測モデルが独自の対象を定義) のみに該当するため、「オプション」タブでヒット条件がユーザー設定されている場合は無視されます。

データ区分データによる分割。レコードを、学習、テスト、および検定用の各サンプルに分割するためにデータ区分フィールドが使用される場合、このオプションを選択すると、各データ区分ごとに、別々の評価グラフが表示されます。 詳しくは、トピック 185 ページの『データ区分ノード』を参照してください。

注 : データ区分を分割する場合、データ区分フィールドにあるヌル値を持つレコードは、評価から除外されます。データ区分ノードは、ヌル値を生成しないため、データ区分ノードを使用している場合は、問題になりません。

作図。ドロップダウン・リストから、グラフにプロットする分位のサイズを選択します。「**4** 分位」、「**5** 分位」、「**10** 分位」、「**20** 分位」、「**100** 分位」、または 「**1000** 分位」 を選択します。(ROC グラフの場合、「プロット」は使用できません。)

スタイル。「線」または「ポイント」を選択します。

ROC グラフを除き、いずれのタイプのグラフにも上記以外のコントロールが用意されており、コスト、収益、および重みを指定することができます。

- コスト : 各レコードに関連付けるコストを指定します。「固定」 または 「変数」 を選択することができます。固定コストの場合はコストの値を指定してください。可変コストの場合は、フィールド・ピッカー・ボタンをクリックして、コスト・フィールドとして使用するフィールドを選択してください。(「コスト」は、ROC グラフには使用できません。)
- 収益 : ヒットを表し各レコードに関連付ける収益を指定します。「固定」 または 「変数」 を選択することができます。固定収益の場合は収益値を指定してください。可変収益の場合は、フィールド・ピッカー・ボタンをクリックして、収益フィールドとして使用するフィールドを選択してください。(「収益」は、ROC グラフには使用できません。)

- 重み**:** データのレコードが複数のユニットからなる場合は、出現頻度の高い重みを使用して結果を調整できます。「固定」 または 「変数」 を選択して、各レコードに関連付ける重みの種類を指定します。重みを固定する場合は、重みの値 (レコードごとのユニット数) を指定してください。重みを変数にする場合は、フィールド・ピッカー・ボタンをクリックして、重みフィールドとして使用するフィールドを選択してください。(「重み」は、ROC グラフには使用できません。)

評価の「オプション」タブ

評価グラフノードの「オプション」タブでは、グラフに表示するヒット、スコアリング基準、およびビジネス・ルールなどを定義することができます。また、モデルの評価結果をエクスポートするためのオプションも設定することができます。

ユーザー定義のヒット **:**ヒットを示す条件を自分で指定する場合に選択します。このオプションは、対象フィールドの種類および値の並びから推論する代わりに対象の結果を定義する場合に役立ちます。

- 条件**:** 「ユーザー定義のヒット」を選択した場合は、ヒット条件の CLEM 式を指定する必要があります。例えば `@TARGET = "YES"` は、「*Yes*」の値を持つ対象フィールドを評価でヒットとしてカウントすることを示す有効な条件です。指定された条件は、すべての対象フィールドに対して使用されます。条件を作成するには、フィールドに式を直接入力するか、または Clem 式ビルダーを使用して条件式を生成します。データがインスタンス化されている場合、Clem 式ビルダーから直接値を挿入することができます。

ユーザー定義のスコア **:**分位に割り当てる前に、ケースのスコアリングに使用する条件を指定する場合に選択します。デフォルトのスコアは、予測値と確信度から算出されます。「式」フィールドを使用して、スコアリング式を作成します。

- 式 **:**スコアリングに使用する CLEM 式を指定します。例えば、0 から 1 の範囲の数値出力を、低い値の方が高い値よりも良好であるように順序付ける場合、ヒットを `@TARGET < 0.5`、関連するスコアを `1 ? @PREDICTED` と定義することができます。スコア式の結果は数値でなければなりません。条件を作成するには、フィールドに式を直接入力するか、または Clem 式ビルダーを使用して条件式を生成します。

ビジネス ルールを含める **:** 目的の基準を表す規則条件を指定する場合に選択します。例えば、`mortgage = "Y" and income >= 33000` であるすべてのケースのルールを表示することができます。ビジネス・ルールは、グラフに描画され、キーにルールとしてラベル付けされます。 (ROC グラフの場合、「ビジネス ルールを含める」はサポートされません。)

- 条件**:** 出力グラフのビジネス・ルールとして定義するために使用する CLEM 式を指定します。フィールドに式を直接入力するか、または Clem 式ビルダーを使用して条件式を生成します。データがインスタンス化されている場合、Clem 式ビルダーから直接値を挿入することができます。

結果をファイルへエクスポート **:**区切り文字で区切られたモデルの評価結果をテキスト・ファイルへエクスポートする場合に選択します。このファイルを読み込んで、計算した値の特殊な分析を行うことができます。次のエクスポート・オプションを設定します。

- ファイル名**:** 出力ファイルのファイル名を入力します。「**...**」 ボタンを使用して、目的のフォルダーを指定することもできます。
- 区切り文字**:** カンマやスペースなど、フィールドの区切り文字として使用する文字を入力します。

フィールド名を含める**:** 出力ファイルの最初の行にフィールド名を表示する場合に選択します。

各レコードの後に改行を入れる**:** 各レコードを新しい行で表示する場合に選択します。

評価の「外観」タブ

グラフ作成前に外観オプションを指定できます。

表題: グラフのタイトルに使用するテキストを入力します。

サブタイトル: グラフのサブタイトルに使用するテキストを入力します。

テキスト。自動生成されたテキスト ラベルを受け入れるか、「ユーザー設定」を選択してラベルを指定します。

X ラベル: 自動的に生成された *x* 座標 (水平) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

Y ラベル: 自動的に生成された *y* 座標 (垂直) ラベルを承認するか、「ユーザー設定」 を選択してユーザー設定ラベルを指定します。

グリッドの表示: デフォルトで選択されているこのオプションは、散布図またはグラフの背景にグリッドを表示する場合に使用します。グリッドを使用すると、領域やバンドの分割点を簡単に決めることができます。グリッド線は、グラフの背景が白でない限り、常に白で表示されます。背景が白の場合は、灰色で表示されます。

モデル評価の結果の読み込み

評価グラフの解釈は、ある程度グラフの種類によって異なりますが、すべての評価グラフに共通する特性もあります。累積グラフの場合、特にグラフの左側において、上部の線ほどすぐれたモデルであることを意味しています。また、複数モデルを比較するときに線がクロスすることがよくあります。つまり、グラフの一部であるモデルの線が上になっており、グラフの別の部分で異なるモデルの線が上になっているような場合です。このような場合に使用するモデルを選択するときは、サンプルのどの部分が必要かを考慮 (*x* 軸上のポイントを定義) しなければなりません。

ほとんどの非累積グラフは非常に似ています。すぐれたモデルの場合、非累積グラフはグラフの左側が高く、右側は低くなります(非累積グラフがのこぎり形になっている場合は、作図する分位数を減らしてグラフを再実行すると、滑らかにすることができます)。グラフの左側が落ち込んでいる場合や右側が鋭くとがっている場合は、モデルのその領域の予測精度が低いことを意味しています。また、グラフ全体を通して線が平坦な場合は、モデルからはほとんど情報が得られないことを意味しています。

ゲイン・グラフ。累積ゲイン・グラフは、常に左から右に向かって進み、0% で始まり 100% で終わります。適切なモデルの場合、ゲイン・グラフは 100% に向けて急勾配で上昇し、その後、水平状態になります。左下から右上に 45 度の線を描くモデルからは情報が得られません (このモデルは、「ベースラインを含める」 を選択した場合にグラフに表示されます)。

リフト・グラフ。累積リフト・グラフは、左から右に向かって進み、1.0 より上から始まって 1.0 に近づくにつれ徐々に下降します。グラフの右端がデータ・セット全体を表します。つまり、データ全体のヒット数に対する累積分位内のヒット数の比率は 1.0 です。すぐれたモデルの場合、リフト・グラフは左端で 1.0 のかなり上から始まったまま、右に移動しても高い状態を維持し、グラフの右側で 1.0 に向かって急激に下降します。グラフ全体にわたって線が 1.0 付近に留まっているモデルからは、情報が得られません(「ベースラインを含める」 を選択すると、グラフの 1.0 のレベルに水平線が参照線として表示されます)。

回答グラフ。累積回答グラフはリフト・グラフと非常に似ていますが、尺度が異なります。回答グラフは、通常 100% 付近から始まり、グラフの右端で全体的な応答率 (総ヒット数 / 総レコード数) に達するまで徐々に下降します。すぐれたモデルの場合、線は左側の 100% 付近から始まったまま、右に移動しても高

い状態を維持し、グラフの右側で全体的な応答率に向かって急激に下降します。グラフ全体にわたって線が全体的な応答率のレベルに留まっているグラフからは、情報が得られません(「ベースラインを含める」 を選択すると、グラフの全体的な応答率のレベルに水平線が参照線として表示されます)。

利益グラフ。累積利益グラフでは、左から右に向かって、選択したサンプルのサイズの増加に伴う利益の合計が示されます。通常、利益グラフは 0 付近から始まり、中央の山形または台形の部分に向かって徐々に増加し、グラフの右端に向かって下降します。すぐれたモデルの場合、グラフの中央付近にくっきりとした山形が見られます。線が比較的真っ直ぐで、適用するコストや収益の構造によって上昇または下降したり、平坦になったりするモデルからは、情報が得られません。

ROI グラフ。累積 ROI (投資収益) グラフは、回答グラフおよびリフト・グラフと似ていますが、尺度が異なります。ROI グラフは通常 0% 付近から始まり、データ・セット全体の ROI に達するまで徐々に下降します (負の値になることもあります)。すぐれたモデルの場合、線は 0% 付近から始まったまま、右に移動しても高い状態を保ち、グラフの右側で全体的な ROI 値に向かって急激に下降します。線が全体的な ROI 値の付近に留まっているモデルからは情報が得られません。

ROC グラフ。 ROC 曲線は一般に、累積ゲイン・グラフの形状を持ちます。曲線は座標 (0,0) で始まり、左から右へ進んで座標 (1,1) で終わります。グラフが座標 (0,1) に向かって急激に上昇した後に水平になる場合は、よい分類子であることを示します。インスタンスを無作為にヒットまたは外れに分類するモデルは、左下から右上に向かう対角線をたどります (対角線は、「ベースラインを含める」を選択するとグラフに表示されます)。モデルの確信度フィールドが提供されていない場合、モデルは単一のポイントとしてプロットされます。最適な分類しきい値を持つ分類子は、グラフの座標 (0,1)、つまり左上の近くに位置します。この位置は、正しくヒットと分類されるインスタンスが多く、誤ってヒットと分類されるインスタンスが少ないことを表します。対角線より上のポイントは、良好な分類結果を表します。対角線より下のポイントは、インスタンスを無作為に分類した場合よりも悪い分類結果を表します。

評価グラフの使用方法

マウスを使った評価グラフの探索は、ヒストグラムや集計棒グラフの場合と同じように行うことができます。*x* 軸は、20 分位や 10 分位など、指定した分位によるモデル・スコアを表しています。

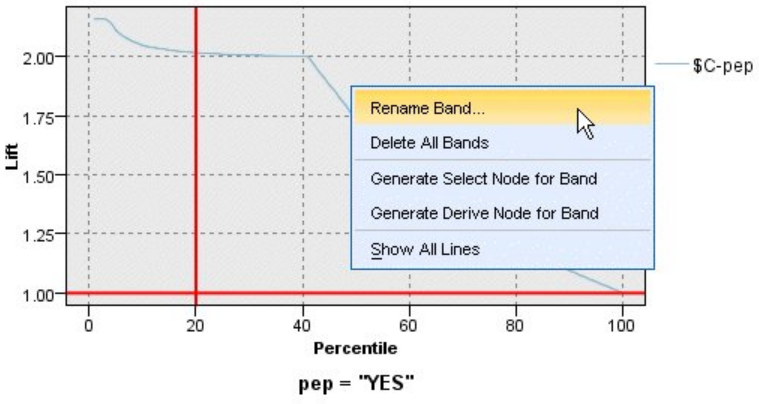


図 *47.* 評価グラフの作業

分割アイコンを使用して軸を等しいバンドに自動分割するオプションを設定して、ヒストグラムのように *x* 軸をバンドに分割することができます。詳しくは、トピック 293 ページの『グラフの検証』を参照してください。バンドの境界を手作業で編集するには、「編集」メニューの 「グラフ バンド」 を選択します。

評価グラフを作成し、バンドを定義して結果を調べたら、「ノードの生成」メニューとコンテキスト・メニューのオプションを使用して、グラフ中の選択項目に基づいて自動的にノードを生成できるようになります。 詳しくは、トピック 300 ページの『グラフからのノードの生成』を参照してください。

評価グラフからノードを生成する場合、グラフで利用できるすべてのモデルから 1 つのモデルを選択するように指示するメッセージが表示されます。

モデルを選択して、「**OK**」 をクリックすると、ストリーム領域に新しいノードが生成されます。

マップ視覚化ノード

マップ視覚化ノードは、複数の入力接続を受け入れて、地理空間データを一連の層としてマップに表示することができます。各層は単一の地理空間フィールドです。例えば、基本層を国のマップとし、その上に道路の層、川の層、町の層を設けることができます。

通常、大部分の地理空間データセットは単一の地理空間フィールドを含みますが、1 つの入力に複数の地理空間フィールドが存在する場合は、表示するフィールドを選択することができます。同じ入力接続からの 2 つのフィールドを同時に表示することはできません。しかし、入力接続をコピーして貼り付け、それぞれの異なるフィールドを表示することは可能です。

マップ視覚化の「プロット」タブ

層(L)

このテーブルには、マップノードへの入力に関する情報が表示されます。層の順序は、マップ プレビューとノード実行時の視覚的出力の両方における層の表示順序を決定します。テーブルの一番上の行が「一番上」の層であり、一番下の行が「基本」層です。つまり、マップの各層は、テーブルでその直下にある層の前に表示されます。

注**:** テーブルの層に 3 次元の地理空間フィールドが含まれる場合は、x 軸と y 軸のみがプロットされます。z 軸は無視されます。

名前　各層の名前は自動的に作成されます。作成に使用される形式は `tag[source node:connected node]` です。デフォルトではタグが数値で表されます。接続される最初の入力が 1 で表され、2 番目の入力が 2 で表されます (以下同様)。必要な場合は、「レイヤーの編集」ボタンを押し、「マップ層オプションの変更」ダイアログボックスでタグを変更します。例えば、タグを「roads」や「cities」に変更して、データ入力を反映させることができます。

データ型

層として選択する地理空間フィールドの尺度タイプを示すアイコンが表示されます。地理空間の尺度タイプを持つ複数のフィールドが入力データに含まれる場合、デフォルトの選択では以下のソート順序が使用されます。

1. ポイント
2. 行ストリング
3. 多角形
4. 複数点
5. 複数行ストリング
6. 多角形群

注: 尺度タイプが同じ 2 つのフィールドがある場合、デフォルトでは (名前のアルファベット順で) 最初のフィールドが選択されます。

記号

注: この列は、ポイントおよび複数点のフィールドの場合にのみ指定します。
ポイントおよび複数点のフィールドの場合に使用する記号が表示されます。必要な場合は、「レイヤーの編集」ボタンを押し、「マップ層オプションの変更」ダイアログボックスで記号を変更します。

色　マップの層を表すために選択されている色が表示されます。必要な場合は、「レイヤーの編集」ボタンを押し、「マップ層オプションの変更」ダイアログボックスで色を変更します。この色は、尺度タイプによって異なる項目に適用されます。

- ポイントまたは複数点の場合は、この色が層の記号に適用されます。
- 行ストリングおよび多角形の場合は、この色が形状全体に適用されます。多角形の輪郭は常に黒になります。この列に表示されている色は、その形状の塗りつぶしに使用される色です。

プレビュー

このペインには、「層」テーブルで現在選択されている入力のプレビューが表示されます。プレビューには、層の順序、記号、色など、層に関連付けられているすべての表示設定が反映され、設定が変更されると (可能であれば) 表示が更新されます。ストリーム (例えば層として使用する地理空間フィールド) で詳細情報を変更したり、関連した集計関数などの詳細を変更したりした場合は、「データの更新」ボタンをクリックしてプレビューを更新する必要がある場合があります。

ストリームの実行前に、「プレビュー」を使用して表示設定の値を設定します。大規模なデータセットの使用によって遅延時間が発生するのを避けるために、プレビューでは各層をサンプリングし、先頭から 100 件のレコードを使用して表示を作成します。

マップ層の変更

「マップ層オプションの変更」ダイアログボックスを使用すると、マップ視覚化ノードの「プロット」タブに表示される層の各種詳細情報を変更できます。

入力の詳細

タグ　デフォルトでは、タグは数値です。この数値を分かりやすいタグに置き換えて、マップの層を区別しやすくすることができます。例えば、タグをデータ入力の名前 (「Cities」など) にすることができます。

レイヤー フィールド**(L)**

入力データに複数の地理空間フィールドがある場合は、このオプションを使用して、マップで層として表示するフィールドを選択します。

デフォルトでは、選択可能な層は以下のソート順序で並んでいます。

- ポイント
- 行ストリング
- 多角形
- 複数点
- 複数行ストリング
- 多角形群

表示の設定

16 進法データ分割の使用**(X)**

注**:** このオプションは、ポイントおよび複数点のフィールドのみに影響します。

六角ビン分割では、(x 座標および y 座標を基準として) 近接ポイントを単一のポイントに結合してマップに表示します。この単一ポイントは六角形として表示されますが、実質的には多角形としてレンダリングされます。

この六角形は多角形としてレンダリングされるため、六角ビン分割がオンに設定されたポイント フィールドはすべて多角形として処理されます。そのため、マップ ノードのダイアログ ボックスで「タイプによる並び順」を選択した場合には、六角データ分割が適用されたポイント層がすべて多角形の層の上にレンダリングされますが、行ストリングおよびポイントの層より下にレンダリングされます。

複数点フィールドに六角データ分割を使用すると、最初に、複数点の値をデータ分割して中心点を計算する方法でフィールドがポイント フィールドに変換されます。中心点は六角データ分割の計算に使用されます。

集計

注**:** 「**16** 進法データ分割の使用」チェック ボックスを選択し、さらに「オーバーレイ」も選択した場合にのみ、この列を使用できます。

六角ビン分割を使用するポイント層に対して「オーバーレイ」フィールドを選択した場合は、六角形の中にあるすべてのポイントについて、そのフィールドのすべての値を集計する必要があります。マップに適用するすべてのオーバーレイ フィールドに使用する集計関数を指定します。使用できる集計関数は尺度タイプによって異なります。

- 実数または整数のストレージを持つ連続型尺度タイプ用の集計関数:
  - 合計
  - 平均値
  - 最小値
  - 最大値
  - 中央値
  - 第 1 四分位数
  - 第 3 四分位数
- 時間、日付、またはタイムスタンプのストレージを持つ連続型尺度タイプ用の集計関数:
  - 平均値
  - 最小値
  - 最大値
- 名義型またはカテゴリ型の尺度タイプ用の集計関数:
  - モード
  - 最小値
  - 最大値
- フラグ型尺度タイプ用の集計関数:
  - いずれかが真の場合は真
  - いずれかが偽の場合は偽

色

このオプションを使用して、地理空間フィールドのすべてのフィーチャーに適用する標準の色を選択するか、オーバーレイ フィールド (データの他のフィールドの値に基づいてフィーチャーに色を付けます) を選択します。

「標準」を選択した場合は、「ユーザー オプション」ダイアログ ボックスの「表示」タブで、「グラフ カテゴリーの色順序」ペインに表示される色のパレットから色を選択できます。

「オーバーレイ」を選択した場合は、「レイヤー フィールド」として選択した地理空間フィールドを含むデータ ソースから、任意のフィールドを選択できます。

- 名義型またはカテゴリ型のオーバーレイ フィールドの場合、選択に使用できる色パレットは、「標準」の色オプションの場合に表示されるものと同じです。
- 連続型および順序型のオーバーレイ フィールドの場合は、2 つ目のドロップダウン リストが表示されます。そのドロップダウン リストで色を選択します。色を選択すると、連続型または順序型のフィールドの値に従ってその色の彩度に変化をもたせることによりオーバーレイが適用されます。最も高い値の場合は、ドロップダウン リストから選択された色が使用され、値が低くなるに従ってそれに応じた低い彩度で値が表示されます。

記号

注: ポイントおよび複数点の尺度の場合にのみ使用可能です。

このオプションを使用して、「標準」の記号を適用する (地理空間フィールドのすべてのレコードに適用されます) か、「オーバーレイ」の記号を適用する (データの他のフィールドの値に基づいてポイントの記号アイコンが変更されます) かを選択します。

「標準」を選択した場合は、マップ上のポイント データを表すために、ドロップダウン リストからデフォルト記号のうちの 1 つを選択できます。

「オーバーレイ」を選択した場合は、「レイヤー フィールド」として選択した地理空間フィールドを含むデータ ソースから、名義型、順序型、またはカテゴリ型のフィールドのうち、任意のフィールドを選択できます。オーバーレイ フィールドの値ごとに異なる記号がマップに表示されます。

例えば、データに店舗の所在地を表すポイント フィールドが含まれており、オーバーレイは店舗タイプのフィールドになる場合があります。この例では、マップですべての食品店を十字記号で識別し、すべての電器店を四角形の記号で識別することができます。

サイズ

注: ポイント、複数点、行ストリング、および複数行ストリングの尺度タイプの場合にのみ使用可能です。

このオプションを使用して、「標準」のサイズを適用する (地理空間フィールドのすべてのレコードに適用されます) か、「オーバーレイ」のサイズを適用する (データの他のフィールドの値に基づいて記号アイコンのサイズや線の幅が変更されます) かを選択します。

「標準」を選択した場合は、ピクセル幅の値を選択できます。選択可能なオプションは 1、2、3、4、5、10、20、または 30 です。

「オーバーレイ」を選択した場合は、「レイヤー フィールド」として選択した地理空間フィールドを含むデータ ソースから、任意のフィールドを選択できます。選択したフィールドの値に応じて線の幅またはポイントの厚みが変化します。

透過度

このオプションを使用して、「標準」の透過度を適用する (地理空間フィールドのすべてのレコードに適用されます) か、「オーバーレイ」の透過度を適用する (データの他のフィールドの値に基づいて記号、線、または多角形の透過度が変更されます) かを選択します。

「標準」を選択した場合は、0% (不透明) から 10% 刻みで 100% (透明) までの透過度を選択肢から選択できます。

「オーバーレイ」を選択した場合は、「レイヤー フィールド」として選択した地理空間フィールドを含むデータ ソースから、任意のフィールドを選択できます。オーバーレイ フィールドの値ごとに異なる透過度でマップに表示されます。透過度は、ポイント、線、または多角形の場合に、色のドロップダウン リストから選択された色に適用されます。

データ ラベル**(D)**

注**:** 「**16** 進法データ分割の使用」チェック ボックスを選択した場合はこのオプションを使用できません。

マップのデータ ラベルとして使用するフィールドを選択するには、このオプションを使用します。例えば、多角形の層に適用した場合、データ ラベルを名前フィールドにして、それぞれの多角形の名前が含まれるようにすることが可能です。名前フィールドを選択すると、その名前がマップに表示されます。

マップ視覚化の「外観」タブ

グラフ作成前に外観オプションを指定できます。

表題**:** グラフのタイトルに使用するテキストを入力します。

サブタイトル**:** グラフのサブタイトルに使用するテキストを入力します。

キャプション**:** グラフのキャプションに使用するテキストを入力します。

---

t-SNE ノード

t 分布 Stochastic Neighbor Embedding (t-SNE)© は、高次元データの視覚化のためのツールです。t-SNE は、データ ポイントの類似性を確率に変換します。元の空間の類似性はガウス ジョイント確率によって表され、埋め込み空間の類似性はスチューデントの t 分布によって表されます。これにより、t-SNE はローカル構造に特に依存します。また、既存の技術に勝る利点が他にもいくつかあります。[1]

- 単一マップ上で、さまざまな尺度で構造を表示する
- 複数の異なる多様体またはクラスターに存在するデータを表示する
- 中央にポイントを集中させる傾向を削減する

SPSS Modeler の t-SNE ノードは Python で実装されており、`scikit-learn`© Python ライブラリーを必要とします。t-SNE および scikit-learn ライブラリーについて詳しくは、以下を参照してください。

- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE
- https://scikit-learn.org/stable/modules/manifold.html#t-sne

「ノード パレット」の「Python」タブには、このノードおよびその他の Python ノードがあります。また、「グラフ」タブには、t-SNE ノードがあります。

[1] 参照資料

van der Maaten, L.J.P.; Hinton, G. "Visualizing High-Dimensional Data using t-SNE". Journal of Machine Learning Research. 9:2579-2605, 2008.

van der Maaten, L.J.P. "t-Distributed Stochastic Neighbor Embedding".

van der Maaten, L.J.P. "Accelerating t-SNE using Tree-Based Algorithms". Journal of Machine Learning Research. 15(Oct):3221-3245, 2014.

t-SNE ノードのエキスパート オプション

t-SNE ノードにどちらのオプションを設定するかに応じて、「シンプル」モードまたは「エキスパート」モードを選択します。

視覚化タイプ: 「**2** 次元」または「**3** 次元」を選択して、2 次元または 3 次元のどちらでグラフを描画するかを指定します。

方法: 「**Barnes Hut**」または「正確確率」を選択します。デフォルトでは、勾配計算アルゴリズムは、正確確率メソッドより確実に速く実行される Barnes-Hut 近似を使用します。Barnes-Hut 近似により、t-SNE 技術を実際の大規模データ セットに適用できます。正確確率アルゴリズムは、最近隣のエラーを回避する上で、より優れています。

初期化 **(Init.):** 埋め込みの初期化について、「無作為」または「**PCA**」を選択します。

対象フィールド**:** 出力グラフのカラー マップとして表示する対象フィールドを選択します。ここで対象フィールドを指定しないと、グラフには 1 色が使用されます。

最適化

Perplexity: Perplexity は、他の多様体学習アルゴリズムで使用される最近隣の数に関連します。通常、データ セットが大きいほど、必要とされる Perplexity も大きくなります。**5** から **50** の間の値を選択することを考慮してください。デフォルトは **30**、範囲は **2** から **9999999** です。

Early exaggeration: この設定は、埋め込み空間における、元の空間の自然クラスターの気密度、およびクラスター間の間隔を制御します。デフォルトは **12**、範囲は **2** から **9999999** です。

学習率**:** 学習率が高すぎる場合、データは、すべてのポイントがその最近傍からほぼ等距離にある 1 つの「ボール」のように表示されることがあります。学習率が低すぎる場合、大部分のポイントは、圧縮された厚い雲のように表示されることがあります。外れ値はほとんどなくなります。誤った局所最小値でコスト関数が停止している場合は、学習率を高くすると改善することがあります。デフォルトは **200**、範囲は **0** から **9999999** です。

最大反復**:** 最適化の最大反復数。デフォルトは **1000**、範囲は **250** から **9999999** です。

角サイズ**:** あるポイントから測定した遠方ノードの角サイズ。**0** から **1** の間の値を入力します。デフォルトは **0.5** です。

ランダム シード

ランダム・シードの設定**:** 乱数発生ルーチンによって使用されるシードを生成するには、このオプションを選択し、「生成」をクリックします。

最適化の中断条件

進捗のない最大反復**:** 最適化を中断するまでに実行する、進捗のない反復の最大数。Early exaggeration を伴う 250 回の初期反復の後に使用されます。進捗は 50 回の反復ごとにしか検査されないため、この値は次の 50 の倍数に丸められることに注意してください。デフォルトは **300**、範囲は **0** から **9999999** です。

最小勾配ノルム **(Min gradient norm):** 勾配ノルムがこの最小しきい値を下回る場合、最適化は中断されます。デフォルトは **1.0E-7** です。

メトリック。 機能配列内のインスタンス間の距離を計算するときに使用するメトリック。メトリックが文字列である場合、メトリックは、scipy.spatial.distance.pdist のメトリック パラメータとして許可されているオプションの 1 つであるか、pairwise.PAIRWISE_DISTANCE_FUNCTIONS にリストされているメトリックである必要があります。使用可能ないずれかのメトリック タイプを選択します。デフォルトは **euclidean** です。

レコード数が次の値より大きい場合**:** 大規模データ・セットの作図の手法を指定します。最大データ・セット・サイズを使用するか、またはデフォルト値 (2,000 ポイント) を使用することができます。「ビン」 または 「サンプル」 を選択すると、大きいデータ・セットに対するパフォーマンスが向上します。代わりに、「すべてのデータを使用」 を選択して、すべてのデータ・ポイントを作図することもできます。ただし、この場合ソフトウェアのパフォーマンスが大幅に低下する可能性があります。

- ビン。データ・セットに格納されているレコード数が、指定した数より大きい場合に、分割を有効にします。分割を行うと、グラフが細かいグリッドに分割されてから、作図や各グリッド・セルに現れる接続数のカウントが実際に行われます。最終的なグラフでは、ビン重心 (ビン中のすべての接続の位置の平均) でセルごとに 1 つの接続が作図されます。
- サンプル。ここに指定した数のレコードまで、無作為にデータのサンプリングを行います。

次の表に、SPSS Modeler t-SNE ノードのダイアログの「エキスパート」タブの設定と、Python t-SNE ライブラリーのパラメータとの間の関係を示します。

表 *36.* ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**Python t-SNE** パラメータ
モード	mode_type	
視覚化タイプ	n_components	n_components
メソッド	method	method
埋め込みの初期化 (Initialization of embedding)	init	init
対象	target_field	target_field
Perplexity	perplexity	perplexity
Early exaggeration	early_exaggeration	early_exaggeration
学習率	learning_rate	learning_rate
最大反復	n_iter	n_iter
角サイズ	angle	angle
ランダム シードの設定	enable_random_seed	
ランダム シード	random_seed	random_state
進捗のない最大反復	n_iter_without_progress	n_iter_without_progress
最小勾配ノルム (Min gradient norm)	min_grad_norm	min_grad_norm
複数の Perplexity で t-SNE を実行	isGridSearch	

t-SNE ノードの出力オプション

「出力」タブで、t-SNE ノードの出力オプションを指定します。

出力名。ノードの実行時に生成される出力の名前を指定します。「自動」を選択すると、出力の名前が自動的に設定されます。

画面に出力。出力を生成し、新規ウィンドウに表示するには、このオプションを選択します。出力は、出力マネージャーにも追加されます。

ファイルに出力。出力をファイルに保存するには、このオプションを選択します。選択すると、「ファイル名」フィールドと「ファイルの種類」フィールドが有効になります。比較する目的で他のフィールドを使用するプロットを作成する場合、あるいは分類モデルまたは回帰モデルで出力を予測値として使用する場合、t-SNE ノードはこの出力ファイルにアクセスする必要があります。t-SNE モデルは、固定長ファイル ソース ノードを使用すると最も容易にアクセスできる、x、y (および z) 座標フィールドから成る結果ファイルを作成します。

次の表に、SPSS Modeler t-SNE ノードのダイアログの「出力」タブの設定と、Python t-SNE ライブラリーのパラメータとの間の関係を示します。

表 37. ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**Python t-SNE** パラメータ
出力名	output_Rename	output_Rename
出力モード	output_to	output_to
ファイル名	full_filename	full_filename
ファイルの種類	output_file_type	output_file_type
対象	target_field	target_field

t-SNE データへのアクセスおよび作図

「ファイルに出力」オプションを使用して t-SNE 出力をファイルに保存すると、比較する目的で他のフィールドを使用するプロットを作成したり、分類モデルまたは回帰モデルで出力を予測値として使用したりできます。t-SNE モデルは、固定長ファイル ソース ノードを使用すると最も容易にアクセスできる、x、y (および z) 座標フィールドから成る結果ファイルを作成します。このセクションでは、サンプル情報を提供します。

1. t-SNE ノードのダイアログで、「出力」タブを開きます。
2. 「ファイルに出力」を選択し、ファイル名を入力します。デフォルトのファイルの種類である HTML を使用します。これにより、モデルを実行すると、出力場所に 3 つの出力ファイルが生成されます。
   - テキスト ファイル (result_xxxxxx.txt)
   - HTML ファイル (指定したファイル名)
   - PNG ファイル (tsne_chart_yyyyyy.png)

   テキスト ファイルには必要なデータが含まれていますが、技術的な理由により、このデータは標準形式である場合と指数形式である場合とがあります。指数形式 (1.11111111e+01) である場合は、この形式を認識する新規ストリームを作成する必要があります。

テキスト ファイルが指数数値形式である場合の t-SNE 作図データへのアクセス

1. 新規ストリームを作成します (「ファイル」 > 「新規ストリーム」)。

2. 「ツール」 > 「ストリームのプロパティー」 > 「オプション」に移動し、「数値の形式」を選択し、数字の表示形式として「科学的表記 **(#.###E+##)**」を選択します。
3. 固定長ファイル ソース ノードをキャンバスに追加し、「ファイル」タブの以下の設定を使用します。
   - スキップするヘッダー行数: 1
   - レコード長: 54
   - tSNE_x 開始: 3、長さ: 16
   - tSNE_y 開始: 20、長さ: 16
   - tSNE_z 開始: 36、長さ: 16
4. 「データ型」タブで、数値は実数として認識されます。「値の読み込み」をクリックすると、以下のようなフィールド値が表示されます。

表 *38.* フィールド値の例

フィールド	尺度	値
tSNE_x	連続	[-7.07176703,7.14338837]
tSNE_y	連続	[-9.2188112,8.89647667]
tSNE_x	連続	[-9.95892882,9.95742482]

5. 条件抽出ノードをストリームに追加して、ファイルの下部にある、ヌルとして読み取られる以下の 2 行のテキストを削除できるようにします。

   ```
   ******************
   Perform t-SNE (total time 9.5s)
   ```

   条件抽出ノードの「設定」タブで、モードとして「破棄」を選択し、条件 `@NULL(tSNE_x)` を使用して行を削除します。
6. データ型ノードおよびフラット ファイル エクスポート ノードをストリームに追加して、元のストリームにコピー アンド ペーストするための可変長ファイル ソース ノードを作成します。

テキスト ファイルが標準数値形式である場合の t-SNE 作図データへのアクセス

1. 新規ストリームを作成します (「ファイル」 > 「新規ストリーム」)。
2. 固定長ファイル ソース ノードをキャンバスに追加します。t-SNE データにアクセスするために必要なノードは、以下の 3 つのみです。

図 *48. Stream for accessing t-SNE plot data in standard numeric format*

3. 固定長ファイル ソース ノードの「ファイル」タブの以下の設定を使用します。
   - スキップするヘッダー行数: 1
   - レコード長: 29
   - tSNE_x 開始: 3、長さ: 12
   - tSNE_y 開始: 16、長さ: 12

4. 「フィルター」タブで、field1 および field2 を tsneX および tsneY に名前変更できます。
5. 「順序」レコード結合方法を使用して、レコード結合ノードを追加し、ストリームに接続します。
6. 作図ノードを使用して、tsneX と tsneY を作図し、調査対象のフィールドを使用して色を付けることができるようになりました。

t-SNE モデル ナゲット

t-SNE モデル ナゲットには、t-SNE モデルが取得したすべての情報が含まれます。以下のタブがあります。

Graph

「グラフ」タブには、t-SNE ノードのグラフ出力が表示されます。pyplot 散布図には、低次元の結果が示されます。t-SNE ノードの「エキスパート」タブで「複数の **Perplexity** で **t-SNE** を実行」オプションを選択しなかった場合、異なる Perplexity を使用した 6 つのグラフではなく、1 つのグラフのみが含まれます。

テキスト出力

「テキスト出力」タブには、t-SNE アルゴリズムの結果が表示されます。t-SNE ノードの「エキスパート」タブで、視覚化タイプに「**2** 次元」を選択した場合、ここに表示される結果は 2 次元のポイント値です。「**3** 次元」を選択した場合、結果は 3 次元のポイント値になります。

E 散布図 (ベータ) ノード

E 散布図 (ベータ) ノードでは、数値フィールド間の関係が示されます。E 散布図 (ベータ) ノードは散布図ノードに類似していますが、オプションは異なり、新規グラフ機能を使用します。SPSS Modeler の新規グラフ機能を利用するには、このノードを使用します。

E 散布図 (ベータ) ノードでは、数値フィールド間の関係を示すために、散布図、折れ線グラフ、および棒グラフが提供されます。このノードの新規グラフ インターフェースは、直感的な最新の機能であり、柔軟にカスタマイズ可能であり、データ グラフはインタラクティブです。詳しくは、 289 ページの『E 散布図グラフの使用方法』を参照してください。

E 散布図 (ベータ) ノードの「作図」タブ

散布図は、*X* フィールドの値に対する *Y* フィールドの値を表しています。多くの場合、これらのフィールドはそれぞれ従属変数と独立変数に対応しています。

X フィールド。リストから、横の *x* 軸に表示するフィールドを選択します。

Y フィールド**:**リストから、縦の *y* 軸に表示するフィールドを選択します。

オーバーレイ**:** データ値のカテゴリーを描くにはさまざまな方法があります。例えば、*maincrop* (主作物) フィールドを色のオーバーレイとして使用して、申請者による主作物の成長に応じた *estincome* (推定所得) と *claimvalue* (申請値) の値を示すことができます。出力でのカラー マッピング、サイズ マッピング、および形状マッピングのフィールドを選択します。インタラクティブ出力に含める、その他の関心のあるフィールドも選択します。詳しくは、トピック 201 ページの『外観、オーバーレイ、パネル、およびアニメーション』を参照してください。

E 散布図のオプションを設定したら、「実行」をクリックして、ダイアログ・ボックスから直接プロットを実行できます。「オプション」タブを使用して、追加オプションを指定することもできます。

E 散布図 (ベータ) ノードの「オプション」タブ

プロットするレコードの最大数。大きいデータ・セットの作図手法を指定します。最大データ・セット・サイズを使用するか、またはデフォルト値 (2,000 レコード) を使用することができます。「サンプル」 を選択すると、大きいデータ・セットに対するパフォーマンスが向上します。「サンプル」オプションは、このテキスト・フィールドに入力した数のレコードまで、無作為にデータのサンプリングを行います。代わりに、「すべてのデータを使用」 を選択して、すべてのデータ・ポイントを作図することもできます。ただし、この場合ソフトウェアのパフォーマンスが大幅に低下する可能性があります。

E 散布図 (ベータ) ノードの「外観」タブ

必要であれば、グラフ作成前にタイトルおよびサブタイトルを指定できます。グラフ作成後にこれらのオプションを指定または変更することもできます。

表題: グラフのタイトルに使用するテキストを入力します。

サブタイトル: グラフのサブタイトルに使用するテキストを入力します。

E 散布図グラフの使用方法

E 散布図 (ベータ) ノードでは、数値フィールド間の関係を示すために、散布図、折れ線グラフ、および棒グラフが提供されます。このベータ ノードで導入された新規グラフ インターフェースには、多くの新機能および改善された機能が含まれています。

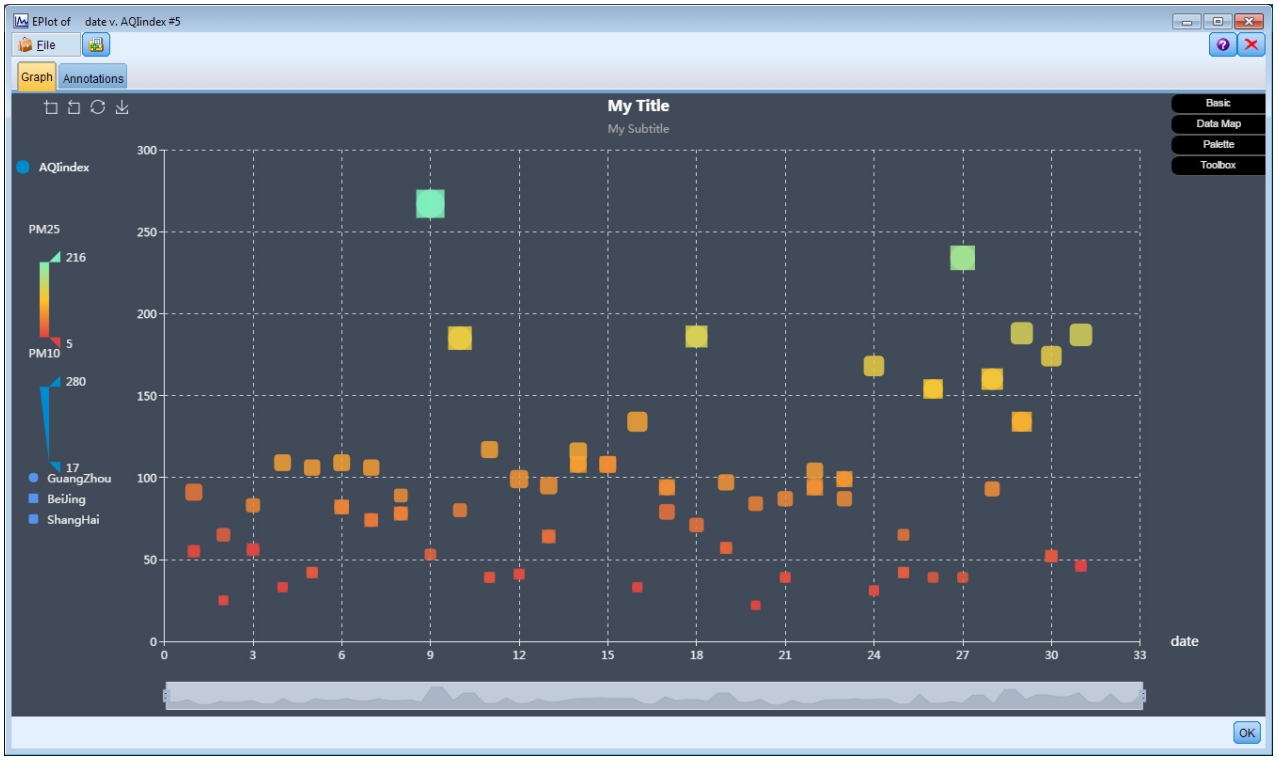


図 *49. E-Plot (Beta) scatterplot graph*

「グラフ」タブの左上隅にはツールバーがあり、このツールバーを使用して、グラフの特定のセクションのズーム イン/ズーム アウト、初期の全画面表示への復帰、グラフの保存 (外部使用のため) を行うことができます。

図 *50. Toolbar*

ウィンドウの下部にあるスライダを使用して、グラフの特定のセクションにズーム インできます。右側および左側にある小さい長方形のコントロールを移動して、ズームします。このスライダを使用するには、最初に「ツールボックス (Toolbox)」オプション領域でこの機能をオンにする必要があります。

図 *51. Zoom slider*

ウィンドウの左側には、表示される値の範囲を変更するためのコントロールがあります。これらのコントロールを使用するには、最初に「データ マップ (Data Map)」オプション領域でオプションを指定する必要があります。下の例では、`PM25` という名前のフィールドがカラー マップとして選択され、`PM10` という名前のフィールドがサイズ マップとして選択され、`City` という名前のフィールドが形状マップとして選択されています。縦方向のカラー バーの上でカーソルを移動すると、対応するグラフ領域を強調表示できます。上部または下部の三角形をスライドすることもできます。

図 *52. Range controls*

ウィンドウの右側にある展開可能なオプションのセットを使用して、データを対話処理し、リアルタイムでグラフの外観を変更することができます。

図 *53. Expandable options*

基本オプション

図 *54. Basic options*

濃い色または薄い色のテーマを選択し、タイトルおよびサブタイトルを指定し、グラフ タイプ (散布図、折れ線グラフ、または棒グラフ) を選択し、Y 軸上に表示される系列を選択します。「折れ線」グラフを選択した場合は、Y 軸上のフィールドのみが表示され、Y 軸上のフィールドのみが「データ マップ (Data Map)」オプションのカラー マップおよびサイズ マップで選択可能です。「棒」グラフを選択した場合は、カラー マップ オプションのみが「データ マップ (Data Map)」オプションで選択可能です。系列については、E 散布図ノードの「作図」タブで選択したすべての「関心のある **(Interested)**」フィールドがここで選択可能です。

「データ マップ **(Data Map)**」オプション

図 *55. Data map options*

「カラー マップ」として連続型フィールドまたはカテゴリ型フィールドを選択します。連続型フィールドが選択された場合は、緑から赤までの色が表示されます。値を小さくするほど、その色は赤に近くなります。値を大きくするほど、その色は緑に近くなります。カテゴリ型フィールドが選択された場合は、定義されたカラー パレットに従って、フィールドの色が表示されます。

「サイズ マップ **(Size Map)**」は、連続型フィールドのみをサポートします。グラフ上で値を小さくするほど、その作図サイズは小さくなります。

「形状マップ **(Shape Map)**」は、カテゴリ型フィールドのみをサポートします。マップ上に表示される形状は、さまざまな形状の要素 (カテゴリーごとに 1 つ) に視覚化を分割するカテゴリー・フィールドによって定義されます。

「パレット」オプション

図 *56. Palette options*

タイトルおよび系列で使用される色をカスタマイズする場合は、「パレット」を使用します。ドロップダウンからタイトルまたは系列を選択し、「事前定義された色の編集 **(Edit Predefined Colors)**」をクリックし、「その他」クリックして、色を選択します。あるいは、RGB フィールドまたは 16 進数フィールドを使用して、正確な色を指定できます。

「ツールボックス **(Toolbox)**」オプション

図 *57. Toolbox options*

「ツールボックス (Toolbox)」オプションを使用して、ズーム スライダをオンまたはオフにし、グリッド プロパティを設定し、マウス トラッキングをオンまたはオフにします。マウス トラッキングは、グラフの上でカーソルを移動すると、正確な座標位置を表示する機能です。

グラフの検証

編集モードを使用するとグラフのレイアウトおよび外観を編集できますが、検証モードを使用するとグラフに表示されたデータおよび値を分析的に検証できます。検証の主な目的は、データを分析し、バンド、領域およびマークを使用して値を識別し、条件抽出ノード、フィールド作成ノード、バランス・ノードを生成します。このモードを選択するには、メニューから「表示」>「探索モード」を選択します (またはツールバーのアイコンをクリックします)。

一部のグラフはすべての検証ツールを使用できますが、1 つのツールのみ受け入れるグラフもあります。検証モードには次の内容が含まれます。

- *x* 軸の尺度に沿って値を分割するために使用されるバンドの定義および編集。 詳しくは、トピック 294 ページの『バンドの使用』を参照してください。
- 四角形の領域内にある値のグループを識別するために使用する領域の定義および編集。 詳しくは、トピック 297 ページの『領域の使用』を参照してください。
- 条件抽出ノードまたはフィールド作成ノードを生成するために使用できた値を手作業で選択する要素のマークおよびマーク解除。 詳しくは、トピック 299 ページの『マークされた要素の使用』を参照してください。
- バンド、領域、マークされた要素、ストリームで使用する Web リンクによって識別される値を使用したノードの生成。 詳しくは、トピック 300 ページの『グラフからのノードの生成』を参照してください。

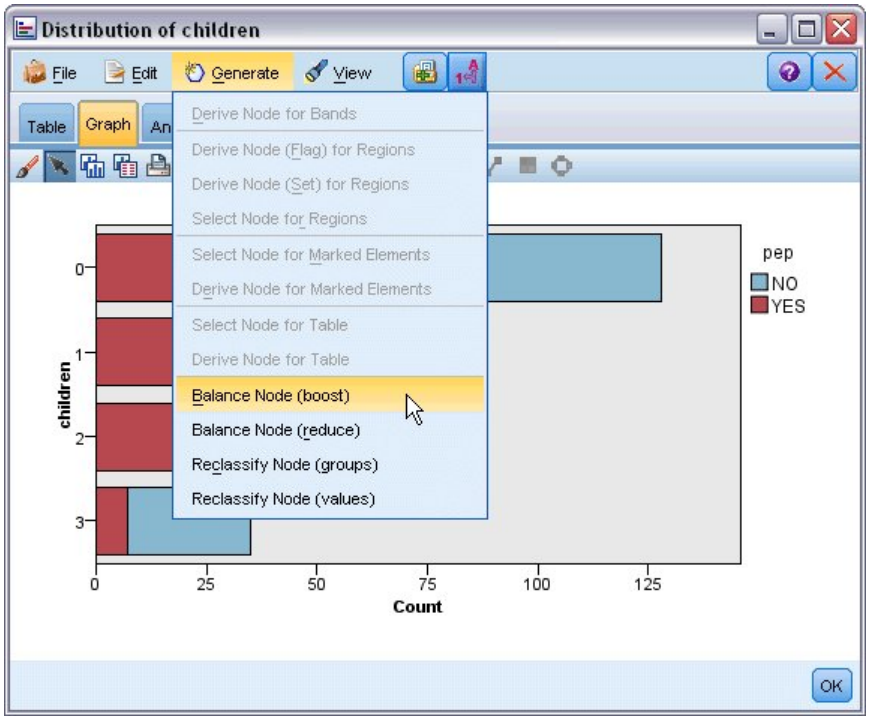


図 *58.* 生成メニューを表示したグラフ

バンドの使用

x 軸上に尺度フィールドを持つグラフでは、垂直なバンドのラインを描いて *x* 軸上の値の範囲を分割できます。グラフに複数のパネルがある場合、1 つのパネルに書かれたバンドのラインは他のパネルにも同様に表示されます。

すべてのグラフでバンドを使用できるわけではありません。バンドを使用できるグラフには、ヒストグラム、棒グラフ、分布図、作図 (線、散布図、時間など)、コレクション、評価グラフがあります。パネルを含むグラフでは、バンドがすべてのパネルに表示されます。また SPLOM では、フィールド/変数のバンドが描かれた軸が表示されるため、水平なバンドのラインが表示される場合があります。

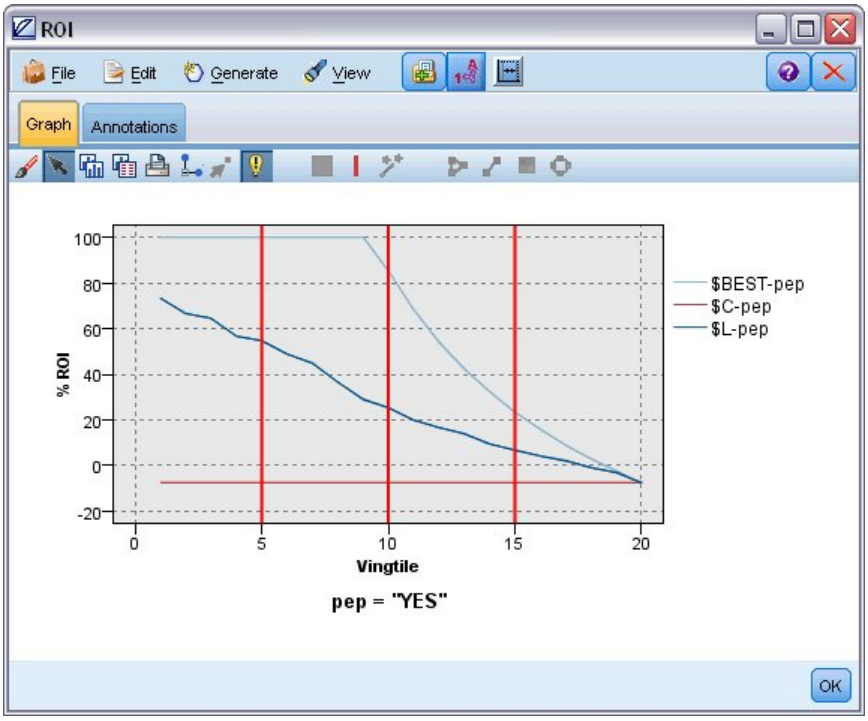


図 *59. 3* つのバンドを使用したグラフ

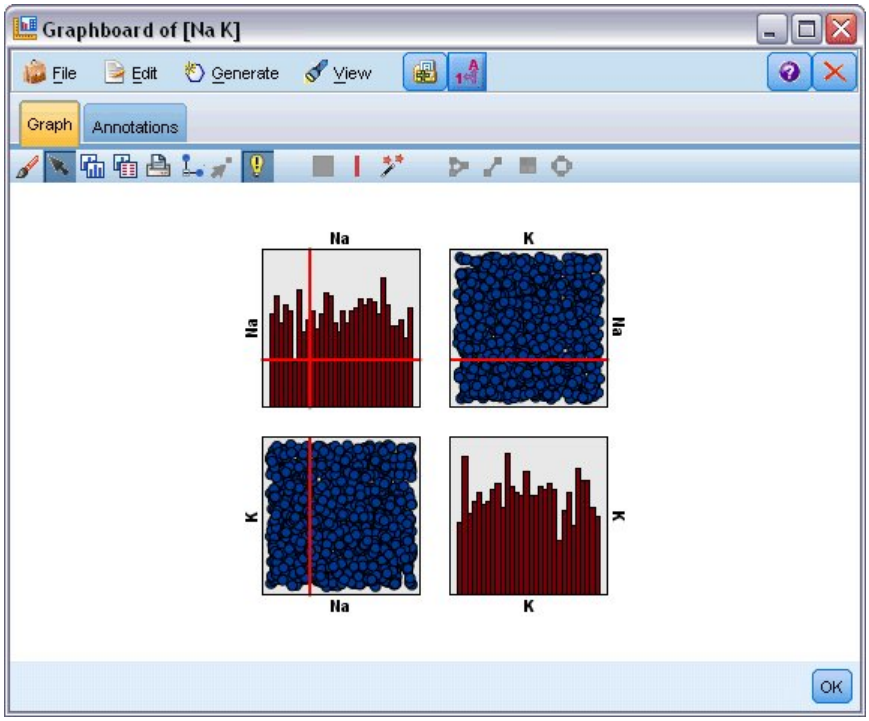


図 *60.* バンドを使用した *SPLOM*

バンドの定義

バンドのないグラフで、バンドのラインを追加するとグラフを 2 つのバンドに分割します。バンドのラインの値は、下限値としても参照される、グラフを左から右へ読んだ場合の 2 番目のバンドの開始点を表します。同様に 2 つのバンドを使用したグラフでは、バンドのラインを追加するとこれらのバンドの 1 つを 2 つに分割し、3 つのバンドを作成します。デフォルトでは、バンドにはバンド *N* という名前が付けられます。*N* には、*x* 軸上で左から右へ順番にバンド数が割り当てられます。

バンドを定義すると、バンドをドラッグ・アンド・ドロップして *x* 軸に再配置できます。バンドの内部を右クリックして、特定のバンドのノードの名前の変更、削除、生成などのタスクのショートカットをさらに表示できます。

バンドを定義する手順は次のとおりです。

1. 探索モードであることを確認します。メニューから、「表示」>「探索モード」を選択します。
2. 検証モードのツールバーで、「バンドを描画」ボタンをクリックします。

図 *61.* 「バンドを描画」ツールバー・ボタン

3. バンドを使用するグラフで、バンドのラインを定義する *x* 軸の値のポイントをクリックします。

注 : 代わりに、「グラフをバンドに分割」 ツールバー・アイコンをクリックして、必要な等しいバンドの数を入力し、「分割」 をクリックします。

図 62. バンドに分割するオプションを表示する分割アイコン

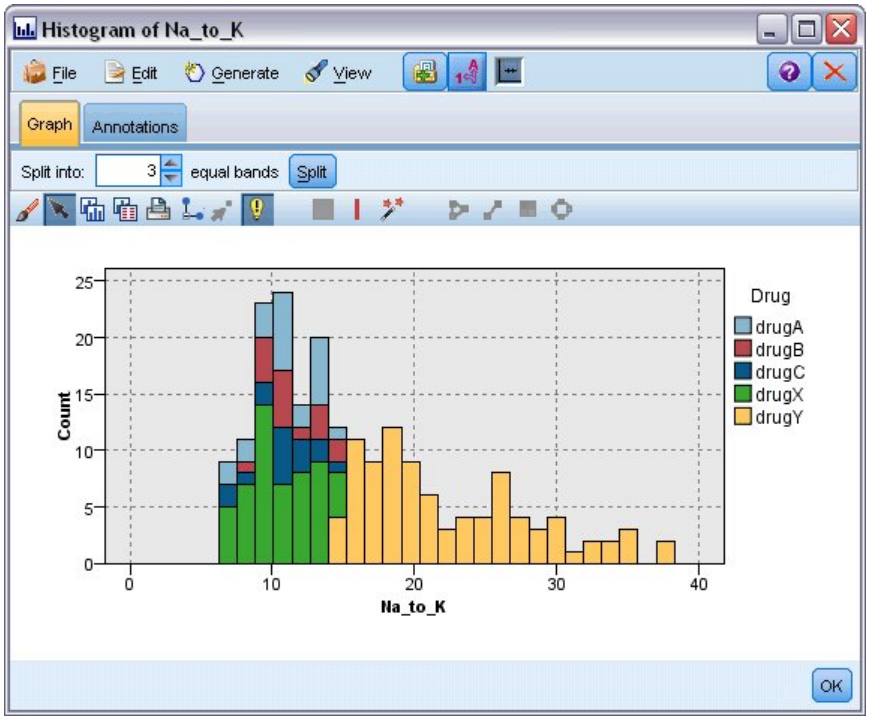


図 63. バンドを有効化した、等しいバンドを作成するツールバー

バンドの編集、名前の変更、削除

「グラフバンドの編集」ダイアログ・ボックスまたはグラフのコンテキスト・メニューで既存のバンドのプロパティーを編集できます。

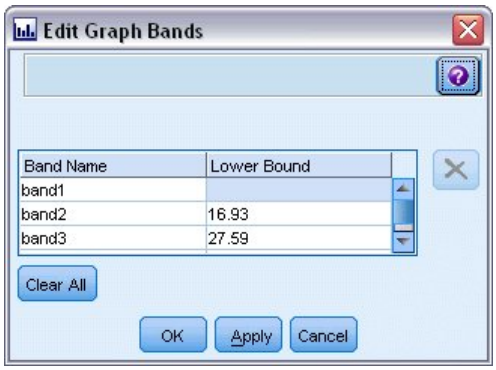


図 64. 「グラフ バンドの編集」ダイアログ・ボックス

バンドを編集する手順は次のとおりです。

1. 探索モードであることを確認します。メニューから、「表示」>「探索モード」を選択します。
2. 検証モードのツールバーで、「バンドを描画」ボタンをクリックします。

3. メニューから 「編集」>「グラフ バンド」 を選択します。「グラフ バンドの編集」ダイアログ・ボックスが表示されます。
4. グラフ (SPLOM グラフなど) に複数のフィールドがある場合、必要なフィールドをドロップダウン・リストで選択できます。
5. 名前および下限を入力して新しいバンドを追加します。Enter キーを押して、新しい行を開始します。
6. 「下限値」 を調整してバンドの境界を編集します。
7. 新しいバンド名を入力して、バンドの名前を変更します。
8. テーブルから削除するラインを選択、「削除」ボタンをクリックして、バンドを削除します。
9. 「**OK**」 をクリックして、変更を適用し、ダイアログ・ボックスを閉じます。

注 : 代わりに、バンドのラインを右クリックしてコンテキスト・メニューからオプションを選択し、グラフのバンドを直接削除したり名前を変更することもできます。

領域の使用

2 つの尺度 (または領域) 軸を持つグラフでは、領域を描画して、描画した四角形の領域内に値をグループ化します。領域とは、*X* と *Y* の最大値と最小値で示されるグラフの領域のことです。グラフに複数のパネルがある場合、1 つのパネルに書かれた領域は他のパネルにも同様に表示されます。

すべてのグラフで領域を使用できるわけではありません。領域を使用できるグラフには、作図 (線、散布図、バブル、時間など)、SPLOM、コレクションがあります。これらの領域は X 領域および Y 領域に描画されます。そのため、1 次元、3 次元またはアニメーション・プロットでは定義できません。パネルを含むグラフでは、領域がすべてのパネルに表示されます。散布図行列 (SPLOM) の場合、1 つの尺度フィールドのみ表示されるため、対応する領域は対角プロットでなく対応する上位プロットに表示されます。

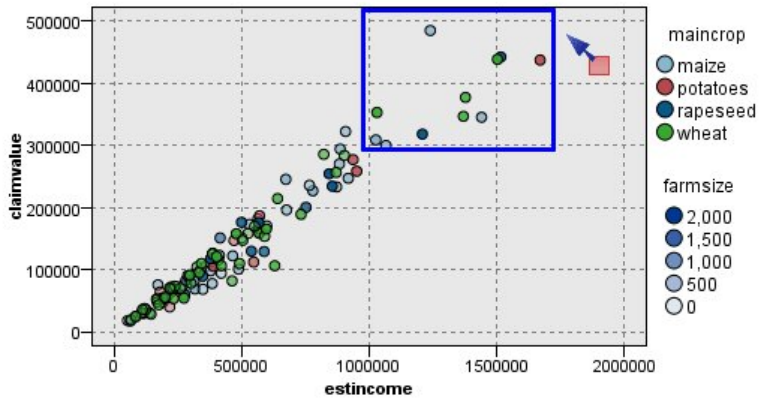


図 *65.* 申請値の高い領域の定義

領域の定義

領域を定義する場合、値のグループを作成します。デフォルトでは、新しい領域は領域 <*N*> と命名されます。*N* には、すでに作成された領域数に対応した数字が入ります。

領域を定義すると、領域のラインを右クリックして基本のショートカットを取得できます。ただし、ライン上ではなく領域の内部を右クリックして、特定の領域の条件抽出ノードおよびフィールド選択ノードの名前の変更、削除、生成などのタスクのショートカットをさらに表示できます。

特定の領域または複数の領域の 1 つに含まれているかどうかを基準にして、レコードのサブグループを選択できるようになります。また、フィールド作成ノードを作成し、領域に含まれているかどうかを基準にレコードにフラグを設定して、レコードの領域情報を組み込むこともできます。 詳しくは、トピック 300 ページの『グラフからのノードの生成』を参照してください。

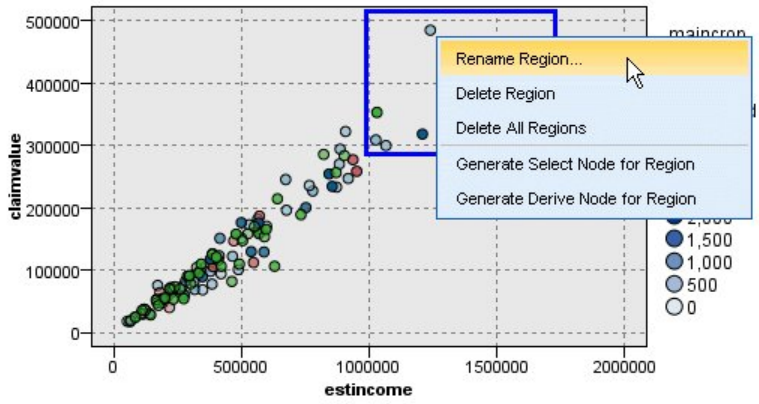


図 *66.* 高い申請値を持つ領域の調査

領域を定義する手順は、次のとおりです。

1. 探索モードであることを確認します。メニューから、「表示」>「探索モード」を選択します。
2. 検証モードのツールバーで、「領域を描画」ボタンをクリックします。

図 *67.* 「領域を描画」ツールバー・ボタン

3. 領域を使用するグラフで、マウスをクリックおよびドラッグして、四角形の領域を描画します。

領域の編集、名前の変更、削除

「グラフ領域の編集」ダイアログ・ボックスまたはグラフのコンテキスト・メニューで既存のバンドのプロパティーを編集できます。

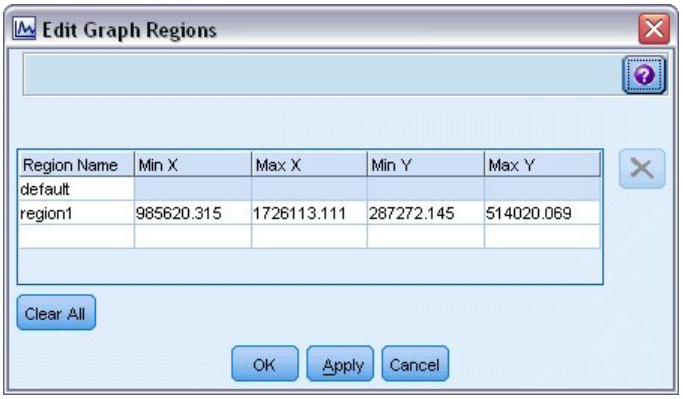


図 *68.* 定義した領域のプロパティーの指定

領域を編集する手順は、次のとおりです。

1. 探索モードであることを確認します。メニューから、「表示」>「探索モード」を選択します。
2. 検証モードのツールバーで、「領域を描画」ボタンをクリックします。
3. メニューから 「編集」>「グラフ領域」 を選択します。「グラフ領域の編集」ダイアログ・ボックスが表示されます。
4. グラフ (SPLOM など) に複数のフィールドがある場合、「フィールド *A*」列および「フィールド *B*」列に領域のフィールドを定義する必要があります。
5. 名前を入力、フィールド名を選択 (該当する場合) および各フィールドの上限および下限を定義して、新しいラインに新しい領域を追加します。Enter キーを押して、新しい行を開始します。
6. *A* および *B* の 「最小」 値および 「最大」 値を調整して既存の領域の境界を編集します。
7. テーブル内の領域の名前を変更して、領域名の変更を行います。
8. テーブルから削除するラインを選択、「削除」ボタンをクリックして、領域を削除します。
9. 「**OK**」 をクリックして、変更を適用し、ダイアログ・ボックスを閉じます。

注 : 代わりに、領域のラインを右クリックしてコンテキスト・メニューからオプションを選択して、グラフの領域を直接削除したり名前を変更することもできます。

マークされた要素の使用

グラフ内のバー、スライス、ポイントなどの要素をマークできます。時系列グラフ、線グラフ、評価グラフのラインはフィールドを参照するため、それ以外のグラフのライン、領域、面をマークできません。要素をマークすると、要素で表示されたすべてのデータは基本的に強調表示されます。同じケースが複数の場所に表示されるグラフ (SPLOM など) では、マークはブラシと同義です。バンド内および領域内であっても、グラフの要素をマークできます。要素をマーク後に編集モードに戻っても、マークは表示されたままです。

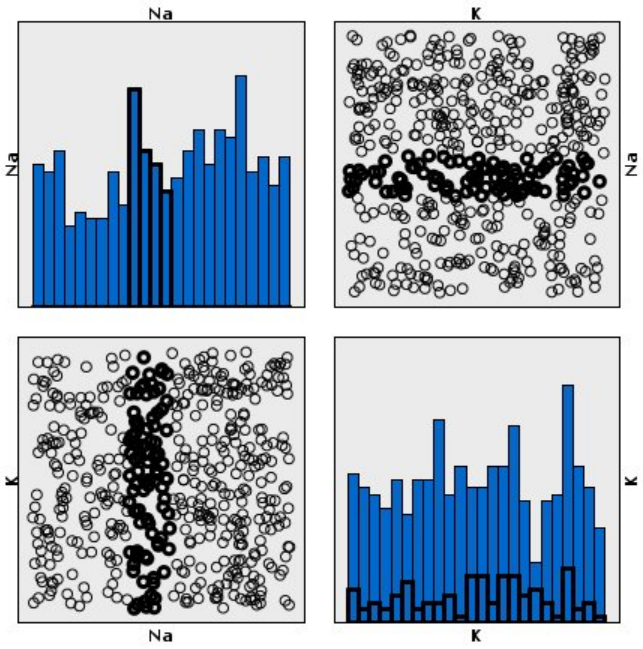


図 *69. SPLOM* の要素のマーク

グラフ内の要素をクリックして要素をマークおよびマーク解除できます。最初に要素をクリックしてマークすると、要素は境界線の色で濃く表示され、マークされていることが示されます。要素をサイドクリックす

ると、境界の色が消え、要素のマークが解除されます。複数の要素をマークするには、Ctrl キーを押したまま要素をクリックするか、「マジック ワンド」を使用してマークするそれぞれの要素の上でマウスをドラッグします。Ctrl キーを押さずに別の領域または要素をクリックすると、以前マークされた要素はすべてクリアされます。

グラフ内のマークされた要素から条件抽出ノードおよびフィールド作成ノードを生成できます。 詳しくは、トピック『グラフからのノードの生成』を参照してください。

要素をマークする手順は、次のとおりです。

1. 探索モードであることを確認します。メニューから、「表示」>「探索モード」を選択します。
2. 検証モードのツールバーで、「要素をマーク」ボタンをクリックします。
3. 必要な要素をクリックするか、マウスでクリック・アンド・ドラッグして複数の要素を含む領域周辺に線を描画します。

グラフからのノードの生成

IBM SPSS Modeler のグラフで提供される最も強力な機能の 1 つは、グラフまたはグラフ内で選択した要素からノードを生成する機能です。例えば、時系列グラフから、データを選択するか領域を指定し、効果的にデータの「サブセットを作成」することで、フィールド作成 (設定とフラグ) ノードと条件抽出ノードを生成することができます。例えば、強力なこの機能を使用し、外れ値を識別してそれを除外することができます。

バンドを描画すると、フィールド作成ノードも生成できます。2 つの尺度軸を持つグラフでは、グラフ内に描画された領域からフィールド作成ノードまたは条件抽出ノードを生成できます。マークされた要素を持つグラフでは、これらの要素からフィールド作成ノード、条件抽出ノードを作成でき、またフィルター・ノードを作成できる場合もあります。バランス・ノード生成は、度数の分布を示すグラフで可能です。

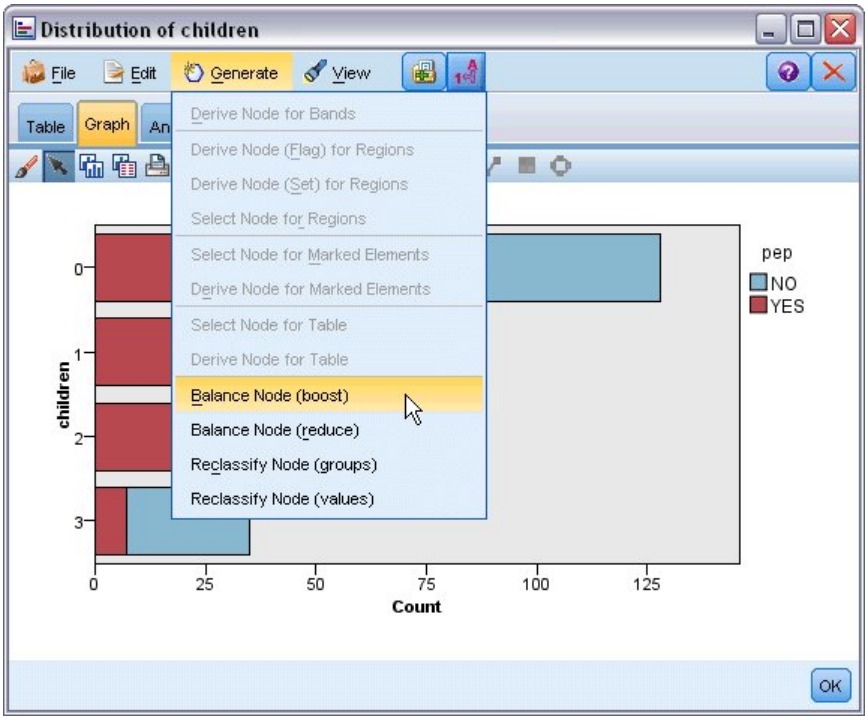


図 *70.* 生成メニューを表示したグラフ

ノードを生成すると、ストリーム領域に直接配置され、ノードを既存のストリームに接続できます。条件抽出ノード、フィールド作成ノード、バランス・ノード、フィルター・ノード、データ分類ノードはグラフから生成できます。

条件抽出ノード

条件抽出ノードを生成して、領域内のレコードの追加および領域外または下流の処理の逆となる全レコードの除外を検定できます。

- バンドの場合: 該当するバンド内のレコードを追加または除外する条件抽出ノードを生成できます。条件抽出ノードで使用するバンドを選択する必要があるため、「バンドのみの条件抽出ノード」 はコンテキスト・メニューでのみ使用できます。
- 領域の場合: 該当する領域内のレコードを追加または除外する条件抽出ノードを生成できます。
- マークされた要素の場合: マークされた要素または Web グラフのリンクに対応するレコードを取得する条件抽出ノードを生成できます。

フィールド作成ノード

フィールド作成ノードは、領域、バンド、およびマークされた要素から生成できます。すべてのグラフで、フィールド作成ノードを作成できます。評価グラフの場合、モデル選択のダイアログ・ボックスが表示されます。Web グラフの場合、「フィールド作成ノード **(**「**AND**」**)**」と「フィールド作成ノード **(**「**OR**」**)**」を指定できます。

- バンドの場合:「バンドの編集」ダイアログ・ボックスにカテゴリー名として表示されたバンド名を使用して、軸上にマークされた各区分のカテゴリーを作成するフィールド作成ノードを生成できます。
- 領域の場合: 領域内のレコードには *T* に設定、全領域外のレコードには *F* に設定されるフラグを持つ *in_region* というフラグ型フィールドを作成するフィールド作成ノード (「フィールド作成ノード **(**フラグ型**)**」) を生成できます。また、領域 と呼ばれる各レコードの新しいフィールドを含む、各領域の値を持つセットを作成するフィールド作成ノード (「フィールド作成ノード **(**セット型**)**」) も生成できます。そのフィールド作成ノードではレコードが分類される領域の名前を値として取得します。どの領域にも属さないレコードの場合は、デフォルトの領域名が表示されます。値の名前は、「領域の編集」ダイアログ・ボックスに表示された領域名となります。
- マークされた要素の場合 :マークされたすべての要素は真、それ以外のすべてのレコードは偽であるフラグを集計するフィールド作成ノードを生成できます。

バランス・ノード

バランス・ノードを生成して、共通する値の頻度を減らす (「バランス ノード **(**減少**)**」 メニュー・オプションの使用)、またはあまり出現しない値の発生を増加させる (「バランス ノード **(**増加**)**」 メニュー・オプションの使用) など、データの不均衡を修正できます。バランス・ノードの生成は、ヒストグラム、点グラフ、集計グラフ、度数の棒グラフ、度数の円グラフ、および線グラフなど、度数の分布を示すグラフで有効です。

フィルター・ノード

フィルター・ノードを生成して、グラフ内でマークされたラインまたはノードに基づいてフィールドの名前を変更またはフィルタリングできます。評価グラフの場合、最良の適合線はフィルター・ノードを生成しません。

データ分類ノード

データ分類ノードを生成して、値を再分類できます。このオプションは、分布図に使用されます。グループ中の選択内容に応じて (Ctrl +「テーブル」タブでグループを選択)、表示されたフィールドの特定の値を再分類するグループのデータ分類ノードを生成できます。また、値のデータ分類ノードを生成して、各種会社の金融データを分析のために結合するためにデータを標準的な値のセットに再分類するなど、多数の値からなる既存のセットにデータを再分類できます。

注**:** 値があらかじめ定義されている場合、IBM SPSS Modeler にフラット・ファイルとして読み込み、棒グラフを使用してすべての値を表示できます。次に、このフィールドのデータ分類 (値) ノードをチャートから直接生成します。これにより、すべての対象値を分類ノードの「新しい値」列 (ドロップダウン・リスト) に分類します。

データ分類ノードのオプションを設定すると、テーブルにより、以下のように古いセット値から指定の新しい値への明確なマッピングを実行できます。

- 元の値**:** 選択したフィールドの既存の値が表示されます。
- 新しい値**:** 新しいカテゴリー値を入力するか、またはドロップダウン・リストから選択します。分布図からの値を使用して、データ分類ノードを自動的に生成する場合、これらの値はドロップダウン・リストに含まれます。これにより、既存の値を既知の値のセット素早くに関連付けることができます。例えば、ネットワークやロケールに基づいて、医療機関は異なる方法で診断をグループ化することがあります。合併または買収の後、すべての機関は新規のみならず既存のデータさえも一貫した方法で分類することが要求されます。長いリストから各対象値を手作業で入力しなくても、値の基本リストから IBM SPSS Modeler に読み込み、「診断」 フィールドの分布図を実行し、分布図から直接このフィールドのデータ分類ノード (値) を生成できます。このプロセスにより、「新しい値」ドロップダウン・リストのすべての対象「診断」値が利用可能になります。

データ分類ノードについて詳しくは、 172 ページの『データ分類ノードのオプション設定』を参照してください。

グラフからのノードの生成

グラフ出力ウィンドウの「ノードの生成」メニューを使用してノードを生成できます。生成されたノードはストリーム領域に配置されます。このノードを使用するには、これを既存のストリームと接続します。

グラフからノードを生成する手順は、次のとおりです。

1. 探索モードであることを確認します。メニューから、「表示」>「探索モード」を選択します。
2. 検証モードのツールバーで、「領域」ボタンをクリックします。
3. ノードの定義に必要なバンド、領域、およびマークされた要素を定義します。
4. 「ノード生成」メニューから、作成するノードの種類を選択します。作成できるノードのみが有効です。

注 : 代わりに、右クリックした後、コンテキスト・メニューから該当する生成オプションを選択して、グラフから直接ノードを生成することもできます。

視覚化の編集

探索モードでは、視覚化によって表現されたデータや値を分析的に検討することができます。一方、編集モードでは、視覚化のレイアウトや外観を変更することができます。例えば、フォントや色を自分の組織のスタイル・ガイドに合わせて変更することが可能です。このモードを選択するには、メニューから「表示」>「編集モード」を選択します (またはツールバーのアイコンをクリックします)。

編集モードには、視覚化のレイアウトのさまざまな要素に影響を与えるいくつかのツールバーがあります。使用しないものがある場合は、そのツールバーを非表示にしてダイアログ・ボックスにおけるグラフの表示領域を増やすことができます。ツールバーの選択または選択解除を行うには、関連するツールバーの名前を「表示」メニューでクリックします。

注：視覚化に詳細を追加する際に、タイトル、脚注、軸ラベルを適用することができます。詳しくは、トピック 314 ページの『表題と脚注の追加』を参照してください。

編集モードには、視覚化を編集するためのオプションがいくつか用意されています。以下を行うことができます。

- テキストを編集して書式を設定する。
- 塗りつぶしの色、透過性、枠とグラフィック要素のパターンを変更する。
- 枠線と線について、色と破線化を変更する。
- ポイント要素を回転し、形状や縦横比を変更する。
- 棒や点などのグラフィック要素のサイズを変更する。
- 余白とパディングを使用して、項目の周囲のスペースを調整する。
- 数値の形式を指定する。
- 軸とスケールの設定を変更する。
- カテゴリー軸のカテゴリーのソート、除外、縮小を行う。
- パネルの方向を設定する。
- 座標系に変換を適用する。
- 統計量、グラフィック要素の種類、衝突変更子を変更する。
- 凡例の位置を変更する。
- 視覚化スタイル・シートを適用する。

以下の各トピックでは、これら各種の作業の方法について説明します。グラフの編集に関する一般的なルールも参照してください。

編集モードへの切り替え方法

メニューから次の項目を選択します。

「表示」 > 「編集モード」

視覚化を編集する場合の一般的なルール

編集モード

すべての編集が、編集モードで実行されます。編集モードを有効にするには、メニューから次の項目を選択します。

「表示」 > 「編集モード」

選択部分

編集で使用可能なオプションは、選択内容によって異なります。選択内容によっては、別のツールバーとプロパティー・パレットのオプションが使用可能になります。使用可能な項目だけが、現在の選択内容に適用されます。例えば軸を選択した場合、プロパティー・パレットで「スケール」タブ、「大分割の目盛り」タブ、「小分割の目盛り」タブが使用可能になります。

ここでは、視覚化の項目を選択する場合のヒントについていくつか説明します。

- 項目を選択するには、その項目をクリックします。
- グラフィック要素 (散布図の点や棒グラフの棒など) の場合は、シングルクリックで選択します。選択してからもう一度クリックすると、選択対象がグラフィック要素のグループまたは単一のグラフィック要素に絞り込まれます。
- Esc を押すと、すべて選択解除されます。

パレット

視覚化で項目を選択すると、各種のパレットが更新され、選択内容が反映されます。これらのパレットには、選択内容を編集するためのコントロールが組み込まれています。パレットは、ツールバーの場合もあれば、複数のコントロールとタブが組み込まれたパネルの場合もあります。編集に必要なパレットだけが表示されるように、パレットは非表示にすることができます。現在表示されているパレットについては、「表示」メニューを確認してください。

パレットの位置を変更するには、ツールバー・パレットの空きスペースまたは他のパレットの左側をクリックしてドラッグします。パレットを移動できる場所が視覚的に表示されます。ツールバー以外のパレットの場合は、「閉じる」ボタンをクリックしてパレットを非表示にしたり、取り外しボタンをクリックしてパレットを別のウィンドウに表示したりすることもできます。ヘルプ・ボタンをクリックすると、そのパレットのヘルプが表示されます。

自動設定

一部の設定には「-自動-」オプションがあります。このオプションは、値が自動的に適用されることを示します。使用される自動設定は、それぞれの視覚化とデータ値によって異なります。値を入力すると、自動設定をオーバーライドすることができます。自動設定を復元する場合は、現在の値を削除して Enter キーを押します。この操作により、設定に再び「-自動-」が表示されます。

項目の除外/非表示

視覚化では、さまざまな項目について、除外/非表示を行うことができます。例えば、凡例や軸ラベルを非表示にすることができます。項目を削除するには、削除したい項目を選択して Delete キーを押します。削除が許可されていない項目の場合、操作は何も実行されません。誤って項目を削除してしまった場合は、Ctrl+Z キーを押して削除を取り消してください。

状態

ツールバーには、現在選択されている項目の状態が反映されるものと反映されないものがあります。プロパティー・パレットには、常に項目の状態が反映されます。ツールバーに項目の状態が反映されない 場合は、そのツールバーの説明が記載されているトピックを参照してください。

テキストの編集と書式設定

テキストをその場で編集したり、テキスト・ブロック全体の書式を変更したりすることができます。データ値に直接リンクされているテキストを編集することはできません。例えば、目盛りラベルの内容はその基礎となるデータから派生しているため、目盛りラベルを編集することはできません。ただし、視覚化では、すべてのテキストの書式を設定することができます。

その場でテキストを編集する方法

1. テキスト・ブロックをダブルクリックします。この操作により、すべてのテキストが選択されます。同時に、すべてのツールバーが使用不可になります。これは、テキストの編集中は、視覚化の他の部分を変更できないためです。
2. 既存のテキストに代わるテキストを入力します。テキストを再度クリックしてカーソルを表示することもできます。任意の位置にカーソルを移動して、追加のテキストを入力します。

テキストの書式設定の方法

1. テキストが含まれている枠を選択します。テキストをダブルクリックしないでください。
2. フォント・ツールバーを使用してテキストの書式を設定します。ツールバーが使用できない場合は、テキストを含む枠 だけが選択されているかどうかを確認してください。テキスト自体が選択されていると、ツールバーが使用不可になります。

以下のようにフォントを変更することができます。

- 色
- ファミリー (Arial や Verdana など)
- サイズ (pc など、別の単位を指定した場合を除き、単位は pt になります)
- 重み
- テキスト枠に対する相対位置による位置合わせ

書式は、枠内のすべてのテキストに適用されます。テキストの特定のブロックにおける個々の文字や単語の書式を変更することはできません。

色、パターン、破線化、透過度の変更

視覚化の多くの項目には、塗りつぶしと枠線が設定されます。最も分かりやすい例として、棒グラフの棒があります。棒の色は塗りつぶし色です。また、棒の周囲を黒い実線の境界線で囲むこともできます。

棒グラフの棒ほど分かりやすくはありませんが、塗りつぶし色が設定される視覚化項目はほかにもあります。塗りつぶし色が透明な場合は、塗りつぶしに気付かないことが場合があります。例として、軸のラベルのテキストを考えてみます。このテキストは「浮いている」ように見えますが、実際には枠内に表示され、枠に対して透明の塗りつぶし色が設定されています。軸のラベルを選択すると枠が表示されます。

視覚化全体を囲む枠を含め、視覚化のすべての枠について、塗りつぶしと枠線のスタイルを設定することができます。また、すべての塗りつぶしには、調整可能な不透明度/透明度が関連付けられています。

色、パターン、破線化、透過度の変更方法

1. 書式を設定したい項目を選択します。例えば、棒グラフの棒やテキストを囲む枠を選択します。視覚化がカテゴリー変数またはカテゴリー・フィールドによって分割されている場合は、個々のカテゴリーに対応するグループを選択することもできます。この方法により、そのグループに割り当てられているデフォルトの外観を変更することができます。例えば、積み上げ棒グラフのいずれかの積み上げグループの色を変更することができます。
2. 塗りつぶし色、枠線の色、塗りつぶしパターンを変更するには、色ツールバーを使用します。

   注 : このツールバーには、現在選択されている項目の状態は反映されません。

   色または塗りつぶしを変更する場合、ボタンをクリックして表示されるオプションを選択することも、ドロップダウンの矢印をクリックして別のオプションを選択することもできます。色の場合、白地に赤い斜線が引かれたように見える色が 1 つあります。これは透明色です。この色を使用して、例えば、ヒストグラムで棒の枠線を非表示にすることができます。

- 最初のボタンは、塗りつぶし色を制御します。連続型または順序型のフィールドに色が関連付けられている場合は、このボタンにより、データ内の最大値に関連付けられた色の塗りつぶし色が変更されます。プロパティー・パレットの「色」タブを使用すると、最小値と欠損データに関連付ける色を変更することができます。要素の色は、基礎となるデータの値が増加するに従い、段階的に「低」の色から「高」の色に変化します。
- 2 番目のボタンは、枠線の色を制御します。
- 3 番目のボタンは、塗りつぶしパターンを制御します。塗りつぶしパターンでは枠線の色が使用されます。そのため、塗りつぶしパターンは、枠線の色が可視の場合のみ可視になります。
- 4 番目のコントロールは、塗りつぶし色とパターンの不透明度を制御するスライダーとテキスト・ボックスです。パーセンテージが低いほど不透明度が低く、透明度が高くなります。100% の場合は完全に不透明になります。

3. 罫線や線の破線化を変更するには、線ツールバーを使用します。

   注 : このツールバーには、現在選択されている項目の状態は反映されません。

   他のツールバーと同様に、ボタンをクリックして表示されるオプションを選択することも、ドロップダウンの矢印をクリックして別のオプションを選択することもできます。

ポイント要素の回転と、形状と縦横比の変更

ポイント要素を回転したり、定義済みの別の形状を割り当てたり、縦横比を変更したりすることができます。

ポイント要素の変更方法

1. ポイント要素を選択します。個別のポイント要素を回転したり、縦横比と形状を変更したりすることはできません。
2. シンボル・ツールバーを使用してポイントを変更します。
   - 最初のボタンを使用すると、ポイントの形状を変更することができます。ドロップダウンの矢印をクリックして、定義済みの形状を選択します。
   - 2 つ目のボタンを使用すると、ポイントを特定のコンパス位置まで回転することができます。ドロップダウンの矢印をクリックして、針を目的の位置にドラッグします。
   - 3 つ目のボタンを使用すると、縦横比を変更することができます。ドロップダウンの矢印をクリックし、表示される長方形をクリックしてドラッグします。この長方形の形が縦横比を表しています。

グラフィック要素のサイズの変更

視覚化では、グラフィック要素のサイズを変更することができます。グラフィック要素には、棒、線、ポイントなどがあります。グラフィック要素のサイズが変数またはフィールドによって決定される場合は、指定されたサイズが最小 サイズになります。

グラフィック要素のサイズの変更方法

1. サイズを変更するグラフィック要素を選択します。
2. スライダーを使用するか、シンボル・ツールバーで使用できるオプションに具体的なサイズを入力します。単位はピクセルです (各単位の省略形については下記を参照してください)。比率で指定することもできます (30% など)。この場合、使用可能なスペースのうちの指定された比率がグラフィック要素によって使用されることになります。使用可能なスペースは、グラフィック要素の種類や個々の視覚化によって異なります。

表 39. 有効な単位の省略形

省略形	単位
cm	センチメートル
含まれる	インチ
mm	ミリメートル
pc	パイカ
pt	ポイント
px	ピクセル

余白とパディングの指定

視覚化において枠の周囲や内側の空白が多すぎる場合や少なすぎる場合は、余白とパディングの設定を変更することができます。余白とは、枠とその周囲に配置されている他の項目との間にあるスペースの量です。パディングとは、枠線と枠の内容 の間にあるスペースの量です。

余白とパディングの指定方法

1. 余白とパディングを指定する枠を選択します。テキスト枠、凡例の周囲の枠、またはグラフィック要素 (棒や点など) を表示するデータ枠を選択できます。
2. プロパティー・パレットの「余白」タブを使用して設定値を指定します。cm や in など、別の単位を指定しない限り、サイズの単位はすべてピクセルになります。

数値の書式設定

連続型の軸の目盛りラベルや、数値を表示するデータ値ラベルの数値の書式を指定することができます。例えば、目盛りラベルに数値が 1000 単位で表示されるように指定することができます。

数値書式の指定方法

1. 数値を含む連続型の軸の目盛りラベル、またはデータ値ラベルを選択します。
2. プロパティー・パレットで「形式」タブをクリックします。
3. 必要な数値書式設定オプションを選択します。

   プレフィックス: 数値の先頭に表示される文字。例えば、数値が米ドルによる給与を表す場合は、ドル記号 ($) を入力します。

   サフィックス: 数値の最後に表示される文字。例えば、数値がパーセンテージを表す場合は、パーセント記号 (%) を入力します。

   整数の最小桁数: 10 進表記の整数部分に表示する最小桁数。実際の値が最小桁数に満たない場合は、値の整数部分に 0 が埋め込まれます。

   整数の最大桁数: 10 進表記の整数部分に表示する最大桁数。実際の値が最大桁数より大きい桁数の場合、値の整数部分はアスタリスクに置き換えられます。

   小数の最小桁数: 10 進表記または指数表記の小数部に表示する最小桁数。実際の値が最小桁数に満たない場合は、値の小数部分に 0 が埋め込まれます。

   小数の最大桁数: 10 進表記または指数表記の小数部に表示する最大桁数。実際の値が最大桁数より大きい桁数の場合、小数部は適切な桁数に丸められます。

指数表記: 数値を指数表記で表示するかどうかを指定します。指数表記は、非常に大きな数値や非常に小さな数値の場合に便利です。「**-**自動**-**」を選択すると、指数表記が適切かどうかがアプリケーションによって判別されます。

尺度変更: スケール係数。元の値がこのスケール係数で除算されます。大きな数値に合わせてラベルが長くなるのを避けたい場合に、スケール係数を使用してください。目盛りラベルの数値書式を変更する場合は、軸のタイトルを編集して、数値の解釈方法を記述してください。例えば、スケール軸に給与を表示し、ラベルが 30,000、50,000、70,000 であるとします。この場合、スケール係数として 1000 を入力すると 30、50、70 が表示されます。次に、スケール軸のタイトルを編集し、スケールが 1000 単位であることを示すテキストを入力します。

-ve の括弧: 負の値を表示するときに括弧で囲むかどうかを指定します。

グループ化: 数値の桁区切り文字を表示するかどうかを指定します。使用しているコンピューターの現在のロケールにより、数値の桁区切りに使用される文字が決まります。

軸とスケールの設定の変更

軸とスケールを変更するためのいくつかのオプションが用意されています。

軸とスケールの設定の変更方法

1. 軸の任意の部分 (軸のラベルや目盛りのラベルなど) を選択します。
2. プロパティー・パレットの「スケール」タブ、「大分割の目盛り」タブ、「小分割の目盛り」タブを使用して、軸とスケールの設定を変更します。

「スケール」タブ

注: 集計済みデータを持つグラフの場合 (ヒストグラムなど)、「スケール」タブは表示されません。

タイプ: スケールを線型にするか変換するかを指定します。スケールの変換は、データを分かりやすくする場合や、統計的推定で必要な仮定を行う場合に役立ちます。散布図では、独立変数と従属変数 (または独立フィールドと従属フィールド) の間の関係が線型ではない場合に、変換後のスケールを使用する場合があります。スケールの変換を使用して、偏ったヒストグラムを対称形にして正規分布に近づけることもできます。変換されるのは、データを表示するスケールだけです。実際のデータが変換されるわけではありません。

- 線型**:** 変換されない線型のスケールを指定します。
- **LOG:** 10 を底とする対数変換を行ったスケールを指定します。ゼロと負の値に対応するために、この変換では修正版の対数関数が使用されます。この「安全な対数」関数は `sign(x) * log(1 + abs(x))` として定義されます。そのため、`safeLog(-99)` は以下の式に等しくなります。

  ```
  sign(-99) * log(1 + abs(-99)) = -1 * log(1 + 99) = -1 * 2 = -2
  ```

- べき乗**:** 指数として 0.5 を使用するべき乗変換を行ったスケールを指定します。負の値に対応するために、この変換では修正版のべき乗関数が使用されます。この「安全なべき乗」関数は `sign(x) * pow(abs(x), 0.5)` として定義されます。そのため、`safePower(-100)` は以下の式に等しくなります。

  ```
  sign(-100) * pow(abs(-100), 0.5) = -1* pow(100, 0.5) = -1 * 10 = -10
  ```

最小**/**最大**/**適切な下限**/**適切な上限**:** スケールの範囲を指定します。「適切な下限」と「適切な上限」を選択すると、データに基づいて適切なスケールをアプリケーションで選択することができます。通常、最小値は最小のデータ値を下回る整数で、最大値は最大のデータ値を上回る整数であるため、これらの値は「適切」

です。例えば、データの範囲が 4 から 92 までである場合、スケールの適切な下限と適切な上限は、実際のデータの最小値と最大値ではなく、0 と 100 になります。設定する範囲が狭すぎると重要な項目が隠れてしまうため、注意してください。また、「ゼロを含める」オプションを選択した場合は、明示的に最小値と最大値を設定することはできません。

下の余白/上の余白: 軸の上端や下端に余白を作成します。余白は、選択した軸に対して直角に表示されます。cm や in など、別の単位を指定しない限り、単位はピクセルになります。例えば、縦軸の「上の余白」を 5 に設定すると、5 px 分の横長の余白がデータ・フレームの上部に配置されます。

反転: スケールを反転するかどうかを指定します。

ゼロを含める: スケールに 0 を含めることを指定します。通常、このオプションは棒グラフの場合に使用します。このオプションを選択すると、棒の起点が、最小の棒の高さに近い値ではなく 0 になります。このオプションを選択した場合、スケールの範囲にカスタムの最小値と最大値を設定できないため、「最小」と「最大」が使用不可になります。

「大分割の目盛り」/「小分割の目盛り」タブ

目盛り (目盛りマーク) とは、軸に表示される線のことです。この線は、特定の間隔またはカテゴリーにおける値を示します。大分割の目盛りとは、ラベル付きの目盛りマークのことです。大分割の目盛りは、他の目盛りマークよりも長くなります。小分割の目盛りとは、大分割の目盛りの間に表示される目盛りマークのことです。一部のオプションは目盛りの種類に特有のものですが、ほとんどのオプションは大分割の目盛りと小分割の目盛りで使用することができます。

目盛りを表示: 大分割の目盛りと小分割の目盛りをグラフに表示するかどうかを指定します。

グリッド線を表示: 大分割の目盛りと小分割の目盛りにグリッド線を表示するかどうかを指定します。グリッド線とは、軸の端から端までグラフ全体を横切る線のことです。

位置: 目盛りマークの、軸に対する相対位置を指定します。

長さ: 目盛りマークの長さを指定します。cm や in など、別の単位を指定しない限り、単位はピクセルになります。

基本: 大分割の目盛りにのみ適用されます。最初の大分割の目盛りを表示する値を指定します。

差分: 大分割の目盛りにのみ適用されます。大分割の目盛り同士の差を指定します。この差分の値を *n* とすると、大分割の目盛りは *n* 番目の値ごとに表示されます。

分割: 小分割の目盛りにのみ適用されます。大分割の目盛りの間における小分割の目盛りの分割数を指定します。小分割の目盛りの数は、分割の数より 1 だけ少なくなります。例えば、0 と 100 の位置に大分割の目盛りがあるとします。この場合、小分割の目盛りの分割数として 2 を入力すると、50 の位置に小分割の目盛りが *1* つ表示され、0 から 100 までの範囲が *2* つに分割されます。

カテゴリーの編集

カテゴリー軸のカテゴリーは、以下に示す方法で編集することができます。

- カテゴリーを表示するためのソート順序を変更する。
- 特定のカテゴリーを除外する。
- データ・セットに出現しないカテゴリーを追加する。
- 複数の小さなカテゴリーを 1 つのカテゴリーに省略/結合する。

カテゴリーのソート順序の変更方法

1. カテゴリー軸を選択します。「カテゴリー」パレットに軸のカテゴリーが表示されます。

   注：パレットが表示されない場合は、パレットが有効になっているかどうかを確認してください。IBM SPSS Modeler の「表示」メニューから「カテゴリー」を選択します。
2. 「カテゴリー」パレットで、ドロップダウン・リストからソート・オプションを選択します。

   ユーザー設定: パレットに表示される順序に基づいてカテゴリーをソートします。カテゴリーをリストの上部に移動するには上矢印ボタンを使用し、カテゴリーをリストの下部に移動するには下矢印ボタンを使用します。

   データ: データ・セットに出現する順序に基づいてカテゴリーをソートします。

   「名前」。パレットに表示される名前のアルファベット順にカテゴリーをソートします。ツールバーのボタンで値とラベルのいずれの表示が選択されているかにより、値とラベルのどちらかを使用してソートされます。

   値: パレットの括弧内に表示される基礎となるデータ値によってカテゴリーをソートします。このオプションがサポートされるのは、メタデータを持つデータ・ソース (IBM SPSS Statistics データ・ファイルなど) だけです。

   統計: 各カテゴリーについて計算された統計量に基づいてカテゴリーをソートします。統計量の例としては、度数、パーセンテージ、平均値などがあります。このオプションを使用できるのは、グラフで統計量を使用する場合だけです。

カテゴリーの追加方法

デフォルトでは、データ・セットに出現するカテゴリーのみ使用することができます。必要な場合は、カテゴリーを視覚化に追加することができます。

1. カテゴリー軸を選択します。「カテゴリー」パレットに軸のカテゴリーが表示されます。

   注：パレットが表示されない場合は、パレットが有効になっているかどうかを確認してください。IBM SPSS Modeler の「表示」メニューから「カテゴリー」を選択します。
2. 「カテゴリー」パレットで、以下の「カテゴリーの追加」ボタンをクリックします。

図 *71.* 「カテゴリーの追加」ボタン

3. 「新しいカテゴリーを追加」ダイアログ・ボックスで、カテゴリーの名前を入力します。
4. 「**OK**」をクリックします。

特定のカテゴリーの除外方法

1. カテゴリー軸を選択します。「カテゴリー」パレットに軸のカテゴリーが表示されます。

   注：パレットが表示されない場合は、パレットが有効になっているかどうかを確認してください。IBM SPSS Modeler の「表示」メニューから「カテゴリー」を選択します。

2. 「カテゴリー」パレットの「含める」リストでカテゴリー名を選択して「X」ボタンをクリックします。カテゴリーを元に戻すには、「除外済み」リストでそのカテゴリーの名前を選択し、リストの右側にある矢印をクリックします。

小さなカテゴリーの省略/結合方法

個別に表示する必要がないほど小さなカテゴリーを結合することができます。例えば、円グラフに多くのカテゴリーがある場合は、パーセンテージが 10 未満のカテゴリーを省略することをお勧めします。省略できるのは、加法的な統計量の場合だけです。例えば、平均値は加法的でないため、平均値同士を加算することはできません。そのため、平均値を使用しているカテゴリーを結合/省略することはできません。

1. カテゴリー軸を選択します。「カテゴリー」パレットに軸のカテゴリーが表示されます。

   注：パレットが表示されない場合は、パレットが有効になっているかどうかを確認してください。IBM SPSS Modeler の「表示」メニューから「カテゴリー」を選択します。
2. 「カテゴリー」パレットで「折りたたみ」を選択し、パーセンテージを指定します。合計のパーセンテージが指定値に満たないすべてのカテゴリーが 1 つのカテゴリーに結合されます。このパーセンテージは、グラフに表示される統計量に基づきます。省略できるのは、度数ベースと合計の統計量の場合だけです。

「方向」パネルの変更

視覚化でパネルを使用する場合は、パネルの方向を変更することができます。

パネルの方向の変更方法

1. 視覚化の任意の部分を選択します。
2. プロパティー・パレットで「パネル」タブをクリックします。
3. 「レイアウト」で、以下のいずれかのオプションを選択します。

テーブル: 表に類似したレイアウトでパネルを配置します。このレイアウトでは、行または列がすべての個別の値に割り当てられます。

入れ換え: 表に類似したレイアウトでパネルを配置しますが、元の行と列が入れ替えられます。このオプションは、グラフ自体の入れ換えとは異なります。このオプションを選択しても、x 軸と y 軸は変更されません。

リスト: リストに類似したレイアウトでペインを配置します。このレイアウトでは、各セルが値の組み合わせを表します。列と行は、個々の値には割り当てられません。このオプションを使用すると、必要に応じてパネルを折り返すことができます。

座標系の変換

多くの視覚化は、平面の直交座標系で表示されます。座標系は、必要に応じて変換することができます。例えば、座標系に極座標変換を適用し、斜めの立体の影付け効果を追加して、軸を入れ替えることができます。これらの変換が既に現在の視覚化に適用されている場合は、元に戻すことができます。例えば、円グラフは極座標系で描画されますが、この極座標変換を元に戻し、円グラフを直交座標系の 1 本の積み上げ棒グラフとして表示することができます。

座標系の変換方法

1. 変換する座標系を選択します。座標系を選択するには、個々のグラフを囲む枠を選択します。
2. プロパティー・パレットで「座標」タブをクリックします。

3. 座標系に適用したい変換を選択します。変換を選択解除して元に戻すこともできます。

   入れ換え: 軸の方向を変更することを入れ替えと呼びます。これは、2 次元の視覚化における縦軸と横軸の入れ替えに類似しています。

   極座標: 極座標変換では、グラフの中心から特定の角度と距離でグラフィック要素が描画されます。円グラフは極座標変換による 1 次元の視覚化であり、個々の棒が特定の角度で描画されます。レーダー・チャートは極座標変換による 2 次元の視覚化であり、グラフの中心から特定の角度と距離でグラフィック要素が描画されます。3 次元の視覚化では深さの次元も追加されます。

   斜交: 斜交変換は、グラフィック要素に 3-D 効果を追加します。この変換はグラフィック要素に深さを追加しますが、深さは単なる装飾にすぎません。特定のデータ値の影響を受けることはありません。

   同一比: 同一比を適用すると、それぞれのスケールの距離がデータ値の差に対応するように指定されます。例えば、両方スケールで 2cm が 1000 の差を表すように指定されます。

   変換前差し込み率: 変換後に軸が切り詰められてしまう場合は、変換の適用前にグラフに差し込みを追加することをお勧めします。この差し込みにより、一定の割合だけ大きさが縮小されてから、座標系に対して変換が適用されます。次元は、最小 x、最大 x、最小 y、最大 y の順に制御することができます。

   変換後差し込み率: グラフの縦横比を変更する場合は、変換の適用前にグラフに差し込みを追加することができます。この差し込みにより、座標系に対する変換が適用されてから、一定の割合だけ大きさが縮小されます。グラフに変換を適用しない場合でも、これらの差し込みを適用することができます。次元は、最小 x、最大 x、最小 y、最大 y の順に制御することができます。

統計量とグラフィック要素の変更

グラフィック要素を別のタイプに変換したり、グラフィック要素の描画で使用される統計量を変更したり、グラフィック要素が重なる場合の処理方法を決定する衝突変更子を指定したりすることができます。

グラフィック要素の変換方法

1. 変換したいグラフィック要素を選択します。
2. プロパティー・パレットで「要素」タブをクリックします。
3. 新しいグラフィック要素のタイプを「タイプ」リストから選択します。

表 *40.* グラフ要素の種類

グラフ要素の種類	説明
ポイント	特定のデータ・ポイントを示すマーカー。ポイント要素は、散布図やその他関連する視覚化で使用されます。
間隔	特定のデータ値で描画され、始点と別のデータ値との間の領域を埋める四角形の領域。区間要素は、棒グラフおよびヒストグラムで使用されます。
線	データ値を接続する線。
パス	データ・セットに表示された順にデータ値を接続する線。
領域	線と始点の間の領域を満たし、データ要素を接続する線。
多角形	データ領域を囲む多角形。多角形要素は、分割された散布図またはマップで使用できます。
スキーマ	外れ値を示すひげとマーカーを持つボックスで構成された要素。スキーマ要素は Boxplot に使用されます。

統計量の変更方法

1. 統計量を変更したいグラフィック要素を選択します。
2. プロパティー・パレットで「要素」タブをクリックします。

衝突変更子の指定方法

衝突変更子は、グラフィック要素が重なる場合の処理方法を決定します。

1. 衝突変更子を指定したいグラフィック要素を選択します。
2. プロパティー・パレットで「要素」タブをクリックします。
3. 「修飾子」ドロップダウン・リストで衝突変更子を選択します。「-自動-」を選択すると、グラフィック要素のタイプと統計量に適した衝突変更子がアプリケーションによって決定されます。

   オーバーレイ: 複数のグラフィック要素が同じ値を持っている場合に、それらのグラフィック要素を相互に重ねて描画します。

   スタック: 複数のグラフィック要素が同じ値を持っている場合に、通常は重ね合わせて描画するグラフィック要素を積み上げて描画します。

   ドッジ: グラフィック要素を重ね合わせるのではなく、同じ値の位置に表示される他のグラフィック要素の横にグラフィック要素を移動します。この場合、グラフィック要素は対称に配置されます。つまり、グラフィック要素が、中央の位置を挟んで正反対の位置に移動されることになります。ドッジは、クラスター化に非常に類似しています。

   積み重ね: グラフィック要素を重ね合わせるのではなく、同じ値の位置に表示される他のグラフィック要素の横にグラフィック要素を移動します。この場合、グラフィック要素は非対称に配置されます。つまり、グラフィック要素が別のグラフィック要素の上部に積み上げられ、一番下のグラフィック要素がスケール上の特定の値の位置に配置されることになります。

   ジッター **(**正規**) (Jitter (normal))**: 同じデータ値の位置にある複数のグラフィック要素を、正規分布を使用してランダムに再配置します。

   ジッター **(**一様**) (Jitter (uniform))**: 同じデータ値の位置にある複数のグラフィック要素を、一様分布を使用してランダムに再配置します。

凡例の位置の変更

グラフに凡例が含まれている場合、通常、凡例はグラフの右側に表示されます。この位置は必要に応じて変更することができます。

凡例の位置の変更方法

1. 凡例を選択します。
2. プロパティー・パレットで「凡例」タブをクリックします。
3. 位置を選択します。

視覚化と視覚化データのコピー

「全般」パレットには、視覚化とそのデータをコピーするためのボタンが用意されています。

図 72. 「視覚化をコピー」ボタン

視覚化のコピー: このアクションは、視覚化を画像としてクリップボードにコピーします。複数の画像形式を使用することができます。画像を別のアプリケーションに貼り付ける際に、「形式を選択して貼り付け」オプションを選択して、貼り付けで使用可能ないずれかの画像形式を選択することができます。

図 73. 「視覚化データをコピー」ボタン

視覚化データのコピー: このアクションは、視覚化の描画で使用された基礎データをコピーします。このデータは、プレーン・テキストまたは HTML 形式のテキストとしてクリップボードにコピーされます。このデータを別のアプリケーションに貼り付ける際に、「形式を選択して貼り付け」オプションを選択して、これらのいずれかの形式を選択して貼り付けることができます。

グラフボード エディタのキーボード ショートカット

表 41. キーボードショートカット

ショートカット・キー	関数
Ctrl+Space	探索モードと編集モードの切り替え
Delete	視覚化の項目の削除
Ctrl+Z	元に戻す
Ctrl+Y	やり直し
[F2] キー	グラフの項目を選択するためのアウトラインの表示

表題と脚注の追加

すべてのグラフのタイプで、独自のタイトル、脚注、または軸ラベルを追加して、グラフに表示されているものを識別することができます。

グラフへのタイトルの追加

1. メニューから「編集」>「グラフ・タイトルの追加」を選択します。**<TITLE>** と入力されているテキスト・ボックスが、グラフの上に表示されます。
2. 編集モードであることを確認します。メニューから「表示」>「編集モード」を選択します。
3. **<TITLE>** テキストをダブルクリックします。
4. 希望のタイトルを入力し、Return キーを押します。

グラフへの脚注の追加

1. メニューから「編集」>「グラフ脚注の追加」を選択します。**<FOOTNOTE>** と入力されているテキスト・ボックスが、グラフの下に表示されます。
2. 編集モードであることを確認します。メニューから「表示」>「編集モード」を選択します。

3. **<FOOTNOTE>** テキストをダブルクリックします。
4. 希望のタイトルを入力し、Return キーを押します。

グラフのスタイル・シートの使用

色、フォント、記号、線の太さなど、グラフの表示に関する基本的な情報はスタイル・シートで管理されます。IBM SPSS Modeler が提供するデフォルトのスタイル・シートがありますが、必要に応じて変更することができます。例えば、グラフで使用するプレゼンテーション用の企業の色を設定することができます。詳しくは、トピック 302 ページの『視覚化の編集』を参照してください。

グラフ作成ノードでは、編集モードを使用してグラフの外観のスタイルを変更できます。「編集」>「スタイル」メニューを使用して、現在のグラフ・ノードから今後生成するすべてのグラフに適用するスタイル・シートとして変更内容を保存することも、IBM SPSS Modeler を使用して生成するすべてのグラフの新しいデフォルトのスタイル・シートとして変更内容を保存することもできます。

「編集」メニューの 「スタイル」 オプションで利用可能なスタイル・シート・オプションが 5 つあります。

- 「スタイルシートを切り替え」。さまざまな格納済みスタイル・シートのリストが表示され、スタイル・シートを選択してグラフの外観を変更できます。詳しくは、トピック『スタイル・シートの適用』を参照してください。
- 「ノードにスタイルを保管」。選択したグラフのスタイルに対する変更を保管します。変更は、現在のストリームで同じグラフ・ノードから作成する今後のグラフに適用されます。
- 「デフォルトとしてスタイルを格納」。選択したグラフのスタイルに対する変更を保管します。変更は、任意のストリームで任意のグラフ・ノードから作成する今後のグラフに適用されます。このオプションを選択すると、「デフォルト スタイルの適用」 を選択して同じスタイルを使用するいかなる既存のグラフも変更することができます。
- 「デフォルト・スタイルの適用」。選択したグラフのスタイルを、現在のデフォルトとして保存されているスタイルに変更します。
- 「元のスタイルを適用」。グラフのスタイルを、元のデフォルトとして提供されたスタイルに戻します。

スタイル・シートの適用

視覚化のスタイル上のプロパティーを指定する視覚化スタイル・シートを適用することができます。例えば、スタイル・シートはその他のオプションから、フォント、ダッシュ、色を定義できます。ある程度まで、スタイル・シートでは、手動で実行する必要がある、編集用のショートカットが用意されています。ただし、スタイル・シートは「スタイル」の変更に限られています。凡例の場所または尺度領域などのその他の変更は、スタイル・シートに保存されません。

スタイル・シートの適用方法

1. メニューから次の項目を選択します。

   「編集」 > 「スタイル」 > 「スタイルシートを切り替え」
2. 「スタイルシートを切り替え」ダイアログ・ボックスを使用して、スタイル・シートを選択します。
3. ダイアログを閉じずに「適用」 をクリックして、スタイル・シートを視覚化に適用します。「**OK**」をクリックして、スタイル・シートを適用し、ダイアログ・ボックスを閉じます。

「スタイル・シートの切り替え/選択」ダイアログ・ボックス

ダイアログ・ボックスの上部のテーブルには、現在使用できる視覚化スタイル・シートがすべて表示されています。事前にインストールされているスタイル・シートがありますが、IBM SPSS Visualization Designer (別製品) で作成されているスタイル・シートもあります。

ダイアログ・ボックスの下部には、サンプル・データを含む視覚化の例が表示されています。スタイル・シートのいずれかを選択して、視覚化の例にスタイルを適用します。これらの例を使用して、実際の視覚化にスタイル・シートがどのように影響を与えるかを確認できます。

ダイアログ・ボックスには、次のオプションもあります。

既存のスタイル。デフォルトでは、スタイル・シートは視覚化のすべてのスタイルを上書きできます。この動作は変更できます。

- すべてのスタイルを上書き。スタイル・シートを適用すると、現在編集中のセッションにおいて変更された視覚化のスタイルを含め、視覚化のすべてのスタイルが上書きされます。
- 変更したスタイルを保持。スタイル・シートの適用時に、現在の編集セッション中に視覚化で変更されなかった スタイルだけが上書きされます。現在の編集セッションで変更されたスタイルは保持されます。

「管理」。コンピューターで視覚化テンプレート、スタイル・シート、およびマップを管理します。視覚化テンプレート、スタイル・シート、およびマップをローカル・マシンでインポート、エクスポート、名前変更、および削除できます。詳しくは、トピック 231 ページの『テンプレート、スタイル・シート、マップ・ファイルの管理』を参照してください。

「場所」。視覚化テンプレート、スタイル・シート、およびマップが保管されている場所を変更します。現在の場所は、ボタンの右側に表示されます。 詳しくは、トピック 230 ページの『テンプレート、スタイル・シート、マップの位置の設定』を参照してください。

グラフの印刷、保存、コピー、およびエクスポート

各グラフには多くのオプションが用意されており、グラフの保存や印刷、別の形式でのエクスポートを行うことができます。これらのオプションのほとんどは「ファイル」メニューから利用できます。さらに、別のアプリケーションで使用するために、「編集」メニューから、グラフ、そのグラフ内のデータ、または Microsoft Office 描画オブジェクトを選択してコピーできます。

印刷中

グラフを印刷するには、「印刷」 メニューを選択するかボタンを使用します。印刷する前に、「ページ設定」 と 「印刷プレビュー」 を使用して、印刷オプションを設定したり、出力をプレビューすることができます。

グラフの保存

グラフを IBM SPSS Modeler 出力ファイル (.cou) に保存するには、メニューから「ファイル」>「保存」または「ファイル」>「名前を付けて保存」を選択します。

または

グラフをリポジトリーに保存するには、「ファイル」>「出力を格納」を選択します。

グラフのコピー

MS Word や MS PowerPoint など他のアプリケーションで使用するグラフをコピーするには、「編集」> 「グラフのコピー」 を選択します。

データのコピー

MS Excel や MS Word など他のアプリケーションで使用するデータをコピーするには、「編集」>「データのコピー」 を選択します。デフォルトでは、データの形式は HTML になります。貼り付け時に他の形式オプションを表示するには、貼り付け先のアプリケーションの 「形式を選択して貼り付け」 を使用します。

Microsoft Office グラフィック オブジェクトのコピー

グラフを Microsoft Office グラフィック オブジェクトとしてコピーし、Excel や PowerPoint などの Microsoft Office アプリケーションで使用できます。グラフをコピーするには、メニューから「編集」>「**Microsoft Office** グラフィック オブジェクトをコピー」を選択します。内容がクリップボードにコピーされ、デフォルトではバイナリ形式になります。貼り付け時に、その他のフォーマット・オプションを指定するには、Microsoft Office アプリケーションで「形式を選択して貼り付け」を使用します。

なお、内容によってはこの機能がサポートされないことがあり、その場合は、「**Microsoft Office** グラフィック オブジェクトをコピー」メニュー・オプションが無効になります。また、Office アプリケーションに貼り付けた後のグラフの外観が異なることがありますが、グラフのデータは同じです。

Excel にコピーして貼り付けできるグラフ出力は、6 種類 (棒グラフ、積み上げ棒グラフ、単純な箱ひげ図 (Boxplot)、クラスタ箱ひげ図、単純な散布図、グループ化散布図) あります。これらのグラフ タイプにパネルおよびアニメーションのオプションを使用している場合、「**Microsoft Office** グラフィック オブジェクトをコピー」オプションは SPSS Modeler で無効になります。その他の設定 (オプションの外観やオーバーレイなど) については、このオプションは部分的にサポートされています。詳しくは、以下の表を参照してください。

表 42. 「*Microsoft* グラフィック オブジェクトのコピー」のサポート

グラフ出力テンプレート	**Modeler** グラフ作成ノード	**Modeler** グラフ タイプ	基本設定	オプションの外観	オーバーレイ	**Microsoft** グラフィック オブジェクトのコピーのサポート	コメント
棒グラフ	グラフボード	バー	はい	いいえ	N/A	はい	
		度数の棒グラフ	はい	いいえ	N/A	はい	
	分布	バー	はい	N/A	いいえ	はい	

表 42. 「*Microsoft* グラフィック オブジェクトのコピー」のサポート (続き)

グラフ出力テンプレート	**Modeler** グラフ作成ノード	**Modeler** グラフ タイプ	基本設定	オプションの外観	オーバーレイ	**Microsoft** グラフィック オブジェクトのコピーのサポート	コメント
積み上げ棒グラフ	グラフボード	バー	はい	はい	N/A	はい (制限付き)	オプションの外観内のカテゴリー変数のみサポート対象。
		度数の棒グラフ	はい	はい	N/A	はい (制限付き)	オプションの外観内のカテゴリー変数のみサポート対象。
	分布	バー	はい	N/A	はい	はい	
Boxplot	グラフボード	Boxplot	はい	いいえ	N/A	はい (制限付き)	Windows でのみサポート。
		Boxplot	はい	はい	N/A	いいえ	
クラスタ箱ひげ図	グラフボード	クラスタ箱ひげ図	はい	いいえ	N/A	はい (制限付き)	Windows でのみサポート。
		クラスタ箱ひげ図	はい	はい	N/A	いいえ	
単純な散布図	グラフボード	バブル・プロット	はい	いいえ	N/A	はい (制限付き)	X と Y の両方のフィールド内の連続変数、およびサイズ内のカテゴリー変数のみサポート対象。
		散布図	はい	いいえ	N/A	はい (制限付き)	X と Y の両方のフィールド内の連続変数のみサポート対象。
	作図	ポイント	はい	N/A	いいえ	はい (制限付き)	X と Y の両方のフィールド内の連続変数のみサポート対象。

表 42. 「*Microsoft* グラフィック オブジェクトのコピー」のサポート (続き)

グラフ出力テンプレート	**Modeler** グラフ作成ノード	**Modeler** グラフ タイプ	基本設定	オプションの外観	オーバーレイ	**Microsoft** グラフィック オブジェクトのコピーのサポート	コメント
グループ化散布図	グラフボード	バブル・プロット	はい	はい	N/A	いいえ	
		散布図	はい	はい	N/A	はい (制限付き)	X と Y の両方のフィールド内の連続変数、およびオプションの外観内のカテゴリー変数のみサポート対象。
	作図	ポイント	はい	N/A	はい	はい (制限付き)	X と Y の両方のフィールド内の連続変数、および「オーバーレイ」オプション内のカテゴリー変数のみサポート対象。

グラフのエクスポート

「グラフのエクスポート」オプションを使用すると、グラフを他のアプリケーションで使用するために、ビットマップ (.bmp)、JPEG (.jpg)、PNG (.png)、HTML (.html)、PDF (.pdf)、または ViZml 文書 (.xml) のいずれかの形式でエクスポートできます。

注: PDF オプションを選択すると、グラフィックのサイズにトリミングされた高解像度の PDF ファイルとして、グラフがエクスポートされます。

グラフをエクスポートするには、「ファイル」>「グラフのエクスポート」 を選択して、形式を選択します。

テーブルのエクスポート

「テーブルのエクスポート」 オプションを使用すると、タブ区切り (tab)、カンマ区切り (csv)、HTML (html) のいずれかの形式でテーブルをエクスポートできます。

テーブルをエクスポートするには、「ファイル」>「テーブルのエクスポート」 を選択して、形式を選択します。

第 6 章 出力ノード

出力ノードの概要

出力ノードを利用すれば、データやモデルに関する情報を取得できます。出力ノードは、他のソフトウェアツールのインターフェースに対応した、さまざまな形式でデータをエクスポートできるメカニズムも備えています。

次の出力ノードを利用できます。

テーブル ノードは、データをテーブル形式で表示します。このデータは、ファイルにも書き込めます。この機能は、データの値を調査したり、データを読みやすい形式でエクスポートする必要がある場合に役立ちます。

クロス集計ノードで、フィールド間の関係を示すテーブルを作成します。一般的にこのノードは、2 つのシンボル値フィールドの関係を示す場合によく使用されますが、フラグ型フィールド間または数値型フィールド間の関係を示すこともできます。

精度分析ノードで、予測モデルの能力を評価して正確な予測を生成します。分析ノードでは、1 つ以上のモデル・ナゲットについて、予測値と実際値をさまざまな方法で比較します。また、分析ノードでは予測モデル同士を比較できます。

データ検査ノードでは、欠損値、外れ値、および極値に関する情報の他、各フィールドの要約統計量、ヒストグラムや棒グラフを含む、データを広範に検査するための手段を提供しています。結果は把握しやすい行列形式で表示され、ソートしたり、フルサイズのグラフやデータ準備ノードを生成することができます。

変換ノードによって、選択フィールドに適用する前に変換の結果を選択し、視覚的に確認することができます。

記述統計ノードでは、数値型フィールドに関する基本的な集計情報が提供されます。このノードで、個々のフィールドの要約統計量とフィールド間の相関が計算されます。

平均比較ノードでは、独立したグループ間で、または関連するフィールドのペア間で著しい違いがあるかどうかを調べるために、平均を比較します。例えば、販売促進活動の前後で平均収益を比較したり、販売促進活動を受けなかった顧客と受けた顧客からの収益を比較することができます。

レポート・ノードで、固定テキスト、およびデータやデータから導かれた他の式を含む、フォーマット済みレポートを作成します。レポートの書式は、固定テキストとデータの出力構成を定義するテキスト テンプレートを使用して指定します。テンプレート内の HTML タグを使用し、また「出力」タブでオプションを設定することで、カスタムのテキスト書式設定を提供できます。テンプレート内の CLEM 式を使用して、データ値やその他の条件出力を含めることができます。

グローバル・ノードで、データを走査し、CLEM 式で使用できる要約値を算出します。例えば、グローバル・ノードを使用して、「年齢」 という名前のフィールドの統計量を算出し、次に CLEM 式に `@GLOBAL_MEAN(年齢)` 関数を挿入して年齢 の全体的な平均を算出することができます。

シミュレーション・フィッティング・ノードは、各フィールドのデータの統計的な分布を調べ、最も適合する分布を各フィールドに割り当ててシミュレーション生成ノードを生成 (または更新) します。この後、シミュレーション生成ノードを使用して、シミュレートするデータを生成することができます。

シミュレーション評価ノードは、指定された予測される対象フィールドを評価し、対象フィールドの分布と相関情報を提供します。

出力の管理

出力マネージャは、IBM SPSS Modeler のセッション中に生成されるチャート、グラフ、テーブルを表示します。マネージャ内でダブルクリックすると、いつでも再表示することができます。該当するストリームやノードに戻る必要はありません。

出力マネージャーを表示するには

「表示」メニューを開き、「マネージャー」 を選択します。「出力」 タブをクリックします。

出力マネージャーから、次のことができます。

- ヒストグラム、評価グラフ、およびテーブルなどの既存の出力オブジェクトを表示する。
- 出力オブジェクトの名前を変更する。
- 出力オブジェクトをディスク、または IBM SPSS Collaboration and Deployment Services Repository (可能ならば) に保存する。
- 現在のプロジェクトに出力ファイルを追加する。
- 現在のセッションから未保存の出力オブジェクトを削除する。

- 出力オブジェクトを IBM SPSS Collaboration and Deployment Services Repository (可能ならば) に保存、またはそこから出力オブジェクトを取得する。

オプションにアクセスするには、「出力」タブの任意の場所で右クリックします。

出力を表示

オンスクリーン出力は、出力ブラウザー・ウィンドウに表示されます。出力ブラウザー・ウィンドウは、それ自体に出力を印刷または保存、または別の形式で出力をエクスポートできるメニュー・セットを備えています。具体的なオプションは、出力のタイプによって大きく異なることに注意してください。

データの印刷、保存、エクスポート：以下に詳しく説明します。

- 出力を印刷するには、「印刷」 メニュー・オプションを選択するかボタンを使用します。印刷する前に、「ページ設定」 と 「印刷プレビュー」 を使用して、印刷オプションを設定したり、出力をプレビューすることができます。
- 出力を IBM SPSS Modeler 出力ファイル (.cou) に保存する場合、「ファイル」メニューから 「保存」 または 「名前を付けて保存」 を選択します。
- テキストや HTML などの他の形式で出力を保存するには、「ファイル」メニューから 「エクスポート」 を選択します。 詳しくは、トピック 325 ページの『出力のエクスポート』を参照してください。

  これらの形式を選択できるのは、出力に含まれるデータをその形式でエクスポートする意味がある場合のみであることに注意してください。例えば、ディシジョン ツリーの内容はテキストとしてエクスポートできますが、K-means モデルの内容はテキストとして意味をなしません。
- ほかのユーザーが IBM SPSS Collaboration and Deployment Services Deployment Portal を使用して出力を表示できるよう共有リポジトリーに出力を保存するには、「ファイル」メニューから 「**Web** に公開」 を選択します。このオプションには、別途 IBM SPSS Collaboration and Deployment Services のライセンスが必要なことに注意してください。

セルと列の選択:「編集」メニューには、現在の出力形式に合わせて、選択、選択解除、セルや列のコピーに関するさまざまなオプションが含まれています。 詳しくは、 325 ページの『セルと列の選択』を参照してください。

ノードの生成:「生成」メニューでは、出力ブラウザーの内容に基づいて新しいノードを生成することができます。オプションは、出力形式や、現在選択されている出力内の項目によって大きく異なります。特殊な出力形式のノード生成オプションの詳細については、該当する出力の説明書を参照してください。

Web に公開

「Web に公開」を選択して、特定の種類のストリームの出力を IBM SPSS Collaboration and Deployment Services の基礎を形成する中央共有 IBM SPSS Collaboration and Deployment Services Repository に公開できます。このオプションを使用すると、出力を表示する必要があるそのほかのユーザーはインターネット・アクセスおよび IBM SPSS Collaboration and Deployment Services アカウントを使用して出力を表示できます。IBM SPSS Modeler をインストールする必要はありません。

次のテーブルに、「Web に公開」機能をサポートする IBM SPSS Modeler ノードを示します。これらのノードからの出力は、IBM SPSS Collaboration and Deployment Services Repository に出力オブジェクト (*.cou*) 形式で保存され、IBM SPSS Collaboration and Deployment Services Deployment Portal で直接表示できます。

その他の種類の出力は、関連するアプリケーション (ストリーム・オブジェクトの場合は IBM SPSS Modeler) が、ユーザーのコンピューターにインストールされている場合にのみ表示できます。

表 *43. Web* への公開をサポートしているノード：

ノード・タイプ	ノード
グラフ作成	all
出力	テーブル
	クロス集計
	データ検査
	変換
	平均値
	分析
	記述統計
	レポート (HTML)
IBM SPSS Statistics	Statistics 出力

出力を Web に公開する

出力を Web に公開する手順は次のとおりです。

1. IBM SPSS Modeler では、テーブルに表示されたノードの 1 つを実行します。ノードの 1 つを実行すると、新しいウィンドウで出力オブジェクトを作成します (テーブル、マトリックス、レポート・オブジェクトなど)。
2. 「出力オブジェクト」ウィンドウから次の項目を選択します。

   「ファイル」 > 「**Web** に公開」

   注：標準の Web ブラウザーで使用するために単純な HTML ファイルをエクスポートする場合、「ファイル」メニューから 「エクスポート」 を選択した後、「**HTML**」 を選択します。
3. IBM SPSS Collaboration and Deployment Services Repository への接続

   正常に接続されると、さまざまなストレージ オプションを提供する「リポジトリー: 保存」ダイアログが表示されます。
4. ストレージ オプションを選択したら、「格納」 をクリックします。

公開出力の Web 表示

この機能を使用するには、IBM SPSS Collaboration and Deployment Services アカウントのセットアップが必要です。表示するオブジェクト・タイプの関連するアプリケーション (IBM SPSS Modeler または IBM SPSS Statistics) がインストールされている場合、出力はブラウザーではなくアプリケーションで表示されます。

公開出力を Web で表示する手順は次のとおりです。

1. ブラウザーに *http://<repos_host>:<repos_port>/peb*

   を指定します。*repos_host* および *repos_port* は IBM SPSS Collaboration and Deployment Services ホストのホスト名およびポート番号です。
2. IBM SPSS Collaboration and Deployment Services アカウントのログインの詳細を入力してください。

3. 「コンテンツ・リポジトリー」 をクリックします。
4. 表示するオブジェクトに移動または検索します。
5. オブジェクト名をクリックします。グラフなど一部のオブジェクト・タイプについて、オブジェクトをブラウザーで表示するときに遅延が生じる場合があります。

HTML ブラウザーで出力結果を表示

線形、ロジスティックおよび因子分析モデル・ナゲットの「詳細」タブで、Internet Explorer などの各ブラウザーに情報を表示することができます。情報は HTML として出力され、保存して、企業のイントラネットやインターネット・サイトなどあらゆる場所で再利用することができます。

ブラウザーで情報を表示するには、「起動」ボタンをクリックします。このボタンは、モデル ナゲットの「詳細」タブの左上、モデル アイコンの下にあります。

出力のエクスポート

出力ブラウズ ウィンドウで、テキストや HTML などの他の形式に出力をエクスポートするよう選択できます。エクスポート形式は、出力形式によって大きく異なりますが、一般にその出力を生成するために使用したノードで 「ファイルに保存」 を選択すると、利用できるファイル形式オプションは似たものとなります。

注**:** これらの形式を選択できるのは、出力に含まれるデータをその形式でエクスポートする意味がある場合のみです。例えば、ディシジョン ツリーの内容はテキストとしてエクスポートできますが、K-means モデルの内容はテキストとして意味をなしません。

出力をエクスポートするには

1. 出力ブラウザーで、「ファイル」メニューを開き、「エクスポート」 を選択します。次に生成するファイル形式を選択します。
   - タブ区切り **(*.tab):** データ値を含む、フォーマット済みのテキスト・ファイルを生成します。このスタイルは、他のアプリケーションにインポートできる、プレーン・テキストで表した情報を生成する場合に役立ちます。このオプションは、テーブル・ノード、クロス集計、および平均値ノードで使用できます。
   - カンマ区切り **(*.dat):** データ値を含む、カンマで区切られたテキスト・ファイルを生成します。このスタイルでは、表計算アプリケーションやデータ分析アプリケーションにインポートできる形式のデータ・ファイルをすばやく生成できます。このオプションは、テーブル・ノード、クロス集計、および平均値ノードで使用できます。
   - 移行タブ区切り **(*.tab):** このオプションは、「タブ区切り」オプションとまったく同じですが、行がフィールドを表し、列がレコードを表すように、データの行列入れ替えが行われます。
   - 移行カンマ区切り **(*.dat):** このオプションは、「カンマ区切り」オプションとまったく同じですが、行がフィールドを表し、列がレコードを表すように、データの行列入れ換えが行われます。
   - **HTML (*.html):** このオプションは、HTML 形式の出力を 1 つ以上のファイルに書き込みます。

セルと列の選択

テーブル・ノード、クロス集計ノード、平均値ノードを含む多くのノードがテーブル形式の出力を生成します。セルの選択、テーブル全体または一部のクリップボードへのコピー、現在の選択を基に新規ノードの生成、およびテーブルの保存と印刷を含む同じような方法でこれらの出力テーブルを表示し、操作することができます。

セルの選択: セルを選択するには、そのセルをクリックします。複数のセルを範囲として選択するには、目的の範囲の一方の角をクリックした後、その範囲の対角までマウスをドラッグしてマウス・ボタンを放します。列全体を選択するには、列見出しをクリックします。複数の列を選択するには、Shift キーまたは Ctrl キーを押しながら列見出しをクリックします。

新しく選択すると、その前の選択は取り消されます。Ctrl キーを押しながら新しく選択すると、その前の選択は取り消されず、既存の選択項目に新しい選択項目が追加されます。この方法を使用して、テーブル内の連続していない複数の領域を選択できます。「編集」メニューにも 「すべて選択」 と 「選択解除」 があります。

列の並び替え: テーブル・ノードと平均値ノードの出力ブラウザーで、列の見出しをクリックして目的の位置にドラッグすると、テーブルの列を移動することができます。列は 1 回に 1 つだけ移動することができます。

テーブル・ノード

テーブル・ノードでは、データ内の値を一覧表示するテーブルを作成します。ストリーム内のすべてのフィールドおよびすべての値が含まれ、データ値を容易に調査したり読みやすい形式でデータ値をエクスポートすることができます。オプションで、特定の条件を満たすレコードを強調表示することができます。

注: 小規模なデータ・セットを処理しているのでない限り、テーブル・ノードに渡すにはデータのサブセットを選択することをお勧めします。レコード数が画面構造に収めることができるサイズを超えている場合 (例えば 1 億行であるなど)、テーブル・ノードでは正しく表示することができません。

テーブル・ノードの「設定」タブ

レコードの強調表示: テーブル内のレコードを強調表示するために、対象のレコードを真にする CLEM 式を入力します。このオプションは、「画面に出力」 が選択されている場合に有効になります。

テーブル・ノードの「形式」タブ

「形式」タブには、形式をフィールド単位で指定するための設定が用意されています。このタブは、データ型ノードと共有です。 詳しくは、トピック 157 ページの『フィールド形式の「設定」タブ』を参照してください。

出力ノードの「出力」タブ

表形式の出力を生成するノードの場合は、「出力」タブで結果の形式と出力先を指定することができます。

出力名。ノードの実行時に生成される出力の名前を指定します。「自動」 は、出力を生成するノードの名前に基づいて名前を選択します。「ユーザー設定」 で別の名前を指定することもできます。

画面に出力 (デフォルト)。オンラインで表示するための出力を作成します。出力ノードを実行すると、出力オブジェクトはマネージャー・ウィンドウの「出力」タブに表示されます。

ファイルに出力。ノードの実行時に、出力をファイルに保存します。このオプションを選択した場合は、ファイル名を入力して (または、ファイル選択ボタンを使用してディレクトリーを参照し、ファイル名を指定して)、ファイル形式を選択してください。ファイル形式には、特定の出力形式に使用できないものがありますので注意してください。

注:

出力ノードからの出力データは、以下の規則に従ってエンコードされます。

- 出力ノードの実行時に、ストリーム・エンコード値 (「ストリーム・オプション (Stream Options)」タブで設定) が出力に設定されます。
- 出力が生成された後、ストリームのエンコード方式が変更されても、テーブル出力のエンコードは変更されません。
- 出力ノード出力のエクスポート時、出力ファイルは現在定義されているストリーム・エンコードでエクスポートされます。出力の作成後は、ストリーム・エンコードを変更しても、生成済みの出力には影響しません。

上記の規則には以下の例外があります。

- すべての HTML エクスポートは UTF-8 形式でエンコードされます。
- 拡張出力ノードからの出力は、カスタム・ユーザー・スクリプトにより生成されます。そのため、エンコード方式はそのスクリプトにより制御されます。

出力をファイルに保存するために、以下のオプションが使用できます。

- データ **(**タブ区切り**) (*.tab):** データ値を含む、フォーマット済みのテキスト・ファイルを生成します。このスタイルは、他のアプリケーションにインポートできる、プレーン・テキストで表した情報を生成する場合に役立ちます。このオプションは、テーブル・ノード、クロス集計、および平均値ノードで使用できます。
- データ **(**カンマ区切り**) (*.dat):** データ値を含む、カンマで区切られたテキスト・ファイルを生成します。このスタイルでは、表計算アプリケーションやデータ分析アプリケーションにインポートできる形式のデータ・ファイルをすばやく生成できます。このオプションは、テーブル・ノード、クロス集計、および平均値ノードで使用できます。
- **HTML (*.html):** このオプションは、HTML 形式の出力を 1 つ以上のファイルに書き込みます。テーブル式出力 (テーブル・ノード、クロス集計、または平均値ノードから) の場合は、一連の HTML ファイルに フィールド名を含む内容パネルと HTML テーブルのデータが格納されます。テーブルの行数が「**1** ページの行数」 の設定値を超えた場合、1 つのテーブルが複数の HTML ファイルに分割されることがあります。このような場合は、すべてのテーブル ページに対するリンクが内容パネルに記載され、これを使用して各テーブルに移動できます。テーブル形式以外の出力の場合は、ノードの結果を含む HTML ファイルが 1 つ作成されます。

  注**:** HTML 出力に最初のページの書式だけが含まれている場合は、「出力のページ分割」 を選択し、1 ページにすべての出力が収まるように 「**1** ページの行数」 の設定を調整してください。または、レポート・ノードのようなノードの出力テンプレートにカスタムの HTML タグがある場合は、形式の種類として 「カスタム」 を選択してください。
- テキスト・ファイル **(*.txt):** 出力データを含むテキスト・ファイルを生成します。このスタイルは、ワープロやプレゼンテーション ソフトウェアなど、他のアプリケーションにインポートできる出力データを生成する場合に便利です。このオプションは、一部のノードでは利用できません。
- 出力オブジェクト **(*.cou):** この形式で保存された出力オブジェクトは、IBM SPSS Modeler で開いて表示したり、プロジェクトへの追加、IBM SPSS Collaboration and Deployment Services Repository を使用して公開し、追跡することができます。

出力ビュー**:** 平均値ノードに対してデフォルトで、シンプルまたは詳細のどちらのモードで出力を表示するかを指定できます。このビューは、生成された出力をブラウズするときに切り替えることができます。 詳しくは、トピック 349 ページの『平均値ノード出力ブラウザー』を参照してください。

形式。 レポート・ノードの場合に、出力を自動的にフォーマットするか、またはテンプレート中の HTML を使用してフォーマットするかを選択することができます。テンプレート中の HTML を使用してフォーマットする場合は、「カスタム」 を選択します。

表題。 レポート・ノードに対して、レポート出力の上部に表示するタイトル テキストを指定することができます。

挿入文字列を強調表示: レポート・ノードの場合に、このオプションを選択すると、レポート テンプレート中の CLEM 式が生成した文字列が強調表示されます。詳しくは、トピック 350 ページの『レポート・ノードの「テンプレート」タブ』を参照してください。形式に 「カスタム」 を指定した場合、このオプションの使用はお勧めできません。

ページ当たりの行数: レポート・ノードの場合に、出力レポートの自動フォーマット時に、各ページに入れる行数を指定します。

データの入れ替え: このオプションは、エクスポート前にデータの行列を入れ替えて、行がフィールドを、列がレコードを表すようにします。

注: 大きなテーブルでは、特にリモート・サーバーを使用しているような場合に、上記のオプションを利用するとパフォーマンスが低下する場合があります。そのような場合は、ファイル出力ノードを使用すれば、パフォーマンスを向上することができます。 詳しくは、トピック 385 ページの『ファイル エクスポート・ノード』を参照してください。

テーブル・ブラウザー

テーブル・ブラウザーにはテーブル形式のデータが表示されます。このブラウザーでは、セルの選択やコピー、列の並べ替え、テーブルの保存や印刷などの標準的な操作を実行することができます。詳しくは、トピック 325 ページの『セルと列の選択』を参照してください。これらは、ノードでデータをプレビューする際に実行できる同じ操作です。

テーブル・データのエクスポート: 次の項目を選択して、テーブル・ブラウザーからデータをエクスポートできます。

「ファイル」 > 「エクスポート」

詳しくは、トピック 325 ページの『出力のエクスポート』を参照してください。

データはシステム・デフォルトの文字コード形式で出力され、Windows のコントロール・パネル、または分散モードで動作している場合はサーバー・コンピューターから指定できます。

テーブルの検索: メイン・ツールバーの検索ボタン (双眼鏡のアイコン) をクリックすると、検索ツールバーがアクティブになり、テーブルで特定の値を検索することができます。テーブル中を前方または後方に検索することができます。また、大文字と小文字を区別するかどうかを指定することもできます (「**Aa**」 ボタン)。検索処理を中断するには、検索中断ボタンをクリックします。

ノードの生成:「ノード生成」メニューには、ノード生成操作に関するメニュー項目があります。

- 条件抽出ノード **(**レコード**):** テーブル内で選択したセルに対するレコードを選択する条件抽出ノードを生成します。
- 条件抽出ノード **(**「**AND**」**):** テーブル内で選択したすべて の値を含むレコードを選択する条件抽出ノードを生成します。

- 条件抽出ノード **(**「**OR**」**):** テーブル内で選択したいずれか の値を含むレコードを選択する条件抽出ノードを生成します。
- フィールド作成ノード **(**レコード**):** 新しいフラグ型フィールドを作成する、フィールド作成ノードを生成します。テーブル内で選択したセルのレコードには *T*、その他のレコードには *F* がフラグ型フィールドに含まれます。
- フィールド作成ノード **(AND):** 新しいフラグ型フィールドを作成する、フィールド作成ノードを生成します。テーブル内で選択した値をすべて 含むレコードには *T*、その他のレコードには *F* がフラグ型フィールドに含まれます。
- フィールド作成ノード **(OR):** 新しいフラグ型フィールドを作成する、フィールド作成ノードを生成します。テーブル内で選択したいずれか の値を含むレコードの場合は *T*、それ以外のレコードの場合は *F* がフラグ型フィールドに格納されます。

クロス集計ノード

クロス集計ノードでは、フィールド間の関係を示す表を作成できます。一般的にこのノードは、2 つのカテゴリー・フィールド (フラグ型、名義型、順序型) の関係を示す場合によく使われますが、連続型 (数値範囲) 型フィールド間の関係を示す場合にも使用できます。

クロス集計ノードの「設定」タブ

「設定」タブでは、行列の構造に関するオプションを指定することができます。

フィールド**:** 次のオプションで、フィールド選択タイプを選択します。

- 選択済み**:** 行列の行と列のカテゴリー・フィールドを 1 つずつ選択できます。行列の行と列は、選択したカテゴリー・フィールドの値のリストで定義されます。行列のセルには、下で選択する項目の要約統計量が入ります。
- すべてのフラグ **(**真の値**):** データ中の各フラグ型フィールドの 1 行 1 列から成る行列を要求します。行列のセルには、フラグの組み合わせごとに、両方に対して真 (true) となるレコードの合計数が入ります。例えば、「パン購入に対応する行とチーズ購入に対応する列がある場合、その行と列が交差するセルには、パン購入とチーズ購入の両方に対して真 (true) となるレコード数が含まれます。
- すべての数値**:** 各数値フィールドの 1 行 1 列で構成される行列を要求します。行列のセルは、対応する一対のフィールドのクロスする値の積の合計を表します。つまり、行列の各セルでは、レコードごとに行フィールドと列フィールドの値が乗算され、レコード全体の合計が計算されます。

欠損値を含める**:** ユーザーによる欠損値 (空白) とシステムによる欠損値 (ヌル) が、行と列の出力に含まれます。例えば、値 *N/A* が選択されたフィールドのユーザー欠損として定義されていた場合、*N/A* とラベル付けされた別の列は、その他のカテゴリーのようにそのテーブルに含まれます。このオプションが選択解除されると、その頻度に関わらず *N/A* 列は除外されます。

注 : 欠損値を含めるオプションは、選択したフィールドがクロス集計されている場合にのみ適用されます。空白値はヌルにマップされ、モードが 「選択」、コンテンツが 「関数」 に設定されている場合は、関数フィールドの集計から除外され、モードが 「すべての数値」 に設定されている場合は、すべての数値の集計から除外されます。

セルの内容**:** 上で 「選択」 フィールドを選択した場合、行列のセルで使う統計量を指定できます。度数の統計量を選択するか、またはオーバーレイ フィールド を選択して、行フィールドと列フィールドの値に基づいた数値フィールドの値を集計します。

- クロス表**:** セル値は、対応する値の組み合わせが含まれているレコードの度数やパーセンテージです。「外観」タブの設定を使用して、どのクロス表要約を使用するかを指定することができます。グローバル カイ 2 乗値も有意値と共に表示されます。 詳しくは、トピック 331 ページの『クロス集計ノードの出力ブラウザー』を参照してください。
- 関数**:** 要約関数を選択すると、適切な行値と列値を含むケースに対して、セル値が、選択したオーバーレイ フィールド値の関数になります。例えば、行フィールドが地域で列フィールドが製品、そしてオーバーレイ フィールドが収益の場合、北東地区行と部品列中のセルには、北東地域で販売された部品の収益の合計 (または平均、最小、または最大) が格納されます。デフォルトの要約関数は 「平均」 になります。関数フィールドを要約する他の関数を選択できます。オプションには、「平均値」、「合計」、「標準偏差」、「最大値」、および 「最小値」 が含まれています。

クロス集計ノードの「外観」タブ

「外観」タブでは、行列のソートや強調表示に関するオプションと、クロス表行列に表示される統計量を制御することができます。

行と列。行列中の行および列見出しを制御します。デフォルトは 「未ソート」 です。行や列見出しをソートするには、「昇順」 または 「降順」 を選択してください。

オーバーレイ。行列中の極値を強調表示できます。値は、セルの度数 (クロス表行列) または計算された値 (関数行列) に基づいて強調表示されます。

- 上位を強調表示。行列の上位 N 個の値を強調表示 (赤) できます。強調表示する値の数を指定します。
- 下位を強調表示。行列の下位 N 個の値を強調表示 (緑) できます。強調表示する値の数を指定します。

注 : これらの強調表示オプションでは、同じ値が複数ある場合、実際に強調表示したい数より多くの値が強調表示されることがあります。例えば、ゼロ値を持つセルが 6 つある行列の場合、「最下位を強調表示」 に 5 を指定すると、6 つのゼロ値すべてが強調表示されます。

クロス表セルの内容。クロス表に対して、クロス表のクロス集計の行列に含める要約統計量を指定できます。これらのオプションは、「すべての数値」 または 「関数」 オプションのどちらかが「設定」タブで選択されている場合には利用できません。

- 度数: セルには、対応する行の値を持つ列値があるレコード数が含まれます。これは、唯一のデフォルトのセルの内容です。
- 期待値。行と列の間には何も関係がないと仮定して、セルにはレコード数に対する期待値が含まれます。期待値は次の式を基準に算出されます。

  p(*row value*) * p(*column value*) * *total number of records*
- 「残差」。確認された値と期待値との差。
- 行のパーセンテージ。対応する列の値を持つ行値があるすべてのレコードのパーセンテージが含まれます。行内のパーセントの合計は 100 になります。
- 列のパーセンテージ。対応する行の値を持つ列値があるすべてのレコードのパーセンテージが含まれます。列内のパーセントの合計は 100 になります。
- 合計パーセンテージ。列値と行値の組み合わせを持つすべてのレコードのパーセンテージが含まれます。行列全体のパーセントの合計は 100 になります。
- 行および列合計を含める。列および行の合計に対して、行列に行と列を加えます。

- 設定値の適用。(出力ブラウザーのみ) 出力ブラウザーを閉じたり再度開いたりすることなく、クロス集計ノード出力の外観を変更することができます。出力ブラウザーの「設定値の適用」を使用して変更し、このボタンをクリックして「クロス集計」タブを選択し、変更の影響を確認します。

クロス集計ノードの出力ブラウザー

クロス集計ブラウザーにはクロス表形式のデータが表示され、セルの選択、行列全体または一部のクリップボードへのコピー、行列の選択に基づく新しいノードの生成、行列の保存や印刷など、行列に関する操作を実行することができます。クロス集計ブラウザーを使用して、Oracle の Naive Bayes モデルなど、特定のモデルからの出力を表示することもできます。

「ファイル」と「編集」メニューは、印刷、保存、出力のエクスポート、データの選択とコピーの通常のオプションを提供します。詳しくは、トピック 323 ページの『出力を表示』を参照してください。

カイ **2** 乗: 2 つのカテゴリー・フィールドのクロス集計表の場合、そのテーブルの下に大域的な Pearson カイ 2 乗も表示されます。この検定は、2 つのフィールドが関係がないことの確率を、確認されたカウントと関係がない場合の期待値との差に基づいて示します。例えば、顧客満足度と店舗の場所の間に関係がない場合、同様の満足度を全店舗に対して期待します。しかし、特定の店舗における顧客が、一貫して他の店舗よりも高い満足度を表明している場合は、一致性に疑念を抱くかもしれません。その差が大きくなればなるほど、ただのサンプリングのミスであったという確率は小さくなります。

- カイ 2 乗テストは、2 つのフィールドが関係がなく、その中で、確認された頻度と期待頻度との間の差は、ただの見込みであることの確率を示します。この確率が非常に小さい場合は (一般的には 5% 未満)、2 つのフィールドの間の関係が有意であると考えることができます。
- 1 列と 1 行しかない場合 (一元カイ 2 乗テスト)、自由度はセル数マイナス 1 です。二元カイ 2 乗の場合の自由度は、(行数 - 1) × (列数 - 1) になります。
- カイ 2 乗の統計量を解釈する場合、すべての期待されるセルの自由度が 5 未満であるかどうか注意します。
- カイ 2 乗テストは、2 つのフィールドがクロス表の場合にのみ利用できます。(「すべてのフラグ」 または 「すべての数値」 が「設定」タブで選択されている場合、このテストは表示されません)。

「ノードの生成」メニュー **:** 大部分のモデル・ナゲットには、「ノードの生成」メニューもあります。「ノード生成」メニューには、ノード生成操作に関するメニュー項目があります。これらの操作は、クロス表行列でしか利用できません。また、行列中のセルを最低 1 つ以上選択している必要があります。

- 条件抽出ノード: 行列で選択した任意のセルと一致するレコードを選択する、条件抽出ノードを生成します。
- フィールド作成ノード **(**フラグ型**):** 新しいフラグ型フィールドを作成する、フィールド作成ノードを生成します。行列内で選択された任意のセルに一致するレコードには *T*、その他のレコードには *F* がフラグ型フィールドに含まれます。
- フィールド作成ノード **(**セット型**):** 新しい名義型フィールドを作成する、フィールド作成ノードを生成します。名義型フィールドには、行列中の選択されたセルのそれぞれの連続セットに対して、1 つのカテゴリーが含まれます。

精度分析ノード

分析ノードでは、正確な予測を生成するためにモデルの能力を評価することができます。分析ノードでは、1 つ以上のモデル・ナゲットについて、予測値と実際値 (対象フィールド) をさまざまな方法で比較します。精度分析ノードを使用して、予測モデル同士を比較することもできます。

精度分析ノードを実行すると、実行ストリーム中のそれぞれのモデル・ナゲットに対して、「要約」タブの「精度分析」に分析結果の要約が自動的に追加されます。精度分析の詳細な結果は、マネージャー・ウィンドウの「出力」タブに表示されます。また、直接ファイルに書き込むこともできます。

注: 精度分析ノードが予測値と実際値を比較するため、この 2 つの値は監視モデル (対象フィールドが必要なモデル) でのみ有用です。クラスタリング・アルゴリズムなどの非監視モデルについては、比較のベースとして利用できる実際の結果はありません。

精度分析ノードの「精度分析」タブ

「分析」タブでは、分析の詳細を指定することができます。

一致行列 **(**シンボル対象またはカテゴリー対象**):** カテゴリー対象 (フラグ型、名義型、または順序型) の各生成 (予測) フィールドとその対象フィールド間の一致パターンを示します。実際値で構成される行と予測値で構成される列から成るテーブルが作成されます。 各セルには、そのパターンを含むレコード数が表示されます。これは、予測時の系統誤差を判別する場合に役に立ちます。異なるモデルによって生成された複数の生成フィールドが、同じ出力フィールドに関連している場合は、これらのフィールドが一致する場合と一致しない場合がカウントされ、合計に表示されます。一致する場合は、別の正/誤統計が表示されます。

パフォーマンス評価**:** カテゴリー出力を行うモデルのパフォーマンス評価統計量を表示します。この統計量は、出力フィールドの各カテゴリーに対して報告され、そのカテゴリーに属するレコードを予測するためにモデルの平均情報量 (ビット数) を測定します。分類の難しさを考慮して、まれなカテゴリーについて正確な予測を行うために、一般的なカテゴリーの予測時よりも高いパフォーマンス評価インデックスが与えられます。あるカテゴリーに関するモデルの予測が推量にすぎない場合は、そのカテゴリーのパフォーマンス評価インデックスは 0 になります。

評価メトリックス **(AUC** と **Gini**、バイナリー分類子のみ**):** バイナリ分類子の場合、このオプションは、AUC (曲線下の領域) および Gini 係数の評価メトリックを報告します。これらの評価メトリックは、いずれもそれぞれの 2 項モデルから一括して計算されます。メトリックの値は、分析出力ブラウザーに表形式で報告されます。

AUC 評価メトリックは ROC (受信者操作特性) 曲線の下の面積として計算され、分類子の予測されるパフォーマンスをスカラー値で表します。AUC は常に 0 と 1 の間であり、数値が大きいほどよい分類子であることを示します。座標 (0,0) と (1,1) の間の ROC 曲線が対角線である場合はランダムな分類子を表し、AUC は 0.5 になります。したがって、現実的な分類子の AUC が 0.5 未満になることはありません。

Gini 係数の評価メトリックは、AUC の代わりとなる評価メトリックとして使用されることがあります。これら 2 つのメトリックは密接に関連しています。Gini 係数は ROC 曲線と対角線の間の面積の 2 倍として計算されます (Gini = 2AUC - 1)。Gini 係数は常に 0 と 1 の間であり、数値が大きいほどよい分類子であることを示します。ROC 曲線が対角線より下に位置するという非現実的な状況では、Gini 係数は負になります。

確信式 **(**ある場合**):** 確信度フィールドを生成するモデルの場合に、確信度の値およびその値と予測値の関係に関する統計量が報告されます。この項目には 2 つの設定があります。

- 閾値**:** 精度が指定されたパーセントに達する確信度レベルを報告します。
- 精度の改善**:** 指定した因子によって精度が改善される確信度レベルを報告します。例えば、全体的な精度が 90% で、このオプションを 2.0 に設定した場合、報告される値は、95% の精度を達成するのに必要な確信度です。

次を使用する予測済み/予測フィールドの検出**:** 予測フィールドが元の対象フィールドにどのように一致するかを指定します。

- モデル出力フィールドのメタデータ**:** モデル・フィールド情報に基づいて、予測フィールドを対象に一致させます。予測フィールドの名前が変更されている場合でも一致は可能です。予測フィールドのモデル・フィールド情報は、データ型ノードを使用して、「値」ダイアログ・ボックスからアクセスすることができます。 詳しくは、トピック 151 ページの『「値」ダイアログ・ボックスの使用』を参照してください。
- フィールド名形式**:** 名前の表記方法に基づいて、フィールドのマッチングを行います。例えば、回答という名の対象に C5.0 モデル・ナゲットが生成した予測値が、*$C-response* というフィールド内にある必要があります。

データ区分によって分割**:** レコードを、学習、テスト、および検定用の各サンプルに分割するためにデータ区分フィールドが使用される場合、このオプションを選択すると、各データ区分ごとに、別々の結果が表示されます。 詳しくは、トピック 185 ページの『データ区分ノード』を参照してください。

注 : データ区分を分割する場合、データ区分フィールドにあるヌル値を持つレコードは、分析から除外されます。データ区分ノードは、ヌル値を生成しないため、データ区分ノードを使用している場合は、問題になりません。

ユーザー定義分析**:** 各自のモデル評価に使用する独自の分析計算式を指定できます。各レコードに対して何を計算するのかを CLEM 式を使用して指定し、さらにレコードレベルのスコアを全体的なスコアに組み込む方法を指定します。また、**@TARGET** 関数と **@PREDICTED** 関数を使用して、対象 (実際の出力) 値と予測値をそれぞれ参照します。

- **If:** 条件に応じて異なる計算を使う必要がある場合は、条件式を指定します。
- **Then:** IF 条件が真 (true) の場合に実行する計算式を指定します。
- **Else:** IF 条件が偽 (false) の場合に実行する計算式を指定します。
- **Use:** 個々のスコアから全体的なスコアを算出するための統計量を選択します。

フィールドによる評価対象分析**:** 精度分析の対象として使用できるカテゴリー・フィールドを表示します。全体的な分析に加えて、各対象フィールドのカテゴリーごとに個別に分析を行うこともできます。

精度分析出力ブラウザー

分析出力ブラウザーには、分析ノードの実行結果が表示されます。通常の保存、エクスポート、および印刷操作は、「ファイル」メニューから行うことができます。 詳しくは、トピック 323 ページの『出力を表示』を参照してください。

精度分析の出力を初めて参照する場合、結果は展開されて表示されています。参照し終わった後に結果を隠すには、項目の左側にある拡張をコントロールを使用して目的の結果を省略するか、または 「すべて閉じる」 ボタンをクリックしてすべての結果を非表示にします。閉じた結果をもう一度表示するには、項目の左側にある拡張コントロールを使用するか、または 「すべて展開」 ボタンをクリックしてすべての結果を表示してください。

出力フィールドの結果。精度分析の出力には、各出力フィールドに対するセクションが含まれています。ここには、生成されたモデルが作成した、対応する予測フィールドが格納されています。

比較。出力フィールド・セクション内には、出力フィールドと関連付けられた各予測フィールドに対するサブセクションがあります。カテゴリー出力フィールドの場合、このセクションの最上位レベルにあるテーブ

ルには、正/誤の予測数とパーセント、およびストリーム中の総レコード数が表示されます。数値出力フィールドの場合、このセクションには次の情報が表示されます。

- 最小誤差。最小誤差を表示します (観測値と予測値の差異)。
- 最大誤差。最大誤差を表示します。
- 平均誤差。すべてのレコードの平均誤差を表示します。これは、モデル中に系統バイアス (過小な推定よりも過大な推定を行う傾向がある、またはその逆) があるかどうかを示します。
- 絶対平均誤差。すべてのレコードの誤差の絶対値の平均を表示します。この値は、方向に関係ない絶対値の平均誤差を示しています。
- 標準偏差。誤差の標準偏差を表示します。
- 線型相関。予測値と実際の値の間の線型相関を表示します。この統計量は -1.0 から 1.0 の範囲で変化します。+1.0 に近い値は強い正の相関を表すため、高い予測値は高い実際値に関連し、低い予測値は低い実際値に関連します。-1.0 に近い値は強い負の相関を表すため、高い予測値は低い実際値に関連し、低い予測値は高い実際値に関連します。0.0 に近い値は弱い相関を表し、予測値と実際の値は多かれ少なかれ独立しています。注 : 空白のエントリーは、実際の値または予測値が定数であるため、この場合は線型相関を計算できないことを示します。
- 頻度数。精度分析に使用したレコード数を表示します。

一致行列。カテゴリー出力フィールドに対して、精度分析オプションで一致行列を要求している場合、ここに行列が表示されます。行は実際の観測値を、列は予測値を表します。テーブル中のセルは、それぞれの予測値と実際の値の組み合わせのレコード数を示しています。

パフォーマンス評価。カテゴリー出力フィールドに対して、精度分析オプションでパフォーマンス評価統計量を要求している場合、ここにパフォーマンス評価の結果が表示されます。各出力カテゴリーが、パフォーマンス評価統計量とともに表示されます。

確信度値レポート。カテゴリー出力フィールドに対して、精度分析オプションで確信度値を要求している場合、その値がここに表示されます。モデルの確信度値に対して、次の統計量が報告されます。

- 範囲: ストリーム・データ中のレコードに対する確信度値の集計範囲 (最小値と最大値) を表示します。
- 正解の平均値。正しく分類されたレコードの確信度の平均値を表示します。
- 誤りの平均値。誤って分類されたレコードの確信度の平均値を表示します。
- 常に正解。予測が常に正解で、ケースのパーセンテージがこの基準を満たす確信度のしきい値を表示します。
- 常に不正解。予測が常に不正解で、ケースのパーセンテージがこの基準を満たす確信度のしきい値を表示します。
- 精度 **X%** 以上。精度が *X*% 以上の確信度レベルを表示します。*X* はほぼ、「精度分析」オプションの 「しきい値」 で指定された値になります。一部のモデルやデータ・セットでは、オプションに指定された正しいしきい値を与える確信度値を選択できません (しきい値近くで同じ確信度値を持つ類似ケースのクラスターのためなど)。報告されるしきい値は、単一の確信度値しきい値から取得できる、指定された精度基準に最も近い値になります。
- 分割 **X** 以上。データ・セットの総体的な精度よりも *X* 倍良好な精度の確信度値を表示します。*X* は、「精度分析」オプションの 「精度の改善」 で指定された値になります。

間の一致。ストリーム中に、同じ出力フィールドを予測する 2 つ以上の生成されたモデルがある場合、モデルが生成した予測間の一致に関する統計量も表示されます。これには、予測が一致したレコード数とパーセンテージ (カテゴリー型出力フィールドの場合)、または誤差要約統計量 (連続型出力フィールドの場合)

が含まれます。カテゴリー・フィールドの場合、モデルが同意した (同じ予測値を生成) レコードのサブセットの実際の値と比較した予測の精度分析が含まれます。

評価メトリック。 バイナリ分類子の場合は、分析オプションで評価メトリックを要求すると、AUC および Gini 係数の評価メトリックの値がこのセクションの表に表示されます。表では、バイナリ分類子モデルごとに 1 行で示されます。評価メトリックの表は、モデルごとではなく、出力フィールドごとに表示されます。

データ検査ノード

データ検査ノードは、IBM SPSS Modeler に取り込むデータを広範に検査するための手段を提供し、把握しやすい行列形式で表示され、簡単にソートしたり、グラフやデータ準備ノードを生成することができます。

- 「検査」タブでは要約統計量、ヒストグラム、棒グラフを含むレポートを表示し、データの予備調査を高めるのに有効です。レポートでは、フィールド名の前にストレージ アイコンも表示します。
- 検査レポートの「欠損値検査」タブは、外れ値、極値、および欠損値に関する情報を示し、これらの値を処理するためのツールを提供します。

データ検査ノードの使用

データ検査ノードは、ソース・ノードやインスタンス化されたデータ型ノードの下流に直接接続することができます。また、結果に基づき、多くのデータ準備ノードを生成することもできます。例えば、フィルター・ノードを生成し、モデル作成で有用な欠損値をあまりに多く含むフィールドを除外することができます。また、スーパー・ノードを作成して維持するいずれかまたはすべてのフィールドに対して欠損値を代入することができます。監査の真価はこのような領域で発揮され、データの現在の状態を評価するだけでなく、評価に基づいて対策をとることができます。

データのスクリーニングまたはサンプリング**:** 「ビッグ・データ」を処理する場合は初期の監査が特に効果的であるため、サンプル・ノードを使用し、レコードのサブセットだけを選択することにより、初期の監査にかかる処理時間を短縮することができます。また、データ検査ノードは、分析の調査段階でフィールド選択や異常値検出ノードなどと連携して使用することができます。

データ検査ノードの「設定」タブ

「設定」タブでは、監査用の基本パラメーターを指定することができます。

デフォルト。ノードをストリームに関連付けて「実行」をクリックするだけで、以下のように、デフォルトの設定に基づいてすべてのフィールドの監査レポートを生成することができます。

- データ型ノードの設定がない場合は、レポートにはすべてのフィールドが含まれます。
- インスタンス化されているかどうかにかかわらず、データ型の設定がある場合は、すべての「入力」、「対象」、および「両方」フィールドが含まれます。「対象」 フィールドが 1 つある場合は、それがオーバーレイ フィールドとして使用されます。複数の 「対象」 フィールドが指定されている場合は、デフォルトのオーバーレイは指定されません。

ユーザー設定フィールドを使用。このオプションを使用すると、フィールドを手動で選択できます。右側にあるフィールド選択ボタンを使用して、フィールドを個別に、またはデータ型により選択します。

オーバーレイ・フィールド。オーバーレイ フィールドは、検査レポートに表示されるサムネイル・グラフの描画に使用されます。連続型 (数値範囲型) フィールドの場合、二変数の統計値 (共分散および相関) も計算されます。データ型ノード設定に基づいて対象フィールドがひとつ表示された場合、上記に表示されて

いる通り、デフォルトのオーバーレイ フィールドとして使用されます。代わりに、「ユーザー設定フィールドを使用」 を選択し、オーバーレイを指定することもできます。

表示。グラフが出力可能かどうかを指定したり、デフォルトで表示する統計量を選択したりすることができます。

- グラフ。選択された各フィールドに、データの必要に応じて棒グラフ、ヒストグラム、散布図のいずれかのグラフを表示します。グラフは、最初のレポートではサムネイルで表示されますが、フルサイズのグラフおよびグラフ・ノードも生成されます。 詳しくは、トピック 337 ページの『データ検査出力ブラウザー』を参照してください。
- 基本/詳細統計量。デフォルトで出力表示される統計量のレベルを指定します。この設定で最初の表示を定義しますが、すべての統計がこの設定と関係なく出力できます。 詳しくは、トピック 338 ページの『統計の表示』を参照してください。

中央値と最頻値。レポートの全フィールドに対し、中央値と最頻値を計算します。大容量のデータ・セットでは、他の統計より計算に時間がかかるため、これらの統計により多くの処理時間がかかることがあります。中央値のみの場合、報告された値は、いくつかのケースのうち (データ・セット全体ではなく) 2000 レコードのサンプルに基づいています。このサンプリングは、メモリが拡張されない場合、フィールドあたりの基準で行われます。サンプリングが実行されている場合、結果は出力 (中央値のみではなくサンプル中央値) などでラベル付けされます。中央値以外の統計値はすべて、常に完全なデータ・セットを使用して計算されます。

空またはデータ型不明のフィールド。インスタンス化されたデータとともに使用した場合、データ型不明のフィールドは検査レポートには含まれていません。データ型不明のフィールド (空のフィールド含む) を含むには、データ型ノードのいずれかの上流で 「すべての値の消去」 を選択します。これにより、データはインスタンス化されず、すべてのフィールドがレポートに含まれます。例えば、すべてのフィールドの完全なリストを取得したり、空のフィールドを除外するフィルター・ノードを生成する場合、このオプションが役に立ちます。 詳しくは、トピック 341 ページの『欠損データを含むフィールドのフィルタリング』を参照してください。

データ検査の「欠損値検査」タブ

データ検査ノードの「欠損値検査」タブは、欠損値や外れ値、極値を処理するためのオプションを提供します。

欠損値

- 有効な値のレコードをカウント: 各評価フィールドに対して、有効な値を持つレコード数を表示する場合に選択します。ヌル (未定義の) 値、空白値、ホワイト スペースや空の文字列は、常に無効な値として処理されます。
- 無効な値のレコードの内訳のカウント: 各フィールドに対して、各種の不正な値を持つレコード数を表示する場合に選択します。

外れ値および極値

外れ値と極値を検出する方法として、以下の 2 つの方法がサポートされています。

平均からの標準偏差: 平均からの標準偏差に基づいて、外れ値や極値を検出します。例えば、フィールドの平均 100 で標準偏差が 10 である場合、 3.0 を指定して、70 未満および 130 を超える値が外れ値として扱われているかを表示することができます。

4 分位範囲**:** 4 分位範囲に基づいて外れ値や極値を検出、その範囲は 2 つの中央分位が存在する範囲です (25 ～ 75 分位の間)。例えば、デフォルト設定の 1.5 の場合は、外れ値の下限しきい値が Q1 - 1.5 * IQR になり、上限しきい値が Q3 + 1.5*IQR になります。このオプションを使用すると、大容量のデータ・セットに対するパフォーマンスが遅くなります。

データ検査出力ブラウザー

データ検査ブラウザーは、データの概要を把握するための強力なツールです。「欠損値検査」タブでは外れ値、極値、欠損値についての情報を表示しますが、「検査」タブでは、全フィールドのサムネイル・グラフやストレージ アイコン、統計を表示します。初期グラフおよび要約統計量に基づいて、数値フィールドの記録、新規フィールドの作成、または名義型フィールドの値の再分類などの作業を行うことができます。また、さまざまな機能を使用してデータを視覚化し、より詳細にデータを探索することもできます。「ノードの生成」メニューを使用して、検査レポート・ブラウザーから直接多くのノードを作成し、データの変換や視覚化に使用することができます。

- 列のヘッダをクリックして列をソートしたり、ドラッグ・アンド・ドロップを使用して列の並べ替えることができます。多くの標準出力処理も、サポートされています。 詳しくは、トピック 323 ページの『出力を表示』を参照してください。
- 「尺度」列または「一意」列のフィールドをダブルクリックして、フィールドの値と範囲を参照する。
- ツールバーまたは「編集」メニューをクリックし、値のレベルを表示または非表示にしたり、表示する統計値を選択することができます。 詳しくは、トピック 338 ページの『統計の表示』を参照してください。
- フィールド名の左側にあるストレージ アイコンを確認します。ストレージは、フィールド中へのデータの格納方法を表しています。例えば、1 と 0 の値をとるフィールドは整数データを格納します。これはデータの使用法を記述する測定の尺度とは異なり、ストレージに影響を与えません。詳しくは、トピック 9 ページの『フィールドのストレージと形式の設定』を参照してください。

グラフの表示および生成

オーバーレイ フィールドが選択されていない場合は、「検査」タブで棒グラフ (名義型またはフラグ型) またはヒストグラム (連続型) を表示します。

名義型またはフラグ型フィールドのオーバーレイの場合、グラフはオーバーレイの値で色分けされます。

連続型フィールドのオーバーレイの場合、1 次元の棒グラフやヒストグラムの代わりに、2 次元の散布図が生成されます。この場合、*x* 軸がオーバーレイ フィールドにマップされ、すべての *x* 軸を同じ尺度で参照することができます。

- フラグ型または名義型フィールドの場合、カーソルをバーの上に当てると、ツール ヒントの潜在的な値またはラベルを表示します。
- フラグ型または名義型フィールドの場合、ツールバーを使用して、サムネイル・グラフの位置を水平方向から垂直方向に切り替えることができます。
- フルサイズのグラフをサムネイルから生成するには、サムネイルをダブルクリックするか、サムネイルを選択して「ノードの生成」メニューから 「グラフ出力」 を選択します。注 : サンプリングされたデータに基づくサムネイル・グラフの場合、元のデータ・ストリームが開かれていれば、生成されるグラフにはすべてのケースが含まれます。

  出力を作成したデータ検査ノードはストリームに接続している場合にのみグラフを生成できます。
- 一致するグラフ・ノードを生成するには、「検査」タブで 1 つ以上のフィールドを選択し、「ノードの生成」メニューから 「グラフ作成ノード」 を選択します。結果ノードがストリーム領域に追加され、そのノードを使用して、ストリームが実行されるごとにグラフを再作成することができます。

- オーバーレイ セットに 100 個を超える値がある場合、警告メッセージが表示され、オーバーレイは入れられません。

統計の表示

「統計の表示」ダイアログ・ボックスでは、「監査」タブに表示される統計量を選択することができます。初期設定は、データ検査ノードで指定されています。 詳しくは、トピック 335 ページの『データ検査ノードの「設定」タブ』を参照してください。

*Minimum (*最小値*).* 数値変数の最小値。

*Maximum (*最大*).* 数値変数の最大値。

*Sum (*合計*).* 欠損値でない値を持つすべてのケースにわたる値の和 (合計)。

*Range (OK (*ファイルオープン時のオプション*)).* 数値変数の最大値と最小値の差。最大値から最小値を引いた値。

*Mean (*平均*).* 中心傾向の指標。算術平均 (合計をケース数で割った値) です。

*Standard Error of Mean (*平均値の標準誤差*).* 同じ分布から抽出したサンプルの間で平均値がどの程度異なるかを示す指標。観測した平均と仮説による値をおおまかに比較するために使用することができます (差と標準誤差の比率が -2 より小さいか +2 より大きい場合は、2 つの値が異なっていると結論付けることができます)。

*standard deviation (*標準偏差*).* 平均の周りの散らばりの指標。分散の平方根に等しくなります。標準偏差は元の変数と同じ単位で表します。

*Variance (*分散 *(*信頼性分析*)).* 平均値の周りの値の散らばりの指標。平均値からの偏差の平方和を、ケース数より 1 少ない値で割ったものに等しくなります。分散の測定単位は、変数自体の単位の 2 乗です。

*Skewness (*歪度*).* 分布の非対称性の指標。正規分布は対称であり、歪度の値は 0 です。歪度が正の大きな値である分布は、右側の裾が長くなります。歪度が負で絶対値が大きい分布は、左側の裾が長くなります。目安として、歪度が標準誤差の 2 倍より大きい場合は、対称分布からずれていると解釈します。

*Standard Error of Skewness (*歪度の標準誤差*).* 標準誤差に対する歪度の比率は、正規性の検定として使用することができます (比率が -2 より小さいか +2 より大きい場合は、正規性を棄却することができます)。歪度が大きな正の値である場合は、右側の裾が長いことを示します。極端な負の値の場合は、左側の裾が長いことを示します。

*Kurtosis (*尖度*).* 外れ値が存在する度合いの指標。正規分布の場合、尖度の統計値は 0 です。尖度が正の場合、そのデータの極端な外れ値は正規分布よりも多いことを示します。尖度が負の場合、そのデータの極端な外れ値は正規分布よりも少ないことを示します。

*Standard Error of Kurtosis (*尖度の標準誤差*).* 標準誤差に対する尖度の比率は、正規性の検定として使用することができます (比率が -2 より小さいか +2 より大きい場合は、正規性を棄却することができます)。尖度が大きな正の値である場合は、分布の裾が正規分布の裾より長いことを示します。尖度が負の値である場合は、裾が短いことを示します (箱形の一様分布に似た形になります)。

*Unique (*固有*).* あらゆる種類の他のすべての効果に適合するように各効果を調整することによって、すべての効果を同時に評価します。

Valid (有効). ユーザー欠損として定義された値もシステム欠損値も持たない有効なケース。 ヌル (未定義の) 値、空白値、ホワイト スペースや空の文字列は、常に無効な値として処理されます。

Median (中央値). この値より上と下それぞれにケースの半数ずつが該当することになる値。50 パーセンタイル。ケース数が偶数の場合の中央値は、昇順または降順にソートしたときに中央に来る 2 つのケースの平均です。中央値は、外れ値に対して敏感でない、中心傾向の指標です。それに対して平均値は、少数の極端に大きいまたは小さい値に影響されることがあります。

Mode (最頻値). 最も多く出現する値。複数の値が最高の頻度で出現し、その頻度が同じである場合は、それぞれが最頻値となります。

パフォーマンス改善のために中央値と最頻値はデフォルトで抑制されていますが、データ検査ノードの「設定」タブで選択することができます。 詳しくは、トピック 335 ページの『データ検査ノードの「設定」タブ』を参照してください。

オーバーレイの統計

連続型 (数値範囲型) のオーバーレイ フィールドが使用されている場合、次の統計も利用可能です。

Covariance (共分散). 2 つの変数の間の、標準化されていない関連度。偏差の積和を N-1 で割った値に等しくなります。

データ検査ブラウザーの「欠損値検査」タブ

データ検査ブラウザーの「品質」タブには、データ品質分析の結果が表示されます。このタブでは、外れ値、極値、欠損値の処理を指定することができます。

値の代入**:** 検査レポートは各フィールドの完全なレコードの割合を有効値、ヌル値、空白値の数とともに一覧表示します。必要時にこのオプションを選択して、特定のフィールドに欠損値を代入し、スーパー・ノードを生成してこれらの変換に適用することができます。

1. 「欠損値の代入」 列では、欠損値が存在した場合に代入する値の種類を指定します。このオプションを選択して、空白値、ヌル値または両方を代入し、代入する値を選択するユーザー設定条件や式を指定します。

   IBM SPSS Modeler の欠損値には、次の 2 種類があります。

   - ヌル値またはシステム欠損値 **:** これらの値は、データベースまたはソース・ファイルに空白のまま残された文字列以外の値であり、入力ノードまたはデータ型ノードで特に「欠損値」として定義されていません。システム欠損値は **$null$**値として表示されます。空の文字列が特定のデータベースでヌルとして処理される場合でも、IBM SPSS Modeler では空の文字列をヌルとは見なさないことに注意してください。
   - 空文字列と空白文字 **:** 空文字の値と空白文字（表示されない文字による文字列）をヌル値の重複レコードとして処理します。空白の文字列は、ほとんどの目的に対してホワイト スペースとして扱われます。例えば、オプションを選択してソースまたはデータ型ノードで空白文字を空白値として扱う場合、この設定は空白の文字列も同様に適用します。
   - 空白値またはユーザー定義の欠損値**:** これらは、ソース・ノードまたはデータ型ノード内で欠損値として明示的に定義されている `unknown`、`99`、`–1` などの値です。オプションでヌルと空白文字を「空白」として処理することもできます。そうすることによって、特別な処理のためにフラグを付けたり、ほとんどの計算から除外することができるようになります。例えば、`@BLANK` 関数を使用して、これらの値を他の欠損値と共に空白値として処理することができます。
2. 「方法」 の列では、使用する方法を指定します。

次の方法で、欠損値の代入ができます。

固定**:** 固定値で置き換えます (指定のフィールド計測、範囲の中間または一定数)。

無作為 **:** 正常または均一分布に基づいたランダム値で置き換えます。

式 **:** ユーザー設定の式を指定することができます。例えば、値をグローバル値設定ノードで作成されたグローバル変数と置き換えることができます。

アルゴリズム: C&RT アルゴリズムの基づいたモデルによって予測された値で置き換えます。この方法で代入された各フィールドに対し、空白値やヌル値をモデルで予測された値と置き換える置換ノードとともに、個別の C&RT モデルが作成されます。フィルター・ノードを使用して、モデルが生成した予測値を削除します。

3. 欠損値スーパー・ノードを生成するには、メニューから次の通り選択します。

   「生成」 > 「欠損値スーパーノード」

   「欠損値スーパーノード」ダイアログ・ボックスが表示されます。
4. 「すべてのフィールド」 または 「選択されたフィールドのみ」 を選択し、必要があれば標本サイズを指定します。(指定のサンプルは割合で、デフォルトで全レコードの 10% がサンプルとなります。)
5. 「**OK**」 をクリックして、生成されたスーパーノードをストリーム領域に追加します。
6. スーパーノードをストリームに接続させ、変換を適用させます。

スーパーノード内では、モデル・ナゲット、置換、および置換ノードの組み合わせが必要に応じて使用されます。スーパーノードを編集して 「ズーム・イン」 をクリックして、どのように動作するか確認することができます。また、スーパーノード内の特定ノードを追加、編集、削除して、動作を調整することができます。

外れ値および極値の処理**:** 検査レポートは多くの外れ値および極値を一覧表示し、データ検査ノードで指定された「検出」オプションに基づいて、各フィールドに一覧表示されます。詳しくは、トピック 336 ページの『データ検査の「欠損値検査」タブ』を参照してください。 必要時にこのオプションを選択して、特定のフィールドにこれらの値を強制、削除、無効化し、スーパーノードを生成してこれらの変換に適用することができます。

1. 「アクション」 列では、必要に応じて特定のフィールドに対する外れ値および極値の処理を指定します。

   外れ値および極値の処理には、次のアクションが有効です。

   - 強制**:** 外れ値および極値を、極端とはみなされない直近の値と置換します。例えば、外れ値が標準偏差 3 を上回るまたは下回ると定義されている場合、外れ値はこの領域内の最大値または最低値と置き換えられます。
   - 破棄**:** 指定されたフィールドに対し、範囲外の値または極値を含むレコードを破棄します。
   - 無効化**:** 外れ値および極値を、ヌル値またはシステム欠損値と置き換えます。
   - 外れ値の強制/極値の廃棄**:** 極値のみを破棄します。
   - 外れ値の強制/極値の無効化**:** 極値のみを無効にします。
2. スーパーノードを生成するには、メニューから次の通り選択します。

   「生成」 > 「外れ値および極値スーパーノード」

   「外れ値スーパーノード」ダイアログ・ボックスが表示されます。

3. 「すべてのフィールド」 または 「選択されたフィールドのみ」 を選択し、「**OK**」 をクリックして、生成されたスーパー・ノードをストリーム領域に追加します。
4. スーパーノードをストリームに接続させ、変換を適用させます。

必要に応じて、スーパー・ノードを編集し、ズーム・インして表示または変更することができます。スーパーノードでは、必要に応じて、一連の条件抽出ノードおよび置換ノードを使用して、値を破棄、強制または無効化します。

欠損データを含むフィールドのフィルタリング**:** データ検査ブラウザーで、「品質からフィルターを生成」ダイアログ・ボックスを使用することにより、品質分析の結果に基づいて新しいフィルター・ノードを作成することができます。

モード**:** 指定したフィールドに対する操作として、「含める」 または 「除外」 を選択します。

- 選択したフィールド**:** フィルター・ノードは、欠損値検査テーブルで選択されたフィールドを含めるか、または除外します。例えば、「非欠損値の割合**(%)**」 でテーブルをソートし、Shift キーを押したままクリックして、最小の完了フィールドを選択し、これらのフィールドを除外するフィルター・ノードを生成することができます。
- 品質パーセンテージが **%** よりも高いフィールド**:** フィルター・ノードは、完全なレコードの割合が指定したしきい値よりも大きい場合に、フィールドを含めるか、または除外します。デフォルトのしきい値は 50 % です。

空のまたはデータ型不明フィールドのフィルタリング

データ値がインスタンス化された後、データ型不明または空のフィールドは検査結果および IBM SPSS Modeler の多くのその他の出力結果から除外されます。これらのフィールドは、モデル作成のためには無視されますが、データを拡大または拡散させる場合があります。その場合、データ検査ブラウザーを使用して、フィルター・ノードを生成し、ストリームからこれらのノードを削除することができます。

1. すべてのフィールド (空のフィールドやデータ型不明のフィールドを含む) を監査の対象とするには、上流のソース・ノードまたはデータ型ノードで「すべての値の消去」を選択するか、すべてのフィールドの値を「<パス>」に設定します。
2. データ検査ブラウザーでは、「非欠損値の割合**(%)**」 列でソートし、ゼロの有効な値 (またはその他のしきい値) を含むフィールドを選択し、「ノード生成」メニューでストリームに追加することのできるフィルター・ノードを生成します。

欠損データを含むレコードの選択**:** データ検査ブラウザーから、欠損値検査の結果に基づいて新しい条件抽出ノードを生成することができます。

1. データ検査ブラウザーで、「品質」タブを選択します。
2. メニューから次の項目を選択します。

   「生成」 > 「欠損値選択ノード」

   「条件抽出ノードの生成」ダイアログ・ボックスが表示されます。

レコードが次の状態の時に選択**:** レコードが 「有効」 または 「無効」 の場合に保持することを指定します。

無効な値の検索場所**:** 無効な値の検索場所を指定します。

- すべてのフィールド**:** 条件抽出ノードは、すべてのフィールドに対して無効な値があるかどうかを検査します。

- テーブルで選択したフィールド: 条件抽出ノードは、現在欠損値検査出力テーブルで選択されているフィールドだけを検査します。
- 品質パーセンテージが % よりも高いフィールド: 完全なレコードの割合が、指定したしきい値よりも大きい場合に、条件抽出ノードはフィールドを検査します。デフォルトのしきい値は 50 % です。

次の場所に無効な値が見つかった場合に、レコードを無効と見なす : レコードを無効と見なす条件を指定します。

- 上記の任意のフィールド: 上で指定したいずれか のフィールドにレコードに対して無効な値が含まれている場合、条件抽出ノードはそのレコードを無効なレコードと見なします。
- 上記のすべてのフィールド: 上で指定したすべて のフィールドにレコードに対して無効な値が含まれている場合のみ、条件抽出ノードはそのレコードを無効なレコードと見なします。

データ準備用のその他のノードの生成

データ分類ノード、データ分割ノード、フィールド作成ノードなど、データの準備に使用するさまざまなノードを、データ検査ブラウザーから直接生成することができます。以下に例を示します。

- *claimvalue* と *farmincome* の値に基づいて新しいフィールドを作成するには、検査レポートからこれらのフィールドを選択し、「生成」メニューの 「フィールド作成」 を選択します。ストリーム領域に新しいノードが追加されます。
- 同様に、検査結果に基づいてより詳細な分析を行うために、パーセンタイルに基づいたビンに *farmincome* を記録することもできます。データ分割ノードを生成するには、表示されているフィールド行を選択した後、「生成」メニューの 「データ分割」 を選択します。

ノードが生成されて、ストリーム領域に追加されたら、そのノードをストリームに接続した後、選択したフィールドに関するオプションを指定する必要があります。

変換ノード

入力フィールドの正規化は、回帰、ロジスティック回帰、判別分析など、既存のスコアリング時技術を使用する前の重要なステップです。これらの技術は、多くの生データ・ファイルにとっては真 (true) ではないデータの正常な分散に関する仮説を実行します。実際のデータの処理に対する 1 つのアプローチは、より正常な分散へ生データの要素を移動させる変換を適用することです。また、正規化フィールドは、用意にお互いを比較することができます。例えば、収入および年齢は生データ・ファイルではまったく異なるスケールですが、正規化されるとそれぞれの関連する影響を、容易に解釈することができます。

変換ノードを使用すると、出力ビューアーで、最高の変換に対する評価を高速かつ視覚的に実行することができます。変数が正常に分散しているか確認したり、必要に応じて適用する変換を選択することができます。複数のフィールドを選択し、フィールドごとに 1 つの変換を実行することができます。

フィールドに対し優先的な変換を選択した後、フィールド作成ノードまたは置換ノードを生成して、変換を実行し、これらのノードをストリームを接続することができます。置換ノードは既存のフィールドを変換しますが、フィールド作成ノードを使用すると新規フィールドを作成できます。 詳しくは、トピック 345 ページの『グラフの生成』を参照してください。

変換ノードの「フィールド」タブ

「フィールド」タブでは、可能な変換を表示し適用させるために使用するデータのフィールドを指定します。数値的フィールドのみが変換できます。フィールド選択ボタンをクリックし、表示されたリストから 1 つ以上の数値フィールドを選択します。

変換ノードの「オプション」タブ

「オプション」タブでは、含める変換の種類を指定することができます。このタブを選択すると、利用可能な変換をすべて含むことができ、または個別に変換を選択することができます。

個別に選択する場合は、逆変換またはログ変換のためにデータをオフセットする数を入力することもできます。この処理は、ゼロを多く含むデータの大部分が、平均値や標準偏差にバイアスがある状況において役に立ちます。

例えば、ゼロの値をいくつか含む 「*BALANCE*」 という名前のフィールドがあり、逆変換を実行したいと仮定します。不要なバイアスを避けるには、逆 **(1/x)** を選択し、「データ・オフセットの使用」 フィールドに 1 を入力します。(このオフセットは、IBM SPSS Modeler の `@OFFSET` シーケンス機能が実行するものとは関連しません。)

すべての公式**:** 利用可能な変換が計算され出力結果に表示されることを指示します。

数式の選択**:** 計算して出力に表示する各種の変換を選択することができます。

- 逆 **(1/x):** 逆変換が出力結果に表示されることを指示します。
- 対数 **(log n):** 対数 $_{n}$ 変換が出力結果に表示されることを指示します。
- 対数 **(log 10):** 対数 $_{10}$ 変換が出力結果に表示されることを指示します。
- 指数**:** 指数変換 (e^{x}) が出力結果に表示されることを指示します。
- 平方根**:** 平方根変換が出力結果に表示されることを指示します。

変換ノードの「出力」タブ

「出力」タブを使用すると、出力形式と出力位置を指定することができます。結果を画面に表示、または結果を標準ファイル形式の 1 つに送ることもできます。詳しくは、トピック 326 ページの『出力ノードの「出力」タブ』を参照してください。

変換ノードの出力ビューアー

出力ビューアーを使用すると、変換ノード実行の結果を表示することができます。ビューアーは、フィールドあたり複数の変換をサムネイルで表示し、フィールドをすばやく比較することができます。「ファイル」メニューのこのオプションを使用して、出力結果を保存、エクスポート、印刷することができます。 詳しくは、トピック 323 ページの『出力を表示』を参照してください。

(選択した変換以外の) 各変換に対し、判例が形式の下に表示されます。

平均 (標準偏差)

変換ノードの作成

出力ビューアーから、データの準備を開始することができます。例えば、「年齢」 のフィールドを正規化して、正常な分布を仮定するロジスティック回帰または判別分析などのスコアリング技術を使用したい場合があります。初期のグラフおよび要約統計に基づき、特定の分布に従って (.log など) 「年齢」 フィールドを変換することを決定します。優先的な分布を選択した後、スコアリングに使用する標準化された変換によって、フィールド作成ノードを生成します。

出力ビューアーから、次のフィールド操作ノードを生成することができます。

- フィールド作成
- 置換

置換ノードは既存のフィールドを変換しますが、フィールド作成ノードを使用すると、希望の変換によって新しいフィールドを作成することができます。ノードはスーパー・ノードの形式で領域に設置されます。

異なるフィールドで同じ変換を選択した場合、フィールド作成ノードまたは置換ノードには、変換が適用されるすべてのフィールドの変換タイプに関する式が含まれます。例えば、次の表に示すフィールドと変換を選択してフィールド作成ノードを生成したとします。

表 *44.* フィールド作成ノードの生成例：

フィールド	変換
AGE	現在の分布
INCOME	Log
OPEN_BAL	逆
BALANCE	逆

次のノードは、スーパーノードに含まれています。

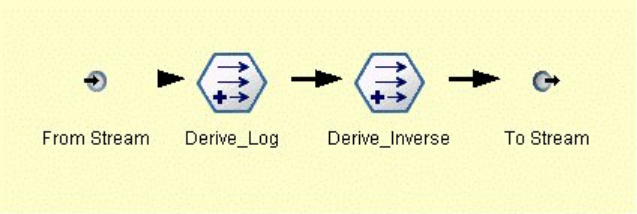


図 *74.* 領域内のスーパーノード

この例では、Derive_Log ノードには 「*INCOME*」 フィールドの対数式、Derive_Inverse ノードには *OPEN_BAL* および 「*BALANCE*」 フィールドに逆数式が含まれています。

ノードの生成方法

1. 出力ビューアーの各フィールドに対し、必要な変換を選択します。
2. 「ノード生成」メニューから、必要に応じてフィールド作成ノードまたは置換ノードを選択します。

   選択すると、「フィールド作成ノードの生成」または「置換ノードの生成」ダイアログ・ボックスが必要に応じて表示されます。

   「標準化されていない変換」 または 「標準化された変換 **(z** スコア**)**」 を必要に応じて選択します。「標準化された変換」オプションは、変換に*z* スコアを適用します。*z* スコアは、標準偏差の変数の平均からの距離を表す機能として、値を表示します。例えば、対数変換を 「年齢」 フィールドに適用し、標準化された変換を選択した場合、生成されたノードの最終式は次のようになります。

   ```
   (log(AGE)-Mean)/SD
   ```

ノードが生成され、ストリーム領域に表示されると、

1. ストリームに接続されます。
2. スーパーノードの場合、必要に応じてノードをダブルクリックし、コンテンツを表示させます。
3. 必要に応じてフィールド作成ノードまたは置換ノードをダブルクリックし、選択したフィールドのオプションを変更します。

グラフの生成**:** 出力ビューアーで、サムネイル ヒストグラムから、フルサイズのヒストグラム出力を生成することができます。

グラフの生成方法

1. 出力ビューアー中のサムネイル・グラフをダブルクリックします。

   または

   出力ビューアー中のサムネイル・グラフを選択します。
2. 「生成」メニューの 「グラフ出力」 を選択します。

正規分布曲線を含むヒストグラムが表示されます。ヒストグラムの表示により、利用可能な各変換がどれほど近く、正規分布に一致するか比較することができます。

注 : 出力を作成した変換ノードはストリームに接続している場合にのみグラフを生成できます。

その他の操作*:* 出力ビューアーから、次のことができます。

- フィールド列で出力グリッドのソート
- 出力結果を HTML ファイルにエクスポート。 詳しくは、トピック 325 ページの『出力のエクスポート』を参照してください。

記述統計ノード

記述統計ノードでは、数値型フィールドに関する基本的な集計情報を得ることができます。このノードから、個々のフィールドの要約統計量とフィールド間の相関についての情報を取得できます。

記述統計ノードの「設定」タブ

検証**:** 要約統計量を個別に算出するフィールドを選択します。複数のフィールドを選択できます。

統計: 報告する統計量を選択します。利用できるオプションには、「カウント」、「平均値」、「合計」、「最小」、「最大」、「集計範囲」、「分散」、「標準偏差」、「平均の標準誤差」、「中央値」、または「最頻値」 があります。

相関関係**:** 相関させるフィールドを選択します。複数のフィールドを選択できます。相関フィールドを選択すると、出力に各「検証」フィールドと相関フィールド間の相関が表示されます。

相関の設定**:** 出力における相関の強さを表示するためのオプションを指定することができます。

相関の設定

IBM SPSS Modeler では、重要な関係を強調するために、相関を詳細ラベルで特徴づけることができます。相関関係は、2 つの連続型 (数値範囲) フィールド間の相関の強さを測定します。値の範囲は -1.0 から 1.0 までです。+1.0 に近い値は強い正の相関を表し、あるフィールドの大きい値が別のフィールドの大きい値と関連付けられ、小さい値が別の小さい値と関連付けられます。-1.0 に近い値は強い負の相関を示し、あるフィールドの大きい値が別のフィールドの小さい値と、そして小さい値が別の大きい値と関連付けられます。0.0 に近い値は弱い相関を示し、2 つのフィールドの値は、独立しているといえます。

「相関の設定」ダイアログ・ボックスを使用すると、相関ラベルの表示の制御、カテゴリーを定義するしきい値の変更、各範囲に使用するラベルの変更を行うことができます。相関値を特徴付ける方法は、問題のドメインに大きく依存しているため、状況に応じて範囲とラベルをカスタマイズすることができます。

出力に相関強度ラベルを表示**:** デフォルトでは、このオプションが選択されます。出力に詳細ラベルを表示しない場合は、このオプションの選択を解除してください。

相関強度**:** 相関強度の定義とラベル付けには次の 2 つのオプションがあります。

- 重要度 **(1-p)** による相関強度の定義**:** 重要度を基に相関にラベル付けをします。 1 マイナス有意、または平均値の差が確率のみで説明できる 1 マイナス確率として定義されます。この値が 1 に近いほど、2 つのフィールドが独立していない、つまりこれらの間になんらかの関係が存在する可能性が高くなります。重要度を基に相関にラベル付けは、データにおけるバラつきを考慮に入れて通常絶対値をお勧めしています。例えば、定数 0.6 は、1 つのデータ・セットでは高い有意性を示しますが、その他のすべてにおいては有意性はありません。デフォルトでは、0.0〜0.9 (絶対値) の重要度の値は 「低い」、0.9〜0.95 は 「中間」、0.95〜1.0 は 「高い」 のラベルが付けられます。
- 絶対値による相関強度の定義**:** Person の相関係数の絶対値を基に相関にラベルを付けます。前述したように、この範囲は -1 から 1 までです。この測定値の絶対値が 1 に近ければ近いほど相関が強くなります。デフォルトでは、0.0〜0.3333 (絶対値) の相関は 「弱い」、0.3333〜0.6666 の相関は 「中間」、0.6666〜1.0 の相関は 「強い」 のラベルが付けられます。ただし、指定された値の有意性をデータ・セットからその他に一般化するのは困難です。そのためほとんどのケースにおいて、絶対値ではなく確率に基づいて相関を定義することをお勧めしています。

記述統計量出力ブラウザー

記述統計ノード出力ブラウザーには、統計分析の結果が表示されます。このブラウザーでは、フィールドの選択、選択内容に基づく新しいノードの生成、結果の保存や印刷などの操作を実行することができます。通常の保存、エクスポート、および印刷関連オプションは「ファイル」メニューから、通常の編集オプションは「編集」メニューから利用できます。 詳しくは、トピック 323 ページの『出力を表示』を参照してください。

記述統計量の出力を初めて参照する場合、結果は展開されて表示されています。参照し終わった後に結果を隠すには、項目の左側にある拡張をコントロールを使用して目的の結果を省略するか、または 「すべて閉じる」 ボタンをクリックしてすべての結果を非表示にします。閉じた結果をもう一度表示するには、項目の左側にある拡張コントロールを使用するか、または 「すべて展開」 ボタンをクリックしてすべての結果を表示してください。

出力には、各 「検証」 フィールド用のセクションが含まれています。このセクションには、要求された統計量がテーブルで表示されています。

- カウント**:** 有効なフィールド値を持つレコード数。
- 平均値**:** すべてのレコードに渡るフィールドの平均値。
- 合計**:**すべてのレコードに渡るフィールドの合計値。
- 最小値**:** フィールドの最小値。
- 最大値**:** フィールドの最大値。
- 範囲: 最小値と最大値の差異。
- 分散**:**フィールドの値における変動の測定値。この値は、各値と総合平均間の差異を二乗して、すべての値を合計したら、それをレコード数で除算して算出します。
- 標準偏差**:**フィールドの値中の変動の測定値で、分散の平方根として算出されます。
- 平均の標準誤差**:**平均が新しいデータに適用されることを前提にしている場合の、フィールドの平均の推定中の不確実性の測定値です。
- 中央値: フィールドの中央値、つまりデータの上半分と下半分を分割する場所にある値です (フィールドの値を基準にして)。

- モード: データ中に最も多く出現する単一値です。

相関係数:相関関係フィールドを指定した場合、出力にも「検証」フィールドと各相関関係フィールド間の Pearson の積率相関係数、および相関関係値の詳細ラベル (オプション) を記載するセクションが含まれます。 詳しくは、トピック 345 ページの『相関の設定』を参照してください。

「ノードの生成」メニュー : 大部分のモデル・ナゲットには、「ノードの生成」メニューもあります。「ノード生成」メニューには、ノード生成操作に関するメニュー項目があります。

- フィルター:他のフィールドとの相関がないフィールドまたは相関が弱いフィールドを除外するフィルター・ノードを生成します。

統計量からのフィルター・ノードの生成

記述統計量出力ブラウザーから生成されたフィルター・ノードでは、他のフィールドとの相関関係に基づいてフィールドがフィルタリングされます。このノードは、絶対値の順序に従って相関関係をソートし、(「統計からフィルターを生成」ダイアログ・ボックスで設定した基準に従って) 最大の相関関係を抽出し、それらの大きな相関関係に現れるすべてのフィールドを通過させるフィルターを作成します。

モード: 相関関係の選択方法を決定します。「含める」 を選択すると、指定した相関関係に現れるフィールドが保持されます。「除外」 を選択すると、フィールドがフィルタリングされます。

次の所に表示されるフィールドを含める/除外する: 相関関係を選択する基準を定義します。

- 最上位数の相関関係: 指定数の相関関係を選択し、それらの相関関係のいずれかに現れるフィールドを含めるか、または除外します。
- 最上位パーセントの相関関係 (%): 指定したパーセンテージ (*n*%) の相関関係を選択し、それらの相関関係のいずれかに現れるフィールドを含めるか、または除外します。
- 次の値よりも大きい相関: 指定されたしきい値よりも絶対値が大きい相関関係を選択します。

平均比較ノード

平均比較ノードでは、独立したグループ間で、または関連するフィールドのペア間で著しい違いがあるかどうかを調べるために、平均を比較します。例えば、販売促進活動の実施前後に平均収入を比較、または販売促進活動を受け容れない顧客からの売上げの比較ができます。

2 つの異なる方法で、以下のデータを基に平均値を比較できます。

- フィールド内のグループ間で比較: 独立グループを比較するには、テストフィールドとグループ化フィールドを選択します。例えば、販売促進活動の対象から「ホールド・アウト」顧客のサンプルを除外でき、ホールド・アウトグループとその他のすべてと平均収入を比較できます。この場合、各顧客の収入を示す単一のテスト・フィールドを、フラグまたはかれらがオファーを受けるかどうかを示す名義型フィールドで指定します。サンプルは各レコードが 1 つのグループ他に割り当てられるとこと、また 1 つのグループの特定のメンバーをその他のグループの特定のメンバーに結びつける方法がない、という意味で独立しています。また、2 つ以上の値で名義型フィールドも指定し、複数のグループの平均値を比較できます。実行時には、ノードは選択したフィールドで一元 ANOVA テストを計算します。2 つのフィールド グループしかない場合、一元 ANOVA の結果は本質的に独立したサンプル *t* テストと同じです。 詳しくは、トピック 348 ページの『独立したグループの平均値を比較』を参照してください。
- フィールドのペア間で比較: 2 つの関連フィールドの平均値を比較する場合、有意の結果を得るために、そのグループ群は何らかの方法でペアを組みます。例えば、販売促進活動の実施前後の同じグループの顧客からの平均収入を比較するか、または夫婦ペア間サービスの利用率を比較して差が見られるか

どうか確認します。各レコードには、独立しているが関連する有意性を比較できる 2 つの方法を含みます。実行時に、ノードは一組のサンプル *t* テストを指定された各フィールドペアで計算します。 詳しくは、トピック『一対のフィールド間の平均値の比較』を参照してください。

独立したグループの平均値を比較

平均比較ノードで 「フィールド内のグループ間で比較」 を選択し、2 つ以上の独立グループを比較します。

グループ化フィールド**:** 比較したいグループにレコードを区切った、例えばオファーを受け取る/受け取らない人などのように区切った、 2 つ以上の重複レコード値を持つ数値フラグまたは名義型フィールドを選択します。試験フィールド数に無関係に、ただひとつグループ化するフィールドを選択できます。

テスト・フィールド**:** 試験の対象となる測定値を含む 1 以上の数値型フィールドを選択します。独立した試験を選択した各フィールドに実施します。例えば、やり方、資金、運動の盛り上げに関する特定の販売促進活動の影響をテストすることができます。

一対のフィールド間の平均値の比較

平均値ノードで 「フィールドのペア間で比較」 を選択し、独立したフィールド間の平均値を比較します。何らかの方法で有意となるようにフィールド同士を関連付ける必要があります (例えば、販売促進活動の前後の収入など)。複数のフィールド ペアを選択することもできます。

フィールド **1:** 比較したい最初の測定値を含む数値フィールドを選択します。前後の検討において、これは「前」のフィールドにあたります。

フィールド **2:** 比較したい 2 番目のフィールドを選択します。

追加**:** 選択したペアをテスト型フィールド ペア リストに追加します。

フィールドの選択を繰り返し、必要な複数のペアをリストに追加します。

相関の設定**:** 相関の強さにラベルを付けるためのオプションを指定することができます。詳しくは、トピック 345 ページの『相関の設定』を参照してください。

平均比較ノードのオプション

「オプション」タブを使用すると、しきい値 *p* を設定し、このしきい値を使用して「重要」、「境界」、「重要ではない」というラベルを結果に付けることができます。また、各ランク付けラベルの編集もできます。重要度 高は、パーセンテージで測定され、大雑把に 1 マイナス結果を取得する確率 (2 つのフィールドの平均値の違いなど) と定義され、観察された確率のみによる結果とほぼ同程度、またはもっと極端と定義します。例えば、0.95 を超える *p* 値は、確率のみで結果を説明できるとすると、5 % 未満の確率であることを示します。

重要度のラベル**:** 出力時に各フィールドペアまたはグループにラベル付けするラベルの編集ができます。デフォルトのラベルは、「重要度 高」、「境界」、「重要度 低」 です。

分割値**:** 各ランクのしきい値を指定します。一般的に、0.95 を超える *p* 値は、重要度 高とランク付けられ、0.9 未満は、重要度 低とランク付けされます。しかしこれらのしきい値を必要に応じて調整することができます。

注：重要度の測定はさまざまなノードで利用できます。具体的な演算、使用する対象フィールドおよび入力フィールドのデータ型はノードによって異なりますが、すべてがパーセンテージによる測定であるため、値は比較することができます。

平均値ノード出力ブラウザー

平均値出力ブラウザーには、クロス集計データが表示されます。このブラウザーでは、テーブルを 1 行ずつ選択 (またはコピー) したり、任意の列でソートしたり、テーブルの保存や印刷を行ったりするなど、標準的な操作を実行することができます。詳しくは、トピック 323 ページの『出力を表示』を参照してください。

表の具体的な情報は、比較のタイプ (フィールド内のグループまたは独立したフィールド) によって異なります。

ソート項目**:** 出力を特定の列でソートすることができます。上下の矢印をクリックしてソートの方向を変更します。または、任意の列見出しをクリックして、列ごとのソートが可能です。(列内のソートの方向を変更するには、もう一度クリックします。)

表示**:** 「シンプル」または「詳細」を選択して、表示の詳細度を制御することができます。詳細表示には、シンプル表示からのすべての情報が含まれますが、その他に詳細な内容が提供されます。

フィールド内でグループを比較する平均値出力

フィールド内のグループを比較する場合、グループ化するフィールド名が出力テーブルの上の方に表示され、平均値や関連する統計値が各グループ別に別々に表示されます。テーブルには、各テストフィールド用に独立した行が含まれています。

また、次の列も表示されます。

- フィールド**:** 選択したテスト・フィールドの名前が表示されます。
- グループ別の平均値**:** グループ化するフィールドの各カテゴリーの平均値が表示されます。例えば、特別なオファーを受けた人 (新規販売促進活動対象) と受けなかった人 (標準) を比較できます。詳細表示で、標準偏差、標準誤差、度数も表示できます。
- 重要度**:** 重要度の値とラベルを表示します。 詳しくは、トピック 348 ページの『平均比較ノードのオプション』を参照してください。

詳細出力

詳細表示では、次の列が追加表示されます。

- **F-**検定**:** このテストは、グループ間の分散と各グループ内の分散の比率を基にします。平均値がすべてのグループで同じとすると、両方とも同じ母集団分散の予測であるため、*F* 率がほぼ 1 になると期待されます。この比率が大きくなると、グループ間の分散が大きくなり、有意差が存在する可能性がより大きくなります。
- **df.** 自由度を表示します。

フィールドのペアを比較する平均値出力

独立したフィールドを比較する場合、出力テーブルには選択したフィールド ペアの行が含まれます。

- フィールド **1/2:** 各ペアの 1 番目と 2 番目のフィールド名を表示します。詳細表示で、標準偏差、標準誤差、度数も表示できます。
- 平均値 **1/2:** 各フィールドの平均値を個別に表示します。

- 相関**:** 相関強度は、2 つの連続型 (数値範囲) フィールド間の関係の強さを測定します。+1.0 に値が近いほど正の関連が強いことを示し、-1.0 に近いほど負の関係が強いことを示します。 詳しくは、トピック 345 ページの『相関の設定』を参照してください。
- 平均差**:** 2 つのフィールドの平均値の差を表示します。
- 重要度**:** 重要度の値とラベルを表示します。 詳しくは、トピック 348 ページの『平均比較ノードのオプション』を参照してください。

詳細出力

詳細出力には次の列が追加されます。

95 % の信頼区間**:** 真 (true) の平均値が、この母集団のこのサイズでのすべての可能なサンプルの 95 % が含まれる範囲の上下の境界。

t 検定**:** *t* 統計値は、平均値の差をその標準誤差で除算すると得られます。この統計値の絶対値が大きくなると、平均値が異なる可能性が大きくなります。

df. 統計値の自由度を表示します。

レポート・ノード

レポート・ノードを使用すると、固定テキスト、データ、データから取得された他の式を含むフォーマット済みレポートを作成することができます。レポートの書式は、固定テキストとデータの出力構成を定義するテキスト テンプレートを使用して指定します。テンプレート中の HTML タグを使用し、「出力」タブで設定を指定することにより、カスタムの書式設定を利用することができます。テンプレート中の CLEM 式を使用して、データ値および他の条件出力がレポートに入れられます。

レポート・ノードの代替

レポート・ノードは通常、一定の条件を満たすすべてのレコードなど、ストリームからのレコードやケース出力を表示するために使用されます。この点で、レポート・ノードは構造化の度合いが少ない、テーブル・ノードの代替物だと考えられます。

- データ自体でなくストリーム内で定義されたフィールド情報やその他の情報 (データ型ノード内で指定されたフィールド定義など) をレポートに表示しようとする場合は、代わりにスクリプトを使用できます。
- ストリームに生成されたモデル、テーブル、およびグラフの集合として複数の出力オブジェクトを含むレポートを生成するには、そしてそのレポートがテキスト、HTML、および Microsoft Word/Office を含む複数の形式で出力できる場合は、IBM SPSS Modeler プロジェクトが使用できます。
- スクリプトを使用しないでフィールド名のリストを作成するには、すべてのレコードを廃棄するサンプル・ノードに先導された、テーブル・ノードが使用できます。これにより、行のないテーブルが作成され、エクスポート時に行列が入れ替えられて、1 列のフィールド名のリストが作成されます。(このためには、テーブル・ノードの「出力」タブで、「データの入れ替え」 を選択します。)

レポート・ノードの「テンプレート」タブ

テンプレートの作成**:** レポートの内容を定義するには、レポート・ノードの「テンプレート」タブでテンプレートを作成します。テンプレートはテキスト行から構成されており、各行にレポートの内容を指定します。内容行の範囲を示す特殊なタグ行も含まれています。各内容行では、その行をレポートに送る前に、大カッコ ([]) で囲まれている CLEM 式が評価されます。テンプレートでは、1 行に付き次の 3 種類の範囲を指定できます。

固定: マークが付けられていない行は固定と見なされます。固定行は、行に含まれる他の式を評価した後に、レポートに 1 行だけ書き込まれます。例えば、次のような行があるとします。

```
This is my report, printed on [@TODAY]
```

この場合、上記の文字列と現在の日付を記載した 1 行がレポートに書き込まれます。

グローバル **(ALL** を反復**):** 入力データのレコードごとに 1 回ずつ、特殊タグ #ALL と # で囲まれた行がレポートにコピーされます。CLEM 式 (かっこで囲まれている) が、各出力行の現在のレコードに基づいて評価されます。例えば、次のような行があるとします。

```
#ALL
For record [@INDEX], the value of AGE is [AGE]
#
```

この場合、各レコードに対して、レコード番号と年齢を示す 1 行が記載されます。

すべてのレコードのリストを生成するには

```
#ALL
[Age] [Sex] [Cholesterol]  [BP]
#
```

条件付き **(WHERE** を反復**):** 指定された条件が真 (true) であるレコードごとに 1 回ずつ、特殊タグ #WHERE *<condition>* と # で囲まれた行がレポートにコピーされます。条件は CLEM 式で指定します。(WHERE 条件内では、大かっこは省略できます。) 例えば、次のような行があるとします。

```
#WHERE [SEX = 'M']
Male at record no. [@INDEX] has age [AGE].
#
```

この場合、性別が *M* の値を持つ各レコードに対して 1 行がファイルに書き込まれます。完全なレポートには、入力データにテンプレートを適用して定義された固定行、グローバル行、および条件行が含まれます。

「出力」タブから、結果の表示や保存に関する設定を指定することができます。この設定は、さまざまな種類の出力ノードに共通しています。 詳しくは、トピック 326 ページの『出力ノードの「出力」タブ』を参照してください。

HTML または XML 形式データを出力する

両方の形式でレポートを書くために、HTML または XML タグを直接テンプレートに含めることができます。例えば、次のテンプレートは HTML タブを生成します。

```
This report is written in HTML.
Only records where Age is above 60 are included.

<HTML>
  <TABLE border="2">
   <TR>
     <TD>Age</TD>
     <TD>BP</TD>
     <TD>Cholesterol</TD>
     <TD>Drug</TD>
   </TR>

  #WHERE Age > 60
    <TR>
       <TD>[Age]</TD>
```

```
      <TD>[BP]</TD>
      <TD>[Cholesterol]</TD>
      <TD>[Drug]</TD>
    </TR>
  #
  </TABLE>
</HTML>
```

レポート・ノード出力ブラウザー

レポート・ブラウザーには、生成されたレポートの内容が表示されます。通常の保存、エクスポート、および印刷関連オプションは「ファイル」メニューから、通常の編集オプションは「編集」メニューから利用できます。 詳しくは、トピック 323 ページの『出力を表示』を参照してください。

グローバル・ノード

グローバル・ノードで、データを走査し、CLEM 式で使用できる要約値を算出します。例えば、グローバル値の設定ノードを使用して *age* というフィールドの統計量を計算してから、CLEM 式に `@GLOBAL_MEAN(age)` 関数を挿入して *age* 全体の平均を使用することができます。

グローバル・ノードの「設定」タブ

作成するグローバル値**:** グローバル値を使用可能にするフィールドを選択します。複数のフィールドを選択できます。各フィールドに対して、目的の統計量がフィールド名の隣の列で選択されていることを確認し、計算する統計量を指定します。

- 平均**:** すべてのレコードに渡るフィールドの平均値。
- 合計**:** すべてのレコードに渡るフィールドの合計値。
- 最小**:** フィールドの最小値。
- 最大**:** フィールドの最大値。
- 標準偏差**:** 標準偏差フィールドの値中の変動の測定値で、分散の平方根として算出されます。

デフォルトの処理**:** ここで選択したオプションが、上のグローバル値リストに新規フィールドを追加する際に使用されます。デフォルトの統計量セットを変更するには、必要に応じて統計量を選択または解除してください。「適用」 ボタンを使用して、リスト中のすべてのフィールドにデフォルトの処理を適用することもできます。

注**:** いくつかの操作は、数値型以外のフィールドに適用できません (例えば、日付/時刻フィールドには「合計」を適用できません)。選択したフィールドで使用できない操作は無効になります。

実行前にすべてのグローバル値を消去**:** 新しい値を計算する前にすべてのグローバル値を取り除く場合に、このオプションを選択します。このオプションを選択しない場合、古い値が新しく計算された値に置き換えられますが、再計算されなかったグローバル値はそのまま使用できます。

実行後に作成されたグローバル値のプレビューを表示**:** このオプションを選択した場合、実行後に「ストリームのプロパティー」ダイアログ・ボックスの「グローバル値」タブが現れ、計算されたグローバル値が表示されます。

シミュレーション適合ノード

シミュレーション適合ノードは、統計分布の候補のセットをデータ内の各フィールドに適合させます。フィールドに対する各分布の適合は、適合度の基準を使用して評価されます。シミュレーション適合ノードが実行されると、シミュレーション生成ノードが作成されます (または、既存のノードが更新されます)。各フィールドは、最も適合する分布に割り当てられます。その後、シミュレーション生成ノードを使用して、フィールドごとにシミュレーション・データを生成することができます。

シミュレーション適合ノードはターミナル・ノードですが、生成されたモデル・パレットへのモデルの追加、「出力」タブへの出力の追加とグラフの追加、データのエクスポートは実行しません。

注: 過去のデータがまばらな場合 (欠損値が多い場合)、分布をデータに適合させるための十分な有効な値を適合コンポーネントによって検出するのが難しくなることがあります。データがまばらな場合、まばらなフィールドを削除するか (不要な場合)、欠損値に代入してから、適合を行う必要があります。データ検査ノードの「欠損値検査」タブのオプションを使用すると、完全なレコードの数を表示し、まばらなフィールドを特定して、代入方法を選択することができます。分布の適合用のレコード数が不足している場合、バランス・ノードを使用してレコード数を増やすことができます。

シミュレーション適合ノードによるシミュレーション生成ノードの自動作成

シミュレーション適合ノードを初めて実行すると、シミュレーション適合ノードへの更新リンクとともにシミュレーション生成ノードが作成されます。シミュレーション適合ノードをもう一度実行すると、更新リンクが削除されている場合のみ、新しいシミュレーション生成ノードが作成されます。また、シミュレーション適合ノードを使用して、接続されているシミュレーション生成ノードを更新することもできます。実行結果は、同じフィールドが両方のノードに存在するかによって異なります。また、フィールドがシミュレーション生成ノード内でロック解除されているかどうかによっても異なります。詳しくは、トピック 56 ページの『シミュレーション生成ノード』を参照してください。

シミュレーション適合ノードが持つことができるのは、シミュレーション生成ノードへの更新リンクだけです。シミュレーション生成ノードへの更新リンクを定義するには、以下の手順を実行します。

1. シミュレーション適合ノードを右クリックします。
2. メニューから「更新リンクの定義」を選択します。
3. 更新リンクの定義対象となるシミュレーション生成ノードをクリックします。

シミュレーション適合ノードとシミュレーション生成ノードの間の更新リンクを削除するには、対象の更新リンクを右クリックして「リンクを削除」を選択します。

分布の適合

統計分布は、変数が取ることができる値の理論的な出現頻度です。シミュレーション適合ノードでは、理論的分布のセットがデータの各フィールドと比較されます。適合で使用できる分布については、 66 ページの『分布』のトピックを参照してください。理論的分布のパラメーターは、適合度 (Anderson-Darling 基準または Kolmogorov-Smirnov 基準) の測定に応じてデータを最適化するために調整されます。シミュレーション適合ノードによる分布の適合の結果には、適合された分布、各分布用のパラメーターの最良の推定値、各分布でのデータの適合度が表示されます。分布の適合の実行中に、数値ストレージ・タイプを持つフィールド間での相関と、カテゴリー分布を持つフィールド間での不測の事態も計算されます。分布の適合の結果を使用して、シミュレーション生成ノードが作成されます。

すべての分布をデータに適合させる前に、最初の 1000 件のレコードに欠損値があるかどうかが検証されます。欠損値が多すぎる場合、分布の適合を行うことはできません。その場合、以下に示す方法のどれが適切かを判断する必要があります。

- 上流ノードを使用して、欠損値が存在するレコードを削除する。
- 上流ノードを使用して、欠損値用の値を代入する。

分布の適合を行っても、ユーザー欠損値は除外されません。データにユーザー欠損値が存在し、これらのユーザー欠損値を分布の適合から除外する場合は、それらの欠損値をシステム欠損値として設定する必要があります。

分布が適合する場合、フィールドの役割は考慮されません。例えば「対象」の役割を持つフィールドは、「入力」、「なし」、「両方」、「データ区分」、「分割」、「度数」、「**ID**」の役割を持つフィールドと同様に処理されます。

分布の適合の実行時は、フィールドのストレージ・タイプと尺度に応じて、フィールドの処理方法が異なります。分布の適合時におけるフィールドの処理方法について、以下の表に示します。

表 *45.* フィールドのストレージ・タイプと尺度に応じた分布の適合

<table>
<tr><td rowspan="2">ストレージ・タイプ</td><td colspan="6">測定水準</td></tr>
<tr><td>連続</td><td>カテゴリー型</td><td>フラグ</td><td>名義</td><td>順序</td><td>不明</td></tr>
<tr><td>文字列</td><td>不可能</td><td colspan="3">カテゴリー分布、ダイス分布、固定分布が適合されます。</td><td rowspan="8">フィールドは無視され、シミュレーション生成ノードには渡されません。</td></tr>
<tr><td>整数</td><td rowspan="6">すべての分布が適合されます。相関と不測の事態が計算されます。</td><td colspan="3" rowspan="6">カテゴリー分布が適合されます。相関は計算されません。</td><td rowspan="6">2 項分布、負の 2 項分布、ポワソン分布が適合され、相関が計算されます。</td></tr>
<tr><td>実数</td></tr>
<tr><td>時間</td></tr>
<tr><td>日付</td></tr>
<tr><td>タイムスタンプ</td></tr>
<tr><td>不明</td><td colspan="5">データにより、適切なストレージ・タイプが決定されます。</td></tr>
</table>

順序型の尺度を持つフィールドは連続型フィールドと同様に処理され、シミュレーション生成ノードの相関テーブルに含まれます。2 項分布、負の 2 項分布、ポワソン分布以外の分布を順序型フィールドに適合させる場合、フィールドの尺度を連続型に変更する必要があります。順序型フィールドの値ごとにラベルが既に定義されている場合、尺度を連続型に変更すると、これらのラベルが失われます。

複数の値を持つフィールドに対する分布の適合の実行時に、単一の値を持つフィールドも同様に処理されます。時間、日付、またはタイム・スタンプのストレージ・タイプを持つフィールドは、数値として処理されます。

分割フィールドに対する分布の適合

データに分割フィールドが含まれていて、分布の適合を分割ごとに個別に実行する場合、上流の再構成ノードを使用してデータを変換する必要があります。再構成ノードを使用して、分割フィールドの値ごとに新規フィールドを生成します。その後、この再構成データを使用して、シミュレーション適合ノードで分布の適合を実行することができます。

シミュレーション適合ノードの「設定」タブ

ソース・ノード名**:** 「自動」を選択すると、生成された (または更新された) シミュレーション生成ノードの名前を自動的に作成することができます。自動生成される名前は、カスタム名が指定されている場合はシミュレーション適合ノードで指定されている名前 (シミュレーション適合ノードでカスタム名が指定されていない場合は「Sim Gen」) になります。横にあるテキスト・フィールドにカスタム名を指定するには、「ユーザー設定」を選択します。このテキスト・フィールドを編集しなかった場合、「Sim Gen」がデフォルトのカスタム名になります。

適合オプション**:** このオプションを使用して、フィールドに対する分布の適合方法と、分布の適合の評価方法を指定することができます。

- サンプリングするケースの数: このオプションは、データ・セット内のフィールドに対して分布を適合させる場合に使用するケースの数を指定します。データ内のすべてのレコードに対して分布を適合させる場合は、「すべてのケース」を選択します。データ・セットのサイズが非常に大きい場合は、分布の適合に使用されるケースの数を制限することをお勧めします。最初の N 件のケースだけを使用する場合は、「初めの **N** ケースに制限」を選択します。使用するケースの数を指定するには、矢印をクリックします。また、上流ノードを使用して、分布の適合用のレコードの無作為サンプルを取得することもできます。
- 適合度基準 **(**連続型フィールドのみ**)**。連続型フィールドで分布をフィールドに適合させる場合は、分布をランク付けする適合度の Anderson-Darling 検定または Kolmogorov-Smirnoff 検定を選択します。Anderson-Darling 検定がデフォルトで選択されます。裾領域での最適な適合度を取得する場合は、特にこの検定をお勧めします。どちらの統計量も、すべての分布候補について計算されますが、分布を並び替えて最も適合する分布を判断する場合は、選択した統計量だけが使用されます。
- ビン **(**経験分布のみ**):** 連続型フィールドの場合、経験的分布が履歴データの累積分布関数になります。これは各値の確率 (値の範囲) であり、データから直接派生します。連続型フィールドの経験分布の計算で使用するビンの数を指定するには、矢印をクリックします。デフォルトは 100 で、最大値は 1000 です。
- 重みフィールド **(**オプション**):** データ・セットに重みフィールドが含まれている場合は、フィールド・ピッカー・アイコンをクリックして、リストから重みフィールドを選択します。この操作により、重みフィールドが分布の適合プロセスから除外されます。このリストには、連続型の尺度を持つデータ・セットのすべてのフィールドが表示されます。選択できる重みフィールドは 1 つだけです。

シミュレーション評価ノード

シミュレーション評価ノードは、指定されたフィールドを評価し、フィールドの分布を提供し、分布と相関のグラフを生成するターミナル・ノードです。このノードは、主に連続型フィールドを評価する目的で使用されます。そのため、このノードは、評価ノードによって生成される、離散型フィールドの評価に役立つ評価グラフを補完します。もう 1 つの相違点は、シミュレーション評価ノードは複数回の反復によって 1 つ

の予測を評価するのに対し、評価ノードは 1 回の反復で複数の予測を評価するという点です。複数の値がシミュレーション生成ノードの分布パラメーターに対して指定されると、反復が生成されます。詳しくは、トピック 66 ページの『反復回数』を参照してください。

シミュレーション評価ノードは、シミュレーション適合ノードとシミュレーション生成ノードから取得されたデータと共に使用されるように設計されています。ただし、このノードを他のノードと併用することもできます。任意の数の処理ステップを、シミュレーション生成ノードとシミュレーション評価ノードの間に配置することができます。

重要: シミュレーション評価ノードでは、少なくとも 1000 レコードの対象フィールドに有効な値が必要です。

シミュレーション評価ノードの「設定」タブ

シミュレーション評価ノードの「設定」タブでは、データ・セットの各フィールドの役割を指定し、シミュレーションによって生成された出力をカスタマイズすることができます。

項目の選択。シミュレーション評価ノードの 3 つのビュー (フィールド、密度関数、出力) を切り替えることができます。

フィールド・ビュー

対象フィールド: これは必須フィールドです。データ・セットの対象フィールドをドロップダウン・リストから選択するには、矢印をクリックします。選択したフィールドには、連続型、順序型、名義型のいずれかの尺度を設定することができますが、日付の尺度や指定されていない尺度を設定することはできません。

反復フィールド (オプション)。データの各レコードが属する反復を示す反復フィールドがデータに含まれている場合、ここでその反復フィールドを選択する必要があります。つまり、各反復は個別に評価されることになります。連続型、順序型、名義型のいずれかの尺度を持つフィールドのみ選択することができます。

入力データは反復より既にソート済み。 このオプションは、「反復フィールド (オプション)」フィールドに反復フィールドが指定されている場合のみ有効です。「反復フィールド (オプション)」に指定されている反復フィールドによって入力データがソートされていることを確認してから、このオプションを選択してください。

プロットする最大反復数。このオプションは、「反復フィールド (オプション)」フィールドに反復フィールドが指定されている場合のみ有効です。プロットする反復回数を指定するには、矢印をクリックします。この反復回数を指定すると、1 つのグラフで大量の反復をプロットすることがなくなります。1 つのグラフで大量の反復をプロットすると、グラフの解釈が難しくなります。最大反復回数として設定できる最小値は 2 で、最大値は 50 です。プロットする最大反復数の初期値は 10 に設定されています。

トルネード相関の入力フィールド。相関トルネードのグラフは、指定された対象値と指定された各入力値の間の相関係数を表示する棒グラフです。シミュレーションされた使用可能な入力値のリストから、トルネード・グラフに表示する入力フィールドを選択するには、フィールド・ピッカー・アイコンをクリックします。選択できるのは、連続型の尺度と順序型の尺度を持つ入力フィールドだけです。名義型、データ型不明、日付の入力フィールドはリスト内では使用不可になるため、選択できません。

「密度関数」ビュー

このビューのオプションを使用すると、連続型目標の確率密度関数と累積分布関数の出力や、カテゴリー型目標の予測値の棒グラフをカスタマイズすることができます。

密度関数。密度関数は、シミュレーションによって生成された一連の結果を調べるための主要な手段です。

- 確率密度関数 **(PDF)**。対象フィールドの確率密度関数を生成するには、このオプションを選択します。確率密度関数は、目標値の分布を表示します。確率密度関数を使用すると、目標が特定の領域内に存在する確率を判断することができます。カテゴリー型目標 (尺度が名義型または順序型の目標) の場合は、目標の各カテゴリーに該当するケースのパーセントを表示する棒グラフが生成されます。
- 累積分布関数 **(CDF)**。対象フィールドの累積分布関数を生成するには、このオプションを選択します。累積分布関数は、目標の値が指定の値以下になる確率を表示します。この関数は、連続型目標でのみ使用することができます。

基準線 **(連続型)**。このオプションは、「確率密度関数 **(PDF)**」または「累積分布関数 **(CDF)**」 (あるいはその両方) を選択した場合に使用可能になります。このオプションを使用すると、確率密度関数と累積分布関数に、さまざまな垂直方向の固定基準線を追加することができます。

- 平均**:** 基準線を対象フィールドの平均値の位置に追加するには、このオプションを選択します。
- 中央値: 基準線を対象フィールドの中央値の位置に追加するには、このオプションを選択します。
- 標準偏差**:** 対象フィールドの平均値を基準に、指定された標準偏差数の分だけプラス方向とマイナス方向の位置に基準線を追加するには、このオプションを選択します。このオプションを選択すると、その横にある「数値」フィールドが使用可能になります。標準偏差数を指定するには、矢印をクリックします。標準偏差数の最小値は 1 で、最大値は 10 です。標準偏差数の初期値は 3 に設定されています。
- パーセンタイル: 基準線を対象フィールドの分布の 2 つのパーセンタイル値の位置に追加するには、このオプションを選択します。このオプションを選択すると、その横にある「下部」テキスト・フィールドと「上部」テキスト・フィールドが使用可能になります。例えば、「上部」テキスト・フィールドに「90」を入力すると、対象の 90 パーセンタイルの位置に基準線が追加され、観測の 90% が該当する値となります。同様に、「下部」テキスト・フィールドに「10」を入力すると、対象の 10 パーセンタイルという意味になり、観測の 10% が該当する値となります。
- ユーザー指定の基準線。横軸に沿って、指定した位置に基準線を追加するには、このオプションを選択します。このオプションを選択すると、その横にある「値」テーブルが使用可能になります。有効な数値を「値」テーブルに入力するたびに、新しい空の行がテーブルの一番下に追加されます。有効な 数値は、対象フィールドの値の範囲内にある数値です。

注**:** (多重反復による) 複数の密度関数または分布関数が 1 つのグラフに表示される場合、カスタム線以外の基準線は各関数に対して個別に適用されます。

カテゴリー対象 **(PDF** のみ**)**。このオプションは、「確率密度関数 **(PDF)**」を選択した場合のみ使用可能になります。

- レポートするカテゴリー値。カテゴリー型の目標フィールドを持つモデルの場合、モデルの結果は、カテゴリーごとの予測確率のセットになります。各カテゴリーに、目標値が含まれます。最も高い確率を持つカテゴリーが予測カテゴリーとなり、このカテゴリーを使用して、確率密度関数の棒グラフが生成されます。 この棒グラフを生成するには、「予測カテゴリー」を選択します。対象フィールドのカテゴリーごとに予測確率の分布を示すヒストグラムを生成するには、「予測確率」を選択します。また、「両方」を選択して、両方のタイプのグラフを生成することもできます。
- 感度分析のグループ化。感度分析の反復を含むシミュレーションは、分析によって定義された反復ごとに、独立した対象フィールド (またはモデルの予測対象フィールド) を生成します。変化する分布パラメーターの値ごとに 1 つの反復が存在します。反復が存在する場合、カテゴリー対象フィールドの予測カテゴリーの棒グラフは、すべての反復の結果が含まれるクラスター棒グラフとして表示されます。「カテゴリーをグループ化する」または「反復をグループ化する」のいずれかを選択してください。

出力ビュー

目標分布のパーセンタイル値。このオプションにより、目標分布のパーセンタイル値のテーブルを作成し、表示するパーセンタイルを指定することができます。

パーセンタイル値のテーブルを作成。連続型対象フィールドの場合、このオプションを選択して、目標分布の指定されたパーセンタイルのテーブルを取得します。パーセンタイルを指定するには、以下に示すいずれかのオプションを選択します。

- **4** 分位: 4 分位は、対象フィールドの分布の 25、50、および 75 パーセンタイルです。観測は、同じサイズの 4 つのグループに分割されます。
- 区間。必要な等サイズ・グループの数が 4 以外の場合は、「区間」を選択します。このオプションを選択すると、その横にある「数値」フィールドが使用可能になります。間隔数を指定するには、矢印をクリックします。最小間隔数は 2 で、最大間隔数は 100 です。間隔数の初期値は 10 に設定されています。
- ユーザー指定のパーセンタイル。個別のパーセンタイル (99 パーセンタイルなど) を指定するには、「ユーザー指定のパーセンタイル」を選択します。このオプションを選択すると、その横にある「値」テーブルが使用可能になります。1 から 100 までの有効な数値を「値」テーブルに入力するたびに、新しい空の行がテーブルの一番下に追加されます。

シミュレーション評価ノードの出力

シミュレーション評価ノードが実行されると、出力マネージャーに出力が追加されます。シミュレーション評価出力ブラウザーには、シミュレーション評価ノードの実行結果が表示されます。通常の保存、エクスポート、および印刷関連オプションは「ファイル」メニューから、通常の編集オプションは「編集」メニューから利用できます。 詳しくは、トピック 323 ページの『出力を表示』を参照してください。「表示」メニューは、いずれかのグラフが選択された場合のみ使用可能になります。分布表と分布情報の出力の場合は、使用できません。「表示」メニューから「編集モード」を選択して、グラフのレイアウトと外観を変更することも、「探索モード」を選択して、グラフで表されるデータと値を探索することもできます。静的モードの場合、グラフの基準線とスライダーが現在の位置に固定されるため、基準線とスライダーを移動することはできません。静的モードは、基準線を持つグラフのコピー、印刷、エクスポートを行う場合に使用できる唯一のモードです。このモードを選択するには、「表示」メニューで「静的モード」をクリックします。

シミュレーション評価出力ブラウザーのウィンドウは、2 つのパネルで構成されています。ウィンドウの左側には、シミュレーション評価ノードの実行時に生成されたグラフのサムネール表現を表示するナビゲーション・パネルが配置されます。サムネールが選択されると、ウィンドウの右側のパネルにグラフ出力が表示されます。

ナビゲーション・パネル

出力ブラウザーのナビゲーション・パネルには、シミュレーションから生成されたグラフのサムネールが表示されます。ナビゲーション・パネルに表示されるサムネールは、対象フィールドの尺度と、「シミュレーション評価ノード (Simulation Evaluation node)」ダイアログ・ボックスで選択されるオプションよって異なります。以下の表に、サムネールの説明を示します。

表 46. ナビゲーション・パネルのサムネール

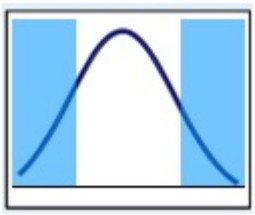
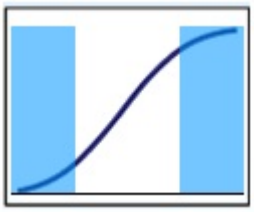
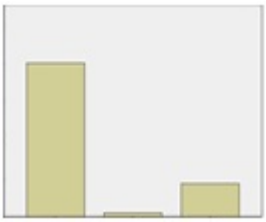
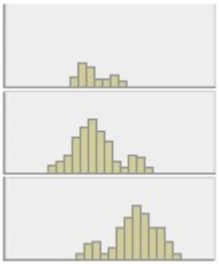
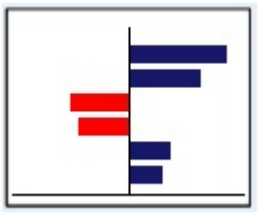
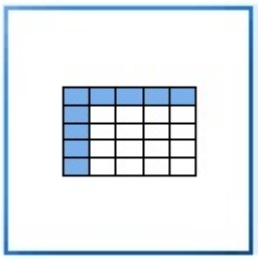
サムネール	説明	コメント
	確率密度関数	このサムネールが表示されるのは、対象フィールドの尺度が連続型で、「確率密度関数 **(PDF)**」が「シミュレーション評価ノード」ダイアログ・ボックスの「密度関数」ビューで選択された場合だけです。 対象フィールドの尺度がカテゴリー型の場合、このサムネールは表示されません。
	累積分布関数	このサムネールが表示されるのは、対象フィールドの尺度が連続型で、「累積分布関数 **(CDF)**」が「シミュレーション評価ノード」ダイアログ・ボックスの「密度関数」ビューで選択された場合だけです。 対象フィールドの尺度がカテゴリー型の場合、このサムネールは表示されません。
	予測カテゴリー値	このサムネールが表示されるのは、対象フィールドの尺度がカテゴリー型で、「確率密度関数 **(PDF)**」が「シミュレーション評価ノード」ダイアログ・ボックスの密度関数ビューで選択され、「予測カテゴリー」または「両方」が「レポートするカテゴリー値」領域で選択された場合だけです。 対象フィールドの尺度が連続型の場合、このサムネールは表示されません。
	予測カテゴリー確率	このサムネールが表示されるのは、対象フィールドの尺度がカテゴリー型で、「確率密度関数 **(PDF)**」が「シミュレーション評価ノード」ダイアログ・ボックスの「密度関数」ビューで選択され、「予測確率」または「両方」が「レポートするカテゴリー値」領域で選択された場合だけです。 対象フィールドの尺度が連続型の場合、このサムネールは表示されません。

表 46. ナビゲーション・パネルのサムネール (続き)

サムネール	説明	コメント
	トルネード・グラフ	このサムネールが表示されるのは、「シミュレーション評価ノード」ダイアログ・ボックスの「フィールド」ビューの「トルネード相関の入力フィールド」フィールドで 1 つ以上の入力フィールドが選択された場合だけです。
	分布表	このサムネールが表示されるのは、対象フィールドの尺度が連続型で、「パーセンタイル値の表を作成」が「シミュレーション評価ノード」ダイアログ・ボックスの「出力」ビューで選択された場合だけです。このグラフでは、「表示」メニューが使用不可になります。 対象フィールドの尺度がカテゴリー型の場合、このサムネールは表示されません。
	情報	このサムネールは常に表示されます。この出力では、「表示」メニューが使用不可になります。

グラフ出力

使用可能な出力グラフのタイプは、反復フィールドを使用するかどうかにかかわらず、対象フィールドの尺度と、「シミュレーション評価ノード (Simulation Evaluation node)」ダイアログ・ボックスで選択されるオプションよって異なります。シミュレーションから生成された各種グラフには、表示をカスタマイズするために使用できるインタラクティブ機能が組み込まれています。「グラフ オプション」をクリックすると、インタラクティブ機能が使用可能になります。すべてのシミュレーション・グラフが、グラフボードでの視覚化に対応しています。

連続型目標の確率密度関数グラフ: このグラフには確率と度数が表示され、左の縦軸に確率のスケール、右の縦軸に度数のスケールが配置されます。グラフには 2 つのスライドする縦の基準線があり、これらの基準線によってグラフが個別の領域に分割されます。グラフの下のテーブルには、各領域内の分布の割合が表示されます。複数の密度関数が (反復のために) 同じグラフ上に表示される場合、各密度関数に関連付けられた確率の独立した行、反復名が含まれた追加の列、および各密度関数に関連付けられた色がテーブルに表示されます。反復は、反復ラベルに応じてアルファベット順でテーブルに表示されます。使用可能な反復ラベルがない場合は、代わりに反復値が使用されます。テーブルを編集することはできません。

各基準線には、線を簡単に移動するためのスライダー (逆三角形) があります。各スライダーには、現在の位置を示すラベルがあります。デフォルトでは、スライダーは、分布の 5 パーセンタイルと 95 パーセン

タイルの位置に配置されます。複数の反復が存在する場合、テーブルに表示された最初の反復の 5 パーセンタイルと 95 パーセンタイルの位置にスライダーが配置されます。線を移動して相互に交差させることはできません。

「グラフ オプション」をクリックすると、各種の追加機能が使用可能になります。具体的には、スライダーの位置を明示的に設定したり、固定基準線を追加したり、グラフの表示を連続曲線からヒストグラムに変更したりすることができます。詳しくは、トピック 362 ページの『グラフ・オプション』を参照してください。グラフのコピーやエクスポートを行うには、グラフを右クリックします。

連続型目標の累積分布関数グラフ: このグラフにも、確率密度関数で説明した 2 つの移動可能な縦の基準線と関連するテーブルがあります。複数の反復が存在する場合、スライダー・コントロールとテーブルは、確率密度関数と同じ動作を行います。各反復に属する密度関数を識別するために使用されるものと同じ色が、分布関数でも使用されます。

このグラフから、「グラフ オプション」ダイアログ・ボックスにアクセスすることもできます。このダイアログ・ボックスで、スライダーの位置を明示的に設定したり、固定基準線を追加したり、累積分布関数を増加関数 (デフォルト) と減少関数のどちらで表示するかを指定したりすることができます。詳しくは、トピック 362 ページの『グラフ・オプション』を参照してください。グラフのコピー、エクスポート、編集を行うには、グラフを右クリックします。「編集」を選択すると、グラフボード・エディターのフローティング・ウィンドウにグラフが表示されます。

カテゴリー対象の予測カテゴリー値のグラフ**:** カテゴリー対象フィールドの場合、棒グラフに予測値が表示されます。予測値は、各カテゴリーに該当することが予測される対象フィールドのパーセントとして表示されます。感度分析の反復があるカテゴリー対象フィールドの場合、予測目標カテゴリーの結果は、すべての反復の結果が含まれるクラスター化された棒グラフとして表示されます。このグラフは、カテゴリー別または反復別にクラスター化されます。どちらの方法でクラスター化されるかは、「シミュレーション評価ノード (Simulation Evaluation node)」ダイアログ・ボックスの「密度関数」ビューの「感度分析のグループ化」領域で選択されたオプションによって異なります。グラフのコピー、エクスポート、編集を行うには、グラフを右クリックします。「編集」を選択すると、グラフボード・エディターのフローティング・ウィンドウにグラフが表示されます。

カテゴリー対象の予測カテゴリー確率のグラフ**:** カテゴリー対象フィールドの場合、ヒストグラムには、対象のカテゴリーごとに予測確率の分布が表示されます。感度分析の反復が存在するカテゴリー対象フィールドの場合、ヒストグラムは、カテゴリー別または反復別に表示されます。どちらの方法で表示されるかは、「シミュレーション評価ノード (Simulation Evaluation node)」ダイアログ・ボックスの「密度関数」ビューの「感度分析のグループ化」領域で選択されたオプションによって異なります。ヒストグラムがカテゴリー別にグループ化されている場合、反復ラベルを含むドロップダウン・リストを使用すると、表示する反復を選択することができます。また、グラフを右クリックして「反復」サブメニューから反復を選択することにより、表示する反復を選択することもできます。ヒストグラムが反復別にグループ化されている場合、カテゴリー名を含むドロップダウン・リストを使用すると、表示する反復を選択することができます。また、グラフを右クリックして「カテゴリー」サブメニューからカテゴリーを選択することにより、表示するカテゴリーを選択することもできます。

このグラフは、モデルのサブセットでのみ使用することができます。モデル・ナゲットで、すべてのグループ確率を生成するオプションを選択する必要があります。例えば、ロジスティック・モデル・ナゲットでは、「すべての確率を追加」を選択する必要があります。このオプションは、以下のモデル・ナゲットでサポートされています。

- ロジスティック、SVM、Bayes、ニューラル・ネットワーク、KNN

- ロジスティック回帰、デシジョン・ツリー、naïve Bayes の Db2/ISW データベース内マイニング・モデル

デフォルトでは、すべてのグループ確率を生成するオプションは、これらのモデル・ナゲットでは選択されません。

トルネード・グラフ: トルネード・グラフは、指定された各入力値に対する対象フィールドの感度を表示する棒グラフです。感度は、対象を各入力値と相関させることよって測定されます。グラフのタイトルには、対象フィールドの名前が表示されます。グラフ上の各バーは、対象フィールドと入力フィールドの間の相関を表します。グラフに表示されるシミュレーション後の入力値は、「シミュレーション評価ノード」ダイアログ・ボックスの「フィールド」ビューの「トルネード相関の入力フィールド」フィールドで選択された入力値です。それぞれの棒には、相関値によってラベルが付けられます。それぞれの棒は、相関の絶対値によって大きい値から小さい値の順に並べ替えられます。反復が存在する場合、反復ごとに個別のグラフが生成されます。各グラフには、反復の名前を含むサブタイトルがあります。

分布表: この表には、指定された観測の割合を上回る対象フィールドの値が表示されます。この表には、「シミュレーション評価ノード (Simulation Evaluation node)」ダイアログ・ボックスの「出力」ビューで指定されたパーセンタイル値ごとに行が表示されます。パーセンタイル値には、4 分位、等間隔に配置された異なる数値のパーセンタイル、または個別に指定されたパーセンタイルを指定することができます。分布表には、反復ごとに列が表示されます。

情報: このセクションには、評価で使用されるフィールドとレコードの全体的な要約が表示されます。また、反復ごとに分割された入力フィールドとレコード件数も表示されます。

グラフ・オプション

「グラフ オプション」ダイアログ・ボックスでは、シミュレーションから生成された確率密度関数と累積分布関数のアクティブなグラフの表示をカスタマイズすることができます。

表示: 「表示」ドロップダウン・リストは、確率密度関数のグラフにのみ適用されます。これを使用して、グラフの表示を連続曲線からヒストグラムに切り替えることができます。(多重反復による) 複数の密度関数が同じグラフ上に表示される場合、この機能は使用できません。複数の密度関数が存在する場合、密度関数は連続曲線としてのみ表示することができます。

順序: 「順序」ドロップダウン・リストは、累積分布関数のグラフにのみ適用されます。このオプションは、累積密度関数を増加関数 (デフォルト) と減少関数のどちらで表示するかを指定します。減少関数で表示する場合、横軸の特定のポイントにある関数の値は、そのポイントの右側に対象フィールドがある確率を示します。

スライダーの位置: 「上限」テキスト・フィールドには、スライドする右側の基準線の現在位置が表示されます。「下限」テキスト・フィールドには、スライドする左側の基準線の現在位置が表示されます。「上限」と「下限」テキスト・フィールドに値を入力することにより、スライダーの位置を明示的に設定することができます。「下限」テキスト・フィールドには、必ず「上限」テキスト・フィールドの値よりも小さな値を入力する必要があります。「負の無限方向」を選択し、基準線の位置を実質的に負の無限大に設定することにより、左側の基準線を削除することができます。この操作を実行すると、「下限」テキスト・フィールドが無効になります。「正の無限方向」を選択し、基準線の位置を実質的に無限大に設定することにより、右側の基準線を削除することができます。この操作を実行すると、「上限」テキスト・フィールドが無効になります。両方の基準線を削除することはできません。「負の無限方向」を選択すると「正の無限方向」チェック・ボックスが無効になり、「正の無限方向」を選択すると「負の無限方向」チェック・ボックスが無効になります。

基準線: 確率密度関数と累積分布関数に、さまざまな垂直方向の固定参照線を追加することができます。

- 平均**:** 基準線を対象フィールドの平均値の位置に追加することができます。
- 中央値: 基準線を対象フィールドの中央値の位置に追加することができます。
- 標準偏差**:** 対象フィールドの平均値を基準に、指定された標準偏差数の分だけプラス方向とマイナス方向の位置に基準線を追加することができます。使用する標準偏差数を横のテキスト・フィールドに入力することができます。標準偏差数の最小値は 1 で、最大値は 10 です。標準偏差数の初期値は 3 に設定されています。
- パーセンタイル: 「下部」テキスト・フィールドと「上部」テキスト・フィールドに値を入力することにより、対象フィールドの分布について、1 または 2 パーセンタイル値の位置に基準線を追加することができます。例えば、「上部」テキスト・フィールドに「95」を入力すると、95 パーセンタイルという意味になり、観測の 95% が該当する値となります。同様に、「下部」テキスト・フィールドに「5」を入力すると、5 パーセンタイルという意味になり、観測の 5% が該当する値となります。「下部」テキスト・フィールドの場合、パーセンタイルの最小値は 0 で最大値は 49 です。「上部」テキスト・フィールドの場合、パーセンタイルの最小値は 50 で最大値は 100 です。
- ユーザー指定の位置: 横軸に沿って、指定した位置に基準線を追加することができます。グリッドからエントリーを削除することにより、ユーザー指定の基準線を削除することができます。

「**OK**」をクリックすると、スライダー、スライダー上部のラベル、基準線、グラフ下部の表が更新され、「グラフ オプション」ダイアログ・ボックスで選択されたオプションが反映されます。何も変更せずにダイアログ・ボックスを閉じる場合は、「キャンセル」をクリックします。基準線を削除するには、「グラフ オプション」ダイアログで関連する選択項目を選択解除し、「**OK**」をクリックします。

注**:** (感度分析の反復結果のために) 複数の密度関数または分布関数が 1 つのグラフに表示される場合、カスタム線以外の基準線は、各関数に対して個別に適用されます。最初の反復の基準線のみ表示されます。基準線ラベルには反復ラベルが含まれます。反復ラベルは、上流 (通常はシミュレーション生成ノード) から取得されます。使用可能な反復ラベルがない場合は、代わりに反復値が使用されます。「平均値」、「中央値」、「標準偏差」、「**100** 分位」の各オプションは、多重反復の累積分布関数の場合は使用不可になります。

拡張出力ノード

拡張出力ノード・ダイアログの「出力」タブで「画面に出力」が選択されている場合は、出力ブラウザー・ウィンドウの画面上に出力が表示されます。出力は、出力マネージャーにも追加されます。出力ブラウザー・ウィンドウは、それ自体に出力を印刷または保存、または別の形式で出力をエクスポートできるメニュー・セットを備えています。「編集」メニューに含まれているのは「コピー」オプションのみです。拡張出力ノードの出力ブラウザーには、テキスト出力が表示される「テキスト出力」タブと、グラフおよび図表が表示される「グラフ出力」タブの 2 つのタブが表示されます。

拡張出力ノード・ダイアログの「出力」タブで「ファイルに出力」が選択されている場合、拡張出力ノードの実行の成功時に出力ブラウザー・ウィンドウは表示されません。

拡張出力ノードの「シンタックス」タブ

シンタックスのタイプ (**R** または **Python for Spark**) を選択します。詳しくは、以下のセクションを参照してください。シンタックスの準備ができたら、「実行」をクリックして、拡張出力ノードを実行できます。出力オブジェクトは出力マネージャーに追加されるか、オプションで、「出力」タブの「ファイル名」フィールドで指定されたファイルに追加されます。

R シンタックス

「**R** シンタックス」。データ分析用のカスタムの R スクリプト・シンタックスを、このフィールドに入力するか、貼り付けることができます。

「フラグ型フィールドの変換」。フラグ型フィールドの処理方法を指定します。「文字列を因子に、整数および実数を倍精度に」および「論理値 **(**真、偽**)**」の 2 つのオプションがあります。「論理値 **(**真、偽**)**」を選択した場合、フラグ型フィールドの元の値は失われます。例えば、フィールドに「Male」および「Female」という値がある場合、これらの値は「真」および「偽」に変更されます。

「欠損値を **R** の欠損値 **(NA)** に変換」。選択すると、欠損値はすべて、R の NA 値に変換されます。R では、欠損値の識別に値 NA が使用されます。使用する R 関数によっては、データに NA が含まれていた場合の関数の動作を制御するために使用される引数が含まれている場合があります。例えば、関数によって NA を含むレコードを自動的に除外することを選択できる場合があります。このオプションが選択されない場合、すべての欠損値はそのまま R に渡されます。これらの欠損値は R スクリプトの実行時にエラーの原因となる可能性があります。

「時間帯を考慮した特殊な制御で日時フィールドを **R** のクラスに変換」。選択すると、日付形式または日付/時刻形式の変数が R の日付/時刻形式に変換されます。次のいずれかのオプションを選択する必要があります。

- 「**R POSIXct**」。日付形式または日付/時刻形式の変数が R の POSIXctオブジェクトに変換されます。
- 「**R POSIXlt (**リスト**)**」。日付形式または日付/時刻形式の変数が R の POSIXltオブジェクトに変換されます。

注**:** POSIX 形式は、拡張オプションです。これらのオプションは、ご使用の R スクリプトで、これらの形式を必要とする方法で日付/時刻フィールドを処理するように指定している場合にのみ使用してください。POSIX 形式は、時刻形式の変数には適用されません。

Python シンタックス

「**Python** シンタックス」。データ分析用のカスタムの Python スクリプト・シンタックスを、このフィールドに入力するか、貼り付けることができます。Python for Spark について詳しくは、Python for Spark およびPython for Spark を使用したスクリプトを参照してください。

拡張出力ノードの「コンソール出力」タブ

「コンソール出力」タブには、「シンタックス」タブの R スクリプトまたは Python for Spark スクリプトが実行されたときに受信するすべての出力が含まれます (例えば、R スクリプトを使用する場合、「シンタックス」タブの「**R** シンタックス」フィールドにある R スクリプトが実行されたときに R コンソールから受信する出力が表示されます)。この出力には、R スクリプトまたは Python スクリプトの実行時に生成される R または Python のエラー・メッセージや警告が含まれる場合があります。出力は、主にスクリプトをデバッグするために使用できます。「コンソール出力」タブには、「**R** シンタックス」フィールドまたは「**Python** シンタックス」フィールドのスクリプトも表示されます。

拡張出力スクリプトが実行されるたびに、R コンソールまたは Python for Spark から受信した出力で「コンソール出力」タブの内容が上書きされます。出力を編集することはできません。

拡張出力ノードの「出力」タブ

出力名。ノードの実行時に作成される出力の名前を指定します。「自動」を選択すると、スクリプト・タイプに応じて、出力の名前は自動的に「R 出力」または「Python 出力 (Python Output)」に設定されます。「ユーザー設定」 で別の名前を指定することもできます。

画面に出力。出力を生成し、新規ウィンドウに表示するには、このオプションを選択します。出力は、出力マネージャーにも追加されます。

ファイルに出力。出力をファイルに保存するには、このオプションを選択します。このオプションを選択すると、「グラフの出力」および「ファイルの出力」ラジオ・ボタンが有効になります。

出力グラフ。「ファイルに出力」が選択された場合にのみ有効になります。拡張出力ノードを実行した結果のグラフをすべてファイルに保存するには、このオプションを選択します。「ファイル名」フィールドで、生成された出力に使用するファイル名を指定してください。特定のファイルと場所を選択するには、省略記号ボタン (「**...**」) をクリックしてください。「ファイルの種類」ドロップダウン・リストでファイルの種類を指定します。次のファイルの種類を利用できます。

- 出力オブジェクト (`.cou`)
- HTML (`.html`)

テキストの出力: 「ファイルに出力」が選択された場合にのみ有効になります。拡張出力ノードを実行した結果のテキスト出力をすべてファイルに保存するには、このオプションを選択します。「ファイル名」フィールドで、生成された出力に使用するファイル名を指定してください。特定のファイルと場所を指定するには、省略記号ボタン (「**...**」) をクリックしてください。「ファイルの種類」ドロップダウン・リストでファイルの種類を指定します。次のファイルの種類を利用できます。

- HTML (`.html`)
- 出力オブジェクト (`.cou`)
- テキスト文書 (`.txt`)

拡張出力ブラウザー

拡張出力ノード・ダイアログ・ボックスの「出力」タブで「画面に出力」が選択されている場合は、出力ブラウザー・ウィンドウの画面上に出力が表示されます。出力は、出力マネージャーにも追加されます。出力ブラウザー・ウィンドウは、それ自体に出力を印刷または保存、または別の形式で出力をエクスポートできるメニュー・セットを備えています。「編集」メニューに含まれているのは「コピー」オプションのみです。拡張出力ノードの出力ブラウザーには、2 つのタブが表示されます。

- 「テキスト出力」タブには、テキスト出力が表示されます。
- 「グラフ出力」タブには、グラフおよび図表が表示されます。

拡張出力ノード・ダイアログ・ボックスの「出力」タブで「画面に出力」の代わりに「ファイルに出力」が選択されている場合、拡張出力ノードの実行の成功時に出力ブラウザー・ウィンドウは表示されません。

拡張出力ブラウザーの「テキスト出力」タブ

「テキスト出力」タブには、拡張出力ノードの「シンタックス」タブにある R スクリプトまたは Python for Spark スクリプトが実行された際に生成される、すべてのテキスト出力が表示されます。

注: 拡張出力スクリプトの実行結果として生成される R または Python for Spark のエラー・メッセージや警告は、拡張出力ノードの「コンソール出力」タブに常に表示されます。

拡張出力ブラウザーの「グラフ出力」タブ

「グラフ出力」タブには、拡張出力ノードの「シンタックス」タブにある R スクリプトまたは Python for Spark スクリプトが実行された際に生成される、すべてのグラフまたは図表が表示されます。例えば、R スクリプトに R の `plot` 関数への呼び出しが含まれる場合、その結果のグラフがこのタブに表示されます。

IBM SPSS Statistics ヘルパー アプリケーション

IBM SPSS Statistics の互換性バージョンがコンピューターにインストールされ、ライセンス供与されると、IBM SPSS Modeler を設定し、Statistics 変換ノード、Statistics モデル・ノード Statistics 出力ノード、または Statistics エクスポート・ノードを使用して、IBM SPSS Statistics の機能によりデータを処理することができます。

現在のバージョンの IBM SPSS Modeler との製品の互換性について詳しくは、当社サポート・サイト (*http://www.ibm.com/support*) を参照してください。

IBM SPSS Modeler で IBM SPSS Statistics や他のアプリケーションを利用できるようにするには、次のメニューを選択します。

「ツール」 > 「オプション」 > 「ヘルパー アプリケーション」

IBM SPSS Statistics インタラクティブ**:** Statistics エクスポート・ノードによって生成されたデータ・ファイルで直接 IBM SPSS Statistics を起動する場合に使用するコマンドのフルパスと名前を入力します (例: *C:¥Program Files¥IBM¥SPSS¥Statistics¥<nn>¥stats.exe*)。詳しくは、トピック 386 ページの『Statistics エクスポート・ノード』を参照してください。

接続**:** IBM SPSS Statistics サーバーが IBM SPSS Modeler Server と同じホストに存在する場合は、これら 2 つのアプリケーションの間の接続を有効にすることができます。これにより、分析中にデータがサーバーに残るため、処理効率が向上します。「サーバー」 を選択すると、下の 「ポート」 オプションが有効になります。デフォルトの設定は 「ローカル」 です。

ポート**:** IBM SPSS Statistics サーバーのサーバー・ポートを指定します。

IBM SPSS Statistics ロケーション・ユーティリティー**:** IBM SPSS Modeler で Statistics 変換ノード、Statistics モデル・ノード、Statistics 出力ノードを使用できるようにするには、ストリームが実行されているコンピューターに IBM SPSS Statistics のコピーがインストールされ、ライセンスが交付されている必要があります。

- ローカル (スタンドアロン) モードで IBM SPSS Modeler を実行中の場合、IBM SPSS Statistics のライセンスが付与されたドライバーをローカル・コンピューターに搭載されている必要があります。このボタンをクリックして、ライセンスに使用するローカル IBM SPSS Statistics インストールの場所を指定します。
- また、リモートの IBM SPSS Modeler Server に対して分散モードで実行する場合は、IBM SPSS Modeler Server ホストでユーティリティーを実行して *statistics.ini* ファイルを作成し、IBM SPSS Modeler Server のインストール・パスを IBM SPSS Statistics に指定する必要があります。ライセンス設定を行うには、Windows の場合、IBM SPSS Modeler Server *bin* ディレクトリーに移動して、コマンド・プロンプトで次の文を実行します。

  ```
  statisticsutility -location=<IBM SPSS Statistics_installation_path>/
  ```

  また、UNIX の場合は、次を実行します。

```
./statisticsutility -location=<IBM SPSS Statistics_installation_path>/bin
```

IBM SPSS Statistics のライセンス認証されたコピーがローカル・マシンにない場合でも、IBM SPSS Statistics サーバーに対して Statistics ファイル・ノードを実行できますが、他の IBM SPSS Statistics ノードを実行しようとするとエラー・メッセージが表示されます。

コメント

IBM SPSS Statistics 手続きノードの実行に何か問題がある場合は、次の事柄を確認してください。

- IBM SPSS Modeler で使用されているフィールド名が 8 文字 (IBM SPSS Statistics 12.0 より前のバージョン) より長い場合、64 文字 (IBM SPSS Statistics 12.0 以降のバージョン) より長い場合、または不正な文字が含まれている場合は、IBM SPSS Statistics に読み込む前にフィールド名を変更するか、または名前を切り詰める必要があります。詳しくは、トピック 388 ページの『IBM SPSS Statistics 用のフィールドの名前変更またはフィルタリング』を参照してください。
- IBM SPSS Modeler の後に IBM SPSS Statistics をインストールした場合、上記で説明したとおり、IBM SPSS Statistics の場所を指定する必要があります。

第 7 章 エクスポート・ノード

エクスポート・ノードの概要

エクスポート・ノードは、他のソフトウェア ツールのインターフェースに対応した、さまざまな形式でデータをエクスポートできるメカニズムも備えています。

次のエクスポート・ノードを利用できます。

データベース・エクスポート・ノードで、データを ODBC 対応のリレーショナル・データ・ソースに書き込みます。ODBC データ・ソースに書き込むには、データ・ソースが存在し、そのデータ・ソースに対する書き込み権限を取得している必要があります。

ファイル・ノードでは、データが区切り文字で区切られたテキスト・ファイルへ出力されます。このことは、他の分析ソフトウェアや表計算ソフトウェアに読み込める形式でデータをエクスポートする場合に、役立ちます。

Statistics エクスポート・ノードでは、IBM SPSS Statistics *.sav* または *.zsav* フォーマットでデータを出力します。*.sav* または *.zsav* ファイルは、IBM SPSS Statistics Base およびその他の製品で読み込むことができます。この形式は、IBM SPSS Modeler のキャッシュ・ファイルでも使用されます。

Data Collection エクスポート・ノードは、Data Collection の市場調査ソフトウェアで使用する形式でデータを出力します。このノードを使用するには、Data Collection Data Library がインストールされている必要があります。

IBM Cognos エクスポート・ノードは、Cognos データベースで読み取ることができる形式でデータをエクスポートできます。

IBM Cognos TM1 エクスポート・ノードは、Cognos TM1 データベースで読み取ることができる形式でデータをエクスポートできます。

SAS エクスポート・ノードで、SAS または SAS 互換ソフトウェア・パッケージで読み込むデータを、SAS 形式で出力できます。3 つの SAS ファイル形式が利用可能です。SAS for Windows/OS2、SAS for UNIX、または SAS バージョン 7/8

Excel エクスポート ノードでは、データを Microsoft Excel .xlsx ファイル形式で出力します。オプションで、ノードが実行されるときに自動的に Excel が起動し、エクスポートするファイルを開けるように選択できます。

XML エクスポート・ノードでは、XML 形式のファイルにデータを出力します。オプションで、エクスポートしたデータをストリームに読み込む XML ソース・ノードを作成できます。

データベース・エクスポート・ノード

データベース・ノードを使用して、データを ODBC 準拠の関連データ・ソースに書き込むことができます。詳細は、データベース・ソース・ノードに記載されています。 詳しくは、トピック 19 ページの『データベース・ソース・ノード』を参照してください。

データベースにデータを書き込むには、次の作業を行います。

1. 使用するデータベースに ODBC ドライバーをインストールして、データ・ソースを構成します。
2. データベース・ノードの「エクスポート」タブで、書き込み先のデータ・ソースとテーブルを指定します。新規テーブルを生成するか、既存のテーブルにデータを挿入します。
3. 必要に応じてオプションを追加します。

これらの作業の詳細は、続く各トピックで説明します。

データベース・ノードの「エクスポート」タブ

注: エクスポート先にできるいくつかのデータベースでは、テーブル内の 30 文字よりも長い列名をサポートしない場合があります。テーブルに正しくない列名があることを示すエラー・メッセージが表示された場合は、名前のサイズを 30 文字よりも短くしてください。

データ・ソース: 選択したデータ・ソースが表示されます。名前を入力するか、ドロップダウン・リストから選択します。リスト中に目的のデータベースが見つからない場合は、「新規データベース接続の追加」を選択して、「データベース接続」ダイアログ・ボックスから目的のデータベースを検索してください。詳しくは、トピック 20 ページの『データベース接続の追加』を参照してください。

テーブル名: データの送信先のテーブル名を入力します。「テーブルへ挿入」 オプションを選択した場合、「選択」 ボタンをクリックして、データベース中の既存のテーブルを選択することができます。

テーブルの作成: 新規データベース・テーブルを作成する場合、または既存のデータベース・テーブルを上書きする場合に選択します。

テーブルへ挿入: 既存のデータベース・テーブルにデータを新しい行として挿入する場合に選択します。

テーブルを結合: (可能な場合) 選択したデータベースの列を、該当する入力データフィールドの値で更新します。このオプションを選択すると 「結合」 ボタンが有効となり、入力データフィールドをデータベース列に関連付けできるダイアログが表示されます。

既存のテーブルを削除: 新しいテーブルの作成時に、同じ名前を持つ既存のテーブルを削除する場合に選択します。

既存の行を削除: テーブルへの挿入時に、エクスポート前に既存の行をテーブルから削除する場合に選択します。

注: 上記の 2 種類のオプションのどちらも選択されていない場合、ノードの実行時に「上書き警告」というメッセージが表示されます。この警告メッセージを表示しない場合は、「ユーザー オプション」ダイアログ・ボックスの「通知」タブにある、「ノードがデータベース テーブルを上書きする時に警告」 を選択します。

デフォルトの文字列サイズ : 上流にあるデータ型ノードでデータ型不明とされたフィールドは、データベースに文字列フィールドとして書き込まれます。ここには、データ型不明フィールドに使う文字列のサイズを指定します。

「スキーマ」 をクリックしてダイアログ・ボックスを開きます。ここでさまざまなエクスポート・オプション (この機能をサポートするデータベース) を設定、SQL データ型をフィールドに設定し、データベース・インデックス用の第 1 入力フィールド・キーを指定します。詳しくは、トピック 373 ページの『データベース・エクスポートのスキーマのオプション』を参照してください。

「インデックス」 をクリックして、エクスポートされたテーブルのインデックスを生成するオプションを指定してデータベースのパフォーマンスを高めます。詳しくは、トピック 375 ページの『データベース・エクスポートのインデックス・オプション』を参照してください。

バルク・ロードおよびデータベースのコミットに関するオプションを指定するには、「詳細」 をクリックします。詳しくは、トピック 377 ページの『データベース・エクスポートの拡張オプション』を参照してください。

表および列名を引用符で囲む: データベースに CREATE TABLE ステートメントを送信するときに使用するオプションを選択します。スペースまたは非標準文字を含むテーブルや列名は引用符で囲む必要があります。

- 必要に応じて: 引用符が必要かどうかを IBM SPSS Modeler が個別に判断して、自動的に引用符で囲む場合に選択します。
- 常時: テーブル名と列名を常に引用符で囲む場合に選択します。
- しない: 引用符を使用しない場合に選択します。

データのインポート ノードを生成: 指定したデータ・ソースとテーブルにエクスポートしたように、データベース・ソース・ノードを生成する場合に選択します。実行時に、ストリーム領域にこのノードが追加されます。

データベース・エクスポート結合オプション

入力データのフィールドを、対象データベース・テーブルの列にマッピングできます。入力データ・フィールドがデータベース列に関連付けされると、ストリーム実行時に列の値が入力データ値に置き換えられます。データベース内の関連付けられていない入力フィールドは、変わらないままです。

フィールドをマッピング: 入力データ・フィールドとデータベース列とのマッピングを指定します。データベースの列と名前が同じ入力データ・フィールドは自動的にマッピングされます。

- マップ。ボタンの左側にあるフィールド リストで選択した入力データ フィールドを、右側のリストで選択したデータベース列にマッピングします。一度に複数のフィールドをマッピングできますが、両方のリストで選択するエントリー数は同じでなければなりません。
- マップ解除。選択された 1 つ以上のデータベース列のマッピングを解除します。ダイアログ右側のテーブルでフィールド列またはデータベース列を選択すると、このボタンが有効になります。
- 追加**:** ボタンの左側にあるフィールド・リストで選択した 1 つまたは複数の入力データ・フィールドを右側のリストに追加すると、マッピングできます。左側のリストでフィールドを選択し、右側のリストにその名前のフィールドがない場合、このボタンが有効になります。このボタンをクリックすると、選択したフィールドを、同じ名前の新しいデータベース列にマッピングします。単語 **<NEW>** がデータベース列の名前の後に表示され、新しいフィールドであることを示します。

行を結合**:** トランザクション *ID* などのキー・フィールドを使用して、同じキー・フィールドの値を持つレコードを結合します。これは、データベースの「等結合」と同じ処理です。キー値はこれらのプライマリ キーでなければなりません。つまり一意でなければならず、Null 値を含むことはできません。

- キーの候補**:**すべての入力データ・ソースから共通するすべてのフィールドが表示されます。このリストから 1 つまたは複数のフィールドを選択して矢印ボタンをクリックすると、レコードを結合するためのキー・フィールドとしてそのフィールドが追加されます。対応するマッピングされたデータベース列を持つマップ・フィールドは、キーとして使用できますが、新しいデータベース列 (名前の後に **<NEW>** と表示) は使用できません。
- 結合キー**:** すべての入力データ・ソースから、キー・フィールドの値に基づいたレコードの結合に使用するすべてのフィールドが表示されます。リストからキーを削除するには、該当するキーを選択して矢印ボタンをクリックし、そのキーを「キーの候補」リストに戻します。複数のキー・フィールドが選択されている場合は、次のオプションが有効になります。
- データベース内に存在するレコードのみを含める**:** 部分結合を実行します。レコードがデータベースとストリーム内にある場合、マップされたフィールドが更新されます。
- レコードをデータベースに追加**:** 外部結合を実行します。データベース内に同じレコードがある場合は、ストリーム内のすべてのレコードが結合され、レコードがデータベース内にない場合は追加されます。

入力データ・フィールドを新しいデータベース列にマッピングするには

1. 「フィールドをマッピング」 下の、左側のリストの入力フィールド名をクリックします。
2. 「追加」 ボタンをクリックして、マッピングを完了します。

入力データ・フィールドを既存のデータベース列にマッピングするには

1. 「フィールドをマッピング」 下の、左側のリストの入力フィールド名をクリックします。
2. 右側の 「データベース列」 の列名をクリックします。
3. 「マップ」 ボタンをクリックして、マッピングを完了します。

マッピングを解除するには

1. 「フィールド」下の右側のリストで、マッピングを削除するフィールドの名前をクリックします。
2. 「マップ解除」ボタンをクリックします。

リストのフィールドの選択を解除するには

Ctrl キーを押したまま、フィールド名をクリックします。

データベース・エクスポートのスキーマのオプション

データベース・エクスポートの「スキーマ」ダイアログ・ボックスでは、データベース (これらのオプションをサポートするデータベース) へのエクスポートのオプションを設定し、フィールドの SQL データ型を設定し、第 1 入力フィールドを指定し、エクスポート時に作成した CREATE TABLE 文をカスタマイズすることができます。

ダイアログ・ボックスは次の部分に分けられます。

- 上部のセクション (表示されている場合) には、これらのオプションをサポートするデータベースへのエクスポートのオプションが表示されます。InfoSphere Warehouse データベースに接続していない場合、このセクションは表示されません。
- 中央部のテキスト・フィールドは、CREATE TABLE コマンドの生成に使用するテンプレートを表示し、デフォルトでは次の形式となっています。

  ```
  CREATE TABLE <table-name> <(table columns)>
  ```

- 下部にあるテーブルを使用して、各フィールドの SQL データ型を指定し、主キーにするフィールドを指定することができます (後述)。このダイアログ・ボックスでは、テーブルの仕様に基づいて自動的に <table-name> パラメーターと <(table columns)> パラメーターが生成されます。

データベース・エクスポート・オプションの設定

このセクションが表示された場合、データベースへのエクスポートのさまざまな設定を指定できます。この機能をサポートするデータベースの種類は次のとおりです。

- SQL Server Enterprise Edition および Developer Edition。詳しくは、トピック 374 ページの『SQL Server のオプション』を参照してください。
- Oracle Enterprise Edition または Personal Edition。詳しくは、トピック 374 ページの『Oracle のオプション』を参照してください。

CREATE TABLE 文のカスタマイズ

このダイアログ・ボックスのテキスト・フィールド部分から、CREATE TABLE 文にデータベース専用オプションを追加することができます。

1. 「**CREATE TABLE** コマンドをカスタマイズ」 のチェック・ボックスを選択し、テキスト・ウィンドウを有効にします。
2. データベース専用オプションを文に追加します。テキストの <table-name> パラメーターと (<table-columns>) パラメーターは IBM SPSS Modeler によって実テーブル名と列の定義に置き換えられるため、必ず保存しておいてください。

SQL データ型の設定

デフォルトでは、IBM SPSS Modeler により、データベース・サーバーで自動的に SQL データ型を割り当てることができます。データ型の自動割り当てを上書きするには、フィールドに対応する行を探して、スキーマ テーブルの「データ型」列にあるドロップダウン・リストから、目的のデータ型を選択します。Shift キーを押しながらクリックすると、複数の行を選択できます。

長さ、精度、または尺度引数を受け取るデータ型の場合 (BINARY、VARBINARY、CHAR、 VARCHAR、NUMERIC、および NUMBER)、データベース・サーバーに自動的に長さを割り当てさせるのではなく、自分で長さを指定する必要があります。例えば、長さに VARCHAR(25) のような意味のある値を指定することにより、IBM

SPSS Modeler 中のストレージ・タイプを確実に上書きすることができます。自動割り当てに優先させるには、「データ型」ドロップダウン・リストで 「指定」 を選択し、データ型定義を適切な SQL タイプ定義文と置換します。

もっとも簡単な方法は、まず目的の定義に似ているデータ型を選択し、次に 「指定」 を選択してその定義を編集することです。例えば、SQL データ型に VARCHAR(25) を設定するには、まず「データ型」ドロップダウン・リストでデータ型を **VARCHAR(length)** に設定してから、次に 「指定」 を選択してテキスト長に 25 を設定します。

プライマリー キー

エクスポートされたテーブルに 1 つ以上の列は、一意の値またはすべての行の値の組合せでなければならず、適用する各フィールドの 「プライマリー キー」 チェック・ボックスをオンにして指示します。ほとんどのデータベースは、無効な第 1 入力フィールド・キー を制限し、自動的に第 1 入力フィールド・キーにインデックスを生成する方法で、テーブルの修正を禁止しています。(オプションで、インデックス・ダイア ログ ボックスでその他のフィールドのインデックスを生成できます)。 詳しくは、トピック 375 ページの『データベース・エクスポートのインデックス・オプション』を参照してください。)

SQL Server のオプション

圧縮を使用: 選択した場合は、圧縮によるエクスポート用のテーブルを作成します。

圧縮レベル: 圧縮のレベルを選択します。

- 行: 行レベルの圧縮を有効にします (例えば、SQL の CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW); と同等)。
- ページ: ページ・レベルの圧縮を有効にします (例えば、SQL の CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);)。

Oracle のオプション

Oracle の設定 - 基本オプション

圧縮を使用: 選択した場合は、圧縮によるエクスポート用のテーブルを作成します。

圧縮レベル: 圧縮のレベルを選択します。

- デフォルト: デフォルトの圧縮を有効にします (例えば SQL の CREATE TABLE MYTABLE(...) COMPRESS;)。この場合、「基本」 オプションと同じ効果があります。
- 基本 : このビューには、予測値間の関係に加え、対象値と最も重要な予測値間の関係を表示するノードのネットワーク グラフが含まれています。基本的な圧縮を有効にします (例えば、SQL の CREATE TABLE MYTABLE(...) COMPRESS BASIC;)。

Oracle の設定 - 高度なオプション

圧縮を使用: 選択した場合は、圧縮によるエクスポート用のテーブルを作成します。

圧縮レベル: 圧縮のレベルを選択します。

- デフォルト: デフォルトの圧縮を有効にします (例えば SQL の CREATE TABLE MYTABLE(...) COMPRESS;)。この場合、「基本」 オプションと同じ効果があります。
- 基本 : このビューには、予測値間の関係に加え、対象値と最も重要な予測値間の関係を表示するノードのネットワーク グラフが含まれています。基本的な圧縮を有効にします (例えば、SQL の CREATE TABLE MYTABLE(...) COMPRESS BASIC;)。

- **OLTP:** OLTP の圧縮を有効にします (例えば、SQL の `CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP;`)。
- クエリー低/高: (Exadata サーバーのみ) クエリーの Hybrid Columnar Compression を有効にします (例えば、SQL の `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW;` や `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH;`)。データ・ウェアハウス環境では QUERY の圧縮が役に立ちます。`HIGH` は、`LOW` より圧縮率が高くなります。
- アーカイブ低/高: (Exadata サーバーのみ) アーカイブの Hybrid Columnar Compression を有効にします (例えば、SQL の `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW;` や `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH;`)。長期間格納されるデータの圧縮には ARCHIVE の圧縮が役に立ちます。`HIGH` は、`LOW` より圧縮率が高くなります。

データベース・エクスポートのインデックス・オプション

インデックス・ダイアログ・ボックスでは、IBM SPSS Modeler からエクスポートされたデータベース・テーブルにインデックスを作成することができます。フィールド・セットに含めるフィールドを指定し、必要に応じて `CREATE INDEX` コマンドをカスタマイズします。

ダイアログ・ボックスには次の 2 つの部分があります。

- 最上部のテキスト・フィールドは、`CREATE INDEX` コマンドの生成に使用するテンプレートを表示し、デフォルトは次の形式となっています。

  ```
  CREATE INDEX <index-name> ON <table-name>
  ```

- ダイアログ・ボックスの下部にあるテーブルでは、作成する各インデックスの仕様を追加することができます。各インデックスに対しては、インデックス名とフィールドまたは含める列を指定します。ダイアログ・ボックスにより、適切な `<index-name>` パラメーターと `<table-name>` パラメーターが自動的に生成されます。

  例えば、フィールド *empid* と *deptid* に関する単一のインデックスに対して生成された SQL はこのようになります。

  ```
  CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)
  ```

  複数のインデックスを生成するために複数の行を追加することができます。独立した `CREATE INDEX` コマンドが各行に生成されます。

CREATE INDEX コマンドのカスタマイズ

オプションで、 `CREATE INDEX` コマンドをすべてのインデックスに対して、または特定のインデックスのみに対してカスタマイズできます。これによって、必要に応じて自由に、固有のデータベースの要件またはオプションを統合し、すべてのインデックスまたは特定のインデックスのみに対してカスタマイズの結果を適用することができるようになります。

- ダイアログ・ボックスの最上部にある 「**CREATE INDEX** コマンドをカスタマイズ」 を選択し、これ以降追加されるすべてのインデックスに使用するテンプレートを修正します。変更結果は、既にテーブルに追加されているインデックスに対して自動的に適用されます。
- テーブルで 1 つ以上の行を選択し、ダイアログ・ボックスの最上部にある 「選択したインデックスを更新」 をクリックして現在のカスタマイズした結果を選択したすべての行に適用します。
- 各行の 「カスタマイズ」 チェック・ボックスをオンにして、そのインデックスのみに対してコマンド・テンプレートを修正します。

<index-name> パラメーターと <table-name> パラメーターの値は、テーブルの仕様に基づいてダイアログ・ボックスによって自動的に生成されます。これらの値を直接編集することはできません。

BITMAP KEYWORD : Oracle のデータベースを使用中に、テンプレートをカスタマイズして、標準のインデックスではなく、次のようにビットマップ・インデックスを生成することができます。

```
CREATE BITMAP INDEX <index-name> ON <table-name>
```

ビットマップ・インデックスは、少数の重複レコード値を持つ列のインデックス生成に役立ちます。生成された SQL はこのようになります。

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

UNIQUE キーワード**:** ほとんどのデータベースは、CREATE INDEX コマンドの UNIQUE キーワードをサポートしています。これは、基本的なテーブルにおける第 1 入力フィールド・キーの制約に類似した、一意性の制約を強制します。

```
CREATE UNIQUE INDEX <index-name> ON <table-name>
```

第 1 入力フィールド・キーとして実際に指定されたフィールドに対して、この指定は必要でないことに注意してください。ほとんどのデータベースは、第 1 入力フィールド・キーとして指定されたフィールドのインデックスを、CREATE TABLE コマンド内に自動的に生成します。したがって、これらのフィールドで明示的に生成するインデックスは必要ありません。 詳しくは、トピック 373 ページの『データベース・エクスポートのスキーマのオプション』を参照してください。

FILLFACTOR キーワード**:** 一部のインデックス用の物理パラメーターを精密に調整することができます。例えば SQL Server では、今後テーブルを変更する場合のメンテナンス・コストとのトレードオフを考慮して、(初期作成後に) インデックス・サイズを選択することができます。

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

その他のコマンド

- 指定された名前を持つ既存のインデックスがある場合、インデックスの生成は失敗します。失敗は警告として処理され、それ以降のインデックスの生成を許可し、それから、すべてのインデックスを試みた後にメッセージ・ログ内にエラーとして再報告されます。
- 最高のパフォーマンスを得るためには、データがテーブルにロードされた後にインデックスを生成すべきです。インデックスには最低 1 行は組まれていなければなりません。
- ノードを実行する前に、メッセージ・ログ内の生成された SQL をプレビューすることができます。
- データベースに書き込まれる一時テーブル (つまり、ノード・キャッシュが有効なとき) に対しては、第 1 入力フィールド・キーを指定するオプションは利用できません。ただし、データが下流ノードで使用される頻度に合わせて、システムが一時テーブルにインデックスを生成します。例えば、キャッシュされたデータがその後 *DEPT* 列ごとに結合された場合、キャッシュされているこの列のテーブルをインデックスする意味があります。

インデックスおよびクエリーの最適化

一部のデータベースの管理システムでは、データベースのテーブルの生成、ロード、インデックスの作成が行われた後に、オプティマイザーがインデックスを利用して新規のテーブル上でのクエリーの高速化ができるようになるまで、あるステップが必要になります。例えば Oracle では、コストベースのクエリー・オプ

ティマイザーが、テーブルのインデックスをクエリーの最適化に使用できるようにする前に、テーブルの解析を要求します。Oracle の ODBC プロパティー・ファイル (ユーザーには見えません) には、これを実行する次のようなオプションが含まれます。

```
# Defines SQL to be executed after a table and any associated indexes
# have been created and populated
table_analysis_sql, 'ANALYZE TABLE <table-name> COMPUTE STATISTICS'
```

このステップは、Oracle 内にテーブルが生成されるときに必ず実行されます (第 1 入力フィールド・キーまたはインデックスが定義済みか否かにかかわらず)。必要に応じて、その他のデータベース用の ODBC プロパティー・ファイルを同様にカスタマイズすることができます。サポートにお問い合わせください。

データベース・エクスポートの拡張オプション

データベース・エクスポート・ノードのダイアログ・ボックスで「拡張」ボタンをクリックすると、結果をデータベースにエクスポートするための技術的な詳細を設定できる新しいダイアログ・ボックスが表示されます。

バッチ コミットを使用。データベースへの行ごとのコミットを無効にする場合に選択します。

バッチ・サイズ。メモリーにコミットする前にデータベースに送信するレコード数を指定します。 この値を減らすと、データの整合性は向上しますが、転送速度が遅くなります。この値を調整して、データベースのパフォーマンスを最適な状態にすることができます。

バルク・ロードを使用。IBM SPSS Modeler からデータベースに直接データをバルク・ロードするための方法を選択します。どのバルク・ロードが特定のシナリオに適当かを選択するには、実験が必要です。

- **ODBC** 経由。通常のデータベースへのエクスポートと比べて大幅に効率的な、ODBC API を使った複数行挿入を実施する場合に選択します。下にあるオプションから、行方向または列方向のバインドを選択してください。
- 外部ローダー経由。データベース固有のカスタム バルク・ローダー・プログラムを使用する場合に選択します。このオプションを選択すると、下にある「外部ローダー・オプション」が有効になります。AIX で UTF-8 データ・エンコードの問題が発生する場合は、`options.cfg` に `locale, en_US.UDF-8` を追加する必要がある場合があります。

詳細 **ODBC** オプション。これらのオプションは、「**ODBC** 経由」 を選択した場合にだけ利用できます。この機能はすべての ODBC ドライバーに対応しているわけではないことに注意してください。

- 行方向。データベースへのデータのロードに、`SQLBulkOperations` コールを使用する場合に選択します。一般的に行方向のバインドは、データをレコード単位に挿入するパラメーター化された INSERT 文を使用する場合に比べて、速度を向上することができます。
- 列方向。データベースへのデータのロードに、列方向のバインドを使用する場合に選択します。列方向のバインドは、各データベースの列 (パラメーター化された `INSERT` 文中の) を *N* 値の配列にバインドすることにより、パフォーマンスを向上します。`INSERT` を 1 回実行すると、データベースに *N* 行が挿入されます。この方法により、大幅にパフォーマンスを向上することができます。

外部ローダー・オプション。「外部ローダー経由」を指定している場合、データ・セットをファイルにエクスポートしたり、そのファイルからデータをデータベースにロードするためのカスタム・ローダー・プログラムの指定と実行を行ったりするための各種オプションが表示されます。IBM SPSS Modeler は多くの一般的なデータベース・システムの外部ローダーと連携できます。このソフトウェアには、技術資料とともにいくつかのスクリプトが付属しています (`scripts` サブディレクトリーに格納されています)。この機能を使用するには、Python 2.7 を IBM SPSS Modeler または IBM SPSS Modeler Server と同じマシンにイ

ンストールし、python_exe_path パラメーターを options.cfg ファイルに設定する必要があります。 詳しくは、トピック『バルク・ローダーのプログラミング』を参照してください。

- 区切り文字を使用。エクスポートするファイルで使用する区切り文字を指定します。タブ文字で区切る場合は 「タブ」 を、スペースで区切る場合は 「スペース」 を選択します。カンマ (,) などの他の文字を指定する場合は、「その他」 を選択します。
- データ ファイルを指定。バルク・ロード時に書き込まれる、データ・ファイルのパスを入力する場合に選択します。デフォルトでは、サーバーの temp ディレクトリーに一時ファイルが作成されます。
- ローダー プログラムを指定。バルク・ローダー・プログラムを指定する場合に選択します。デフォルトでは、IBM SPSS Modeler インストール・ディレクトリー中の /scripts サブディレクトリーから、データベースに対して実行される Python スクリプトが検索されます。このソフトウェアには、技術資料とともにいくつかのスクリプトが付属しています (scripts サブディレクトリーに格納されています)。
- ログの生成。指定したディレクトリーにログを生成する場合に選択します。ログ・ファイルには、バルク・ロード操作が失敗した場合などに役立つエラー情報が含まれています。
- テーブル サイズの検査。IBM SPSS Modeler からエクスポートされる行数に対応してテーブル・サイズを増やすために、テーブル検査を実施する場合に選択します。
- ローダーの付加オプション。ローダー・プログラムに対する引数を指定します。スペースを含む引数には二重引用符を使用してください。

円記号を前に付けることにより、オプションの引数中に二重引用符を指定することができます。例えば、-comment "This is a ¥"comment¥"" と指定されたオプションには、-comment フラグと、This is a "comment" として表示されるコメント自体が含まれています。

円記号を 1 つ指定するには、もう 1 つの円記号を付けて指定します。例えば、-specialdir "C:¥¥Test Scripts¥¥" と指定されたオプションには、-specialdir フラグと、C:¥Test Scripts¥ として表示されるディレクトリーが含まれています。

バルク・ローダーのプログラミング

データ・エクスポート・ノードは、「詳細オプション」ダイアログ・ボックスにバルク・ロードのオプションがあります。バルク・ローダー・プログラムを使用して、データをテキストからデータベースにロードすることができます。

オプション 「**Use bulk loading - via external loader**」は IBM SPSS Modeler を次の 3 つに構成します。

- 要求されたすべてのデータベース・テーブルを生成する。
- データベースをテキスト・ファイルにエクスポートする。
- バルク・ローダー・プログラムを起動し、このファイルからデータをデータベースにロードする。

一般に、バルク・ローダー・プログラムはデータベース・ロード ユーティリティーそのもの (例えば、Oracle の sqlldr ユーティリティー) ではありませんが、正しい引数を形成し、データベース固有の補助ファイル (コントロール・ファイルなど) を生成し、次にデータベース・ロード ユーティリティーを起動する小型のスクリプトまたはプログラムです。以下のセクションの情報は、既存のバルク・ローダーを編集できます。

またはバルク・ロードの独自のプログラムを作成することもできます。 詳しくは、トピック 383 ページの『バルク・ローダー・プログラムの開発』を参照してください。ただし、これは標準のテクニカル サポート契約の範囲外であり、IBM サービス担当員に連絡して支援を受ける必要があることに注意してください。

バルク・ロードのスクリプト

IBM SPSS Modeler には、Python スクリプトを使用して実装されるさまざまなデータベースのさまざまなバルク・ローダーが用意されています。「外部ローダー経由」 オプションを選択してデータベース・エクスポート・ノードを含むストリームを実行すると、IBM SPSS Modeler は ODBC 経由でデータベース・テーブルを (必要に応じて) 作成し、IBM SPSS Modeler Server を実行するホストの一時ファイルにデータをエクスポートし、バルク・ロード スクリプトを起動します。このスクリプトは DBMS ベンダーが提供するユーティリティーを実行し、一時ファイルのデータをデータベースにアップロードします。

注 : IBM SPSS Modeler のインストールには Python ランタイム インタープリターが含まれていないため、Python を別途インストールする必要があります。 詳しくは、トピック 377 ページの『データベース・エクスポートの拡張オプション』を参照してください。

以下の表に示すデータベースのスクリプトが用意されています (IBM SPSS Modeler インストール・ディレクトリーの *¥scripts* フォルダー)。

表 47. 提供されるバルク・ロード スクリプト :

データベース	スクリプト名	詳細情報
IBM Db2	db2_loader.py	詳しくは、トピック『IBM Db2 データベースへのデータのバルク・ロード』を参照してください。
IBM Netezza	netezza_loader.py	詳しくは、トピック 380 ページの『データの IBM Netezza データベースへのバルク・ロード』を参照してください。
Oracle	oracle_loader.py	詳しくは、トピック 381 ページの『データの Oracle データベースへのバルク・ロード』を参照してください。
SQL Server	mssql_loader.py	詳しくは、トピック 381 ページの『データの SQL Server データベースへのバルク・ロード』を参照してください。
Teradata	teradata_loader.py	詳しくは、トピック 382 ページの『データの Teradata データベースへのバルク・ロード』を参照してください。

IBM Db2 データベースへのデータのバルク・ロード

「DB エクスポートの詳細オプション」ダイアログ・ボックスの「外部ローダー・オプション」を使用して、IBM SPSS Modeler から IBM Db2 データベースへのバルク・ロードを設定するには、以下のポイントに留意してください。

Db2 コマンド・ライン・プロセッサー (CLP) ユーティリティーをインストールする

スクリプト db2_loader.py は Db2 LOAD コマンドを起動します。コマンド・ライン・プロセッサー (UNIX の場合は db2、Windows の場合は db2cmd) を db2_loader.py が実行されるサーバー (通常、IBM SPSS Modeler Server を実行するホスト) にインストールします。

ローカル・データベースのエイリアス名が実際のデータベース名と同じかどうかを確認する

Db2 ローカル・データベースのエイリアスは、ローカルまたはリモートの Db2 インスタンスのデータベースを参照するために Db2 クライアント・ソフトウェアによって使用される名前です。ローカル・データベースのエイリアスがリモート・データベースの名前と異なる場合、追加のローダー・オプションを指定します。

```
-alias <local_database_alias>
```

例えば、ホスト GALAXY のリモート・データベースの名前が STARS であっても、IBM SPSS Modeler Server を実行するホストの Db2 ローカル・データベースのエイリアスは STARS_GALAXY の場合があります。追加のローダー・オプションを使用

```
-alias STARS_GALAXY
```

ASCII 文字以外のデータ・エンコード

ASCII 形式でないデータをバルク・ロードしている場合、db2_loader.py の設定セクションのコードページ変数が正しく設定されている必要があります。

空白の文字列

空白の文字列は NULL 値としてデータベースにエクスポートされます。

データの IBM Netezza データベースへのバルク・ロード

以下のポイントによって、IBM SPSS Modeler からのバルク・ロードをデータベース・エクスポートの「詳細オプション」ダイアログ・ボックスの「外部ローダー」オプションを使用する IBM Netezza データベースに設定することができます。

Netezza nzload ユーティリティーをインストールする

スクリプト *netezza_loader.py* が Netezza ユーティリティー *nzload* を起動します。*nzload* を、*netezza_loader.py* を実行するサーバーにインストールして正しく設定します。

非 ASCII データのエクスポート

エクスポートに ASCII 形式ではないデータが含まれている場合は、「DB エクスポートの詳細オプション」ダイアログ・ボックスの「ローダーの付加オプション」フィールドに -encoding UTF8 を追加しなければならないことがあります。ASCII 以外のデータは、正しくアップロードする必要があります。

日付、時刻、タイムスタンプ形式のデータ

ストリームのプロパティーで、日付の形式を 「**DD-MM-YYYY**」 に、時刻の形式を 「**HH:MM:SS**」 に設定します。

空白の文字列

空白の文字列は NULL 値としてデータベースにエクスポートされます。

データを既存のテーブルに挿入する場合のストリームおよび対象テーブルの列の異なる順序

ストリームの列の順序が対象テーブルの列の順序と異なる場合、日付の値が誤った列に挿入されます。フィールドの順序変更ノードを使用して、ストリームの列の順序が対象テーブルの順序と一致するようにします。 詳しくは、トピック 193 ページの『フィールド順序ノード』を参照してください。

nzload の進捗状況の追跡

IBM SPSS Modeler をローカル・モードで実行している場合、-sts を「DB エクスポートの詳細オプション」ダイアログ・ボックスの 「ローダーの付加オプション」 フィールドに追加して、*nzload* ユーティリティーで開いたコマンド ウィンドウに 1000 行ごとのステータス メッセージが表示されるようにします。

データの Oracle データベースへのバルク・ロード

以下のポイントによって、IBM SPSS Modeler からのバルク・ロードをデータベース・エクスポートの「詳細オプション」ダイアログ・ボックスの「外部ローダー」オプションを使用する Oracle データベースに設定することができます。

Oracle sqlldr ユーティリティーをインストールする

スクリプト *oracle_loader.py* が Oracle ユーティリティー *sqlldr* を起動します。*sqlldr* は Oracle Client に自動的に含まれません。*sqlldr* は、*oracle_loader.py* が実行されるサーバーにインストールされます。

データベース SID またはサービス名を指定します

データをローカル以外の Oracle サーバーにエクスポートしている場合、またはローカルの Oracle サーバーに複数のデータベースがある場合、「DB エクスポートの詳細オプション」ダイアログ・ボックスの「ローダーの付加オプション」 フィールドで以下を指定し、SID またはサービス名で渡す必要があります。

```
-database <SID>
```

oracle_loader.py の設定セクションの編集

UNIX (およびオプションで、Windows) システムで、*oracle_loader.py* スクリプトの始めの設定セクションを編集します。ここで、ORACLE_SID、NLS_LANG、TNS_ADMIN および ORACLE_HOME 環境変数の値を、*sqlldr* ユーティリティーのフル・パスとともに必要に応じて指定できます。

日付、時刻、タイムスタンプ形式のデータ

ストリームのプロパティーで、通常は日付の形式を 「**YYYY-MM-DD**」 に、時刻の形式を「**HH:MM:SS**」 に設定する必要があります。

上記と異なる日付および時刻の形式を使用する必要がある場合、Oracle のマニュアルを参照し、*oracle_loader.py* スクリプト・ファイルを編集してください。

ASCII 文字以外のデータ・エンコード

ASCII 形式でないデータをバルク・ロードしている場合、環境変数 NLS_LANG が正しく設定されている必要があります。Oracle ローダー・ユーティリティー *sqlldr* によって読み込まれます。例えば、Windows の Shift-JIS の NLS_LANG に対する正しい値は Japanese_Japan.JA16SJIS です。NLS_LANG の詳細については、Oracle のマニュアルを参照してください。

空白の文字列

空白の文字列は NULL 値としてデータベースにエクスポートされます。

データの SQL Server データベースへのバルク・ロード

以下のポイントによって、IBM SPSS Modeler からのバルク・ロードをデータベース・エクスポートの「詳細オプション」ダイアログ・ボックスの「外部ローダー」オプションを使用する SQL Server データベースに設定することができます。

SQL Server bcp.exe ユーティリティーをインストールする

スクリプト *mssql_loader.py* が SQL Server ユーティリティー *bcp.exe* を起動します。*bcp.exe* は、*mssql_loader.py* が実行されるサーバーにインストールされます。

区切り文字としてのスペースの使用は無効
: 「DB エクスポートの詳細オプション」ダイアログボックスでは、スペースを区切り文字として使用しないでください。

「テーブル サイズの検査」オプションの推奨
: 「DB エクスポートの詳細オプション」ダイアログボックスで 「テーブル サイズの検査」 オプションを有効にすることをお勧めします。バルク・ロード・プロセスの障害は常に検知されるわけではないため、このオプションを有効にして追加のチェックを行い、適切な数の行が読み込まれていることを確認します。

空白の文字列
: 空白の文字列は NULL 値としてデータベースにエクスポートされます。

完全修飾 **SQL** サーバーの名前付きインスタンスの指定
: 修飾されていないホスト名が原因で SPSS Modeler が SQL Server にアクセスできず、以下のエラーが表示される場合があります。

  外部バルク ローダーの実行中にエラーが発生しました。ログ ファイルに詳細が記録されています。

  このエラーを修正するには、二重引用符も含めて以下の文字列を「ローダーの付加オプション」フィールドに追加します。

  ```
  "-S mhreboot.spss.com¥SQLEXPRESS"
  ```

データの Teradata データベースへのバルク・ロード

以下のポイントによって、IBM SPSS Modeler からのバルク・ロードをデータベース・エクスポートの「詳細オプション」ダイアログ・ボックスの「外部ローダー」オプションを使用する Teradata データベースに設定することができます。

Teradata fastload ユーティリティーをインストールする

スクリプト *teradata_loader.py* が Teradata ユーティリティー *fastload* を起動します。*fastload* を、*teradata_loader.py* を実行するサーバーにインストールして正しく設定します。

データのバルク・ロードは空のテーブルにのみ可能

バルク・ロードのターゲットとしては空のテーブルだけが使用できます。バルク・ロード前の対象テーブルにデータが含まれている場合、操作は失敗します。

日付、時刻、タイムスタンプ形式のデータ

ストリームのプロパティーで、日付の形式を 「**YYYY-MM-DD**」 に、時刻の形式を 「**HH:MM:SS**」 に設定します。

空白の文字列

空白の文字列は NULL 値としてデータベースにエクスポートされます。

Teradata プロセス ID (tdpid)

デフォルトで、*fastload* は `tdpid=dbc` によってデータを Teradata システムにエクスポートします。通常、`dbccop1` が Teradata サーバーの IP アドレスと関連する HOSTS ファイルにエントリーがあります。異なるサーバーを使用する場合、「DB エクスポートの詳細オプション」ダイアログ・ボックスの「ローダーの付加オプション」 フィールドで以下を指定し、このサーバーの `tdpid` を渡します。

```
-tdpid <id>
```

テーブル名および列名のスペース

テーブル名または列名にスペースが含まれる場合、バルク・ロード操作は失敗します。可能な場合は、テーブル名または列名の名前を変更してスペースを削除してください。

バルク・ローダー・プログラムの開発

このトピックでは、IBM SPSS Modeler から実行して、テキスト・ファイルのデータをデータベースに読み込むバルク・ローダー・プログラムの開発方法について説明します。 ただし、これは標準のテクニカルサポート契約の範囲外であり、IBM サービス担当員に連絡して支援を受ける必要があることに注意してください。

Python を使用したバルク・ローダー・プログラムの作成

デフォルトでは、IBM SPSS Modeler は、データベースのタイプに基づいてデフォルトのバルク・ローダー・プログラムを検索します。 379 ページの表 47を参照してください。

スクリプト *test_loader.py* を使用して、バッチ ローダー・プログラムの開発を支援します。 詳しくは、トピック 385 ページの『バルク・ローダー・プログラムのテスト』を参照してください。

バルク・ローダー・プログラムに渡されるオブジェクト

IBM SPSS Modeler は、バルク・ローダー・プログラムに渡される 2 つのファイルを作成します。

- データ・ファイル: 読み込まれるデータ (テキスト形式) が含まれます。
- スキーマ ファイル: 列の名前と種類を説明する XML ファイルで、データ・ファイルをどのように書式設定するかに関する情報が提供されます (フィールド間の区切り文字として使用される文字など)。

また、IBM SPSS Modeler は、バルク・ロード・プログラム起動時に、テーブル名、ユーザー名、パスワード名など他の情報を引数として渡します。

注 : IBM SPSS Modeler に正常に完了したことを通知するために、バルク・ローダー・プログラムはスキーマ ファイルを削除する必要があります。

バルク・ローダー・プログラムに渡される引数

プログラムに渡される引数を以下の表に示します。

表 *48.* バルク・ローダーに渡される引数 :

引数	説明
schemafile	スキーマ ファイルのパス。
data file	データ・ファイルのパス。
servername	DBMS サーバーの名前。空白の場合あり。
databasename	DBMS サーバー内のデータベースの名前。空白の場合あり。
username	データベースにログインするユーザー名。
password	データベースにログインするパスワード。
tablename	読み込むテーブルの名前。
ownername	テーブル所有者の名前 (スキーマ名)。
logfilename	ログファイルの名前 (空白の場合、ログ・ファイルは作成されません)。
rowcount	データ・セット内の行数。

「DB エクスポートの詳細オプション」ダイアログボックスの 「ローダーの付加オプション」 フィールドで指定されたオプションは、これらの標準引数の後でバルク・ローダー・プログラムに渡されます。

データ・ファイルの形式

データはテキスト形式でデータ・ファイルに書き込まれ、各フィールドは「DB エクスポートの詳細オプション」ダイアログボックスで指定された区切り文字で区切られます。以下に、タブ区切りのデータ・ファイルがどのように表示されるかの例を示します。

```
48  F  HIGH    NORMAL  0.692623  0.055369  drugA
15  M  NORMAL  HIGH    0.678247  0.040851  drugY
37  M  HIGH    NORMAL  0.538192  0.069780  drugA
35  F  HIGH    HIGH    0.635680  0.068481  drugA
```

ファイルは、IBM SPSS Modeler Server で使用されるローカル エンコードで作成されます (IBM SPSS Modeler Server に接続していない場合は IBM SPSS Modeler)。いくつかの形式は、IBM SPSS Modeler ストリーム設定で制御されます。

スキーマ ファイルの形式

スキーマ ファイルは、データ・ファイルを記述する XML ファイルです。以下に、先行するデータ・ファイルに伴う例を示します。

```
<?xml version="1.0" encoding="UTF-8"?>
<DBSCHEMA version="1.0">
 <table delimiter="¥t" commit_every="10000" date_format="YYYY-MM-DD" time_format="HH:MM:SS"
append_existing="false" delete_datafile="false">
  <column name="Age" encoded_name="416765" type="integer"/>
  <column name="Sex" encoded_name="536578" type="char" size="1"/>
  <column name="BP" encoded_name="4250" type="char" size="6"/>
  <column name="Cholesterol" encoded_name="43686F6C65737465726F6C" type="char" size="6"/>
  <column name="Na" encoded_name="4E61" type="real"/>
  <column name="K" encoded_name="4B" type="real"/>
  <column name="Drug" encoded_name="44727567" type="char" size="5"/>
 </table>
</DBSCHEMA>
```

以下の 2 つの表に、スキーマ・ファイルの <table> 要素と <column> 要素の属性を示します。

表 *49. <table>* 要素の属性：

属性	説明
delimiter	フィールド区切り文字 (TAB は ¥t として示されます)。
commit_every	バッチ サイズの間隔 (「DB エクスポートの詳細オプション」ダイアログボックスで指定されているとおり)
date_format	日付の表示に使用する形式。
time_format	時刻の表示に使用する形式。
append_existing	ロードするテーブルにデータが含まれている場合は true、含まれていない場合は false。
delete_datafile	ロードの完了時にバルク・ロード・プログラムでデータ・ファイルを削除する場合は true。

表 50. <column> 要素の属性：

属性	説明
name	列名。
encoded_name	データ・ファイルと出力と同じエンコードに 2 桁の 16 進数で変換される列の名前。
type	列のデータ型。integer、real、char、time、date、および datetime のいずれか。
size	char データ型の場合、列の最大幅の文字数。

バルク・ローダー・プログラムのテスト

IBM SPSS Modeler インストール・ディレクトリーの *¥scripts* フォルダーに格納されているテスト・スクリプト *test_loader.py* を使用して、バルク・ロードのテストを実行することができます。このテストは、IBM SPSS Modeler で使用するバルク・ロード・プログラムまたはスクリプトの開発、デバッグ、トラブルシューティングを行う場合に役立ちます。

テスト・スクリプトを使用するには、以下の手順に従います。

1. *test_loader.py* スクリプトを実行して、スキーマ・ファイルとデータ・ファイルを *schema.xml* ファイルと *data.txt* ファイルにコピーし、Windows バッチ・ファイル (*test.bat*) を作成します。
2. *test.bat* ファイルを編集して、テストするバルク・ローダー・プログラムまたはスクリプトを選択します。
3. コマンド・シェルから *test.bat* ファイルを実行して、バルク・ロード・プログラムまたはスクリプトをテストします。

注 : *test.bat* を実行してもデータは実際にデータベースには読み込まれません。

ファイル エクスポート・ノード

フラット・ファイル・エクスポート・ノードを使用すると、区切り文字で区切られた形式のテキスト・ファイルにデータを書き込むことができます。他の分析ソフトや表計算ソフトに読み込める形式でデータをエクスポートする場合に役に立ちます。

データに地理空間情報が含まれる場合は、フラット ファイルとしてエクスポートできます。また、同じストリームの中で使用する可変長ファイル・ソース・ノードを生成した場合、すべてのストレージ、測定、および地理空間メタデータが新しいソース・ノードで保持されます。しかし、データをエクスポートしてから別のストリームにインポートする場合は、追加のステップを実行して、新しいソース・ノードで地理空間メタデータを設定する必要があります。詳しくは、 28 ページの『可変長ファイル・ノード』のトピックを参照してください。

注**:** IBM SPSS Modeler がキャッシュ・ファイルに古いキャッシュ形式を使用しないため、古いキャッシュ形式のファイルを書き込むことはできません。IBM SPSS Modeler のキャッシュ・ファイルは IBM SPSS Statistics *.sav* 形式で保存され、Statistics エクスポート・ノードを使用して書き込みできます。 詳しくは、 386 ページの『Statistics エクスポート・ノード』のトピックを参照してください。

ファイル・ノードの「エクスポート」タブ

エクスポート・ファイル**:** ファイルの名前を指定します。ファイル名を入力するか、またはファイル選択ボタンをクリックしてファイルの場所を指定します。

書き込みモード: 「上書き」を選択すると、指定されたファイル内の既存のデータが上書きされます。「レコード追加」 を選択すると、出力データが既存ファイルの末尾に追加され、既存データはそのまま保存されます。

- フィールド名を含める: このオプションを選択すると、出力ファイルの 1 行目にフィールド名が書き込まれます。このオプションは、保存モードで 「上書き」 を選択した場合にだけ利用できます。

各レコードの後に改行を入れる: このオプションを選択すると、各レコードが出力ファイル中の新しい行に書き込まれます。

フィールド区切り文字: 生成するテキスト・ファイルで、フィールド値の間に挿入する文字列を選択します。「カンマ」、「タブ」、「スペース」、または 「その他」 を選択することができます。「その他」を選択した場合は、テキスト・ボックスに適切な区切り文字を入力してください。

シンボル値の引用符: シンボル値フィールドの値に対して使用する引用符の種類を選択します。「なし」(値に引用符を付けない)、「単一 **(')**」、「二重 **(")**」、または 「その他」 を選択できます。「その他」を選択した場合は、テキスト・ボックスに適切な引用文字を入力してください。

エンコード: 使用するテキストのエンコード方法を指定します。サーバー・デフォルト、システム・デフォルト、UTF-8 から選択できます。

- システム・デフォルトは、Windows のコントロール・パネル (分散モードで実行している場合はサーバー・コンピューター) で指定できます。
- デフォルトは、「ストリーム・プロパティー」ダイアログ・ボックスで指定されます。

小数点記号: データ内で小数点記号をどのように表すかを指定します。

- ストリームのデフォルト: 現在のストリームのデフォルト設定で定義された小数点区切り文字が使用されます。これは、通常、コンピューターのロケールの設定で定義された小数点区切り文字になります。
- ピリオド **(.)**: 小数点区切り文字として、ピリオドを使用します。
- カンマ **(,)**: 小数点区切り文字として、カンマを使用します。

データのインポート ノードを生成: エクスポートされたデータ・ファイルを読み込む可変長ファイル・ソース・ノードを自動的に生成するには、このオプションを選択します。詳しくは、トピック 28 ページの『可変長ファイル・ノード』を参照してください。

Statistics エクスポート・ノード

Statistics エクスポート・ノードを使用すると、IBM SPSS Statistics の *.sav* 形式でデータをエクスポートすることができます。IBM SPSS Statistics *.sav* ファイルは、IBM SPSS Statistics Base およびその他のモジュールで読み込むことができます。この形式は、IBM SPSS Modeler キャッシュ・ファイルでも使用されます。

IBM SPSS Statistics の変数名は 64 文字までに制限されており、スペース、ドル記号 ($)、ダッシュ (-) など一部の文字を使用できないため、IBM SPSS Modeler のフィールド名を IBM SPSS Statistics の変数名にマップするとエラーが発生することがあります。この問題に対処するには、次の 2 通りの方法があります。

- 「フィルター」タブをクリックして、IBM SPSS Statistics 変数名の要件に準拠したフィールド名に変更することができます。 詳しくは、トピック 388 ページの『IBM SPSS Statistics 用のフィールドの名前変更またはフィルタリング』を参照してください。
- IBM SPSS Modeler でフィールド名とラベルをエクスポートします。

注: IBM SPSS Modeler は、*.sav* ファイルを Unicode の UTF-8 形式で書き込みます。リリース 16.0 以降の IBM SPSS Statistics でサポートしているのは Unicode UTF-8 形式だけです。データの破損の可能性を回避するために、Unicode エンコードで保存した *.sav* ファイルはリリース 16.0 以前の IBM SPSS Statistics で使用することはできません。詳細は、IBM SPSS Statistics のヘルプを参照してください。

複数回答設定: ストリームに定義された複数の回答セットは、ファイルがエクスポートされると自動的に保存されます。「フィルター」タブで、ノードの複数の回答セットを表示および編集することができます。詳しくは、トピック 161 ページの『複数回答セット編集』を参照してください。

Statistics エクスポート・ノードの「エクスポート」タブ

エクスポート・ファイル: ファイルの名前を指定します。ファイル名を入力するか、またはファイル選択ボタンをクリックしてファイルの場所を指定します。

ファイルの種類: ファイルを通常の *.sav* または圧縮の *.zsav* のいずれの形式で保存するかを選択します。

パスワードでファイルを暗号化: パスワードを使用してファイルを保護するには、このボックスを選択します。別のダイアログ・ボックスで「パスワード」の入力および確認を要求するプロンプトが出されます。

注: パスワード保護されたファイルを開くことができるのは、SPSS Modeler バージョン 16 以降、または SPSS Statistics バージョン 21 以降のみです。

フィールド名のエクスポート: SPSS Modeler から SPSS Statistics の *.sav* または *.zsav* ファイルにエクスポートする場合の変数名とラベルの処理方法を指定します。

- 「名前と変数ラベル」 SPSS Modeler のフィールド名とフィールド・ラベルの両方をエクスポートする場合に選択します。名前は SPSS Statistics の変数名としてエクスポートされ、ラベルは SPSS Statistics の変数ラベルとしてエクスポートされます。
- 「変数ラベルとして使用」 SPSS Modeler のフィールド名を SPSS Statistics の変数ラベルとして使用する場合に選択します。SPSS Modeler は SPSS Statistics 変数名では無効であるフィールド名の文字を使用できます。無効な SPSS Statistics 名を作成しないように、「変数ラベルとして使用」 を選択するか、フィールド名を調整するための「フィルター」タブを利用します。

アプリケーションの起動: SPSS Statistics がコンピューターにインストールされている場合、このオプションを選択することにより、保存したデータ・ファイルに対してこのアプリケーションを直接起動できます。アプリケーションを起動するためのオプションは、「ヘルパー アプリケーション」ダイアログ・ボックスで指定する必要があります。詳しくは、トピック 366 ページの『IBM SPSS Statistics ヘルパー アプリケーション』を参照してください。外部プログラムで起動しないで、単に SPSS Statistics *.sav* または *.zsav* ファイルを作成する場合は、このオプションの選択を解除してください。

注: SPSS Modeler と SPSS Statistics をサーバー (分散) モードで一緒に実行すると、データを書き込んで SPSS Statistics セッションを起動しても、SPSS Statistics クライアントが自動的に開かないため、アクティブ・データ・セットに読み込まれるデータ・セットが表示されません。この問題を回避するには、SPSS Statistics クライアントを起動して、手動でデータ・ファイルを開きます。

データのインポート ノードを生成: エクスポートされたデータ・ファイルを読み込む、Statistics ファイル・ソース・ノードを自動的に生成する場合に選択します。詳しくは、トピック 34 ページの『Statistics ファイル・ノード』を参照してください。

IBM SPSS Statistics 用のフィールドの名前変更またはフィルタリング

IBM SPSS Modeler から IBM SPSS Statistics などの外部アプリケーションにデータをエクスポートまたは展開する前に、フィールド名を変更したり、調節しなければならないこともあります。Statistics 変換、Statistics 出力、および Statistics エクスポート ダイアログ・ボックスには、「フィルター」タブが用意されており、ここからこの処理を簡単に行うことができます。

「フィルター」タブの機能に関する基本的な説明は、他の場所で説明されています。詳しくは、 159 ページの『フィルタリング・オプションの設定』を参照してください。ここでは、IBM SPSS Statistics にデータを読み込む場合のヒントについて説明します。

IBM SPSS Statistics の命名規則に準ずるよう、ファイル名を調整する手順は次のとおりです。

1. 「フィルター」タブで、「フィルター・オプション・メニュー」ツールバー・ボタン (ツールバーの最初のボタン) をクリックします。
2. 「IBM SPSS Statistics 用に名前変更」を選択します。
3. 「IBM SPSS Statistics 用に名前変更」ダイアログで、ファイル名の無効な文字を「ハッシュ **(#)**」文字または「下線 **(_)**」のいずれに置き換えるかを選択することができます。

複数回答セットの名前を変更: Statistics ファイル入力ノードを使用して、IBM SPSS Modeler にインポートできる複数回答セットの名前を変更する場合、このオプションを選択します。調査の回答など、ケースごとに複数の値があるデータを記録するのに使用されます。

Data Collection エクスポート・ノード

Data Collection エクスポート・ノードは、Data Collection Data Model に基づき、Data Collection の市場調査ソフトウェアで使用する形式でデータを保存します。この形式は、調査中に収集された質問に対する実際の回答であるケース・データを、ケース・データが収集され整理されたメタデータと区別します。メタデータは、質問テキスト、変数名とその説明、複数の回答セット、種々のテキストの翻訳、ケース・データの構造の定義などの情報から構成されます。 詳しくは、トピック 35 ページの『Data Collectionノード』を参照してください。

メタデータ・ファイル**:** メタデータが保存される質問定義ファイルの名前 (*.mdd*) を指定します。デフォルトの質問は、フィールドのデータ型情報に基づいて作成されます。例えば、名義型 (セット型) フィールドは、定義された各値の質問テキストおよび個別のチェック・ボックスとして使用するフィールド説明を含む単一の質問として表されます。

メタデータ結合**:** メタデータが既存のバージョンを上書きするか、既存のメタデータと結合するかを指定します。結合オプションを選択した場合、ストリームを実行するごとに新しいバージョンが作成されます。これにより、変更が行われるごとに質問のバージョンを記録することができます。各バージョンは、ケース・データの特定のセットを収集するために使用されるメタデータのスナップショットと見なすことができます。

システム変数を使用**:** システム変数がエクスポートされた *.mdd* ファイルに含まれるかどうかを指定します。*Respondent.Serial*、*Respondent.Origin*、*DataCollection.StartTime* などの変数が含まれます。

ケース・データの設定**:** ケース・データがエクスポートされる IBM SPSS Statistics データ (*.sav*) ファイルを指定します。変数名および値名のすべての制限がここで適用されるため、例えば「フィルター」タブに切り替えて、「フィルター」オプション・メニューで「IBM SPSS Statistics 用に名前を変更する」オプションを使用し、フィールド名の無効な文字を修正する必要があります。

データのインポート ノードを生成: エクスポートされたデータ・ファイルを読み込む Data Collection ソース・ノードを自動的に生成するには、このオプションを選択します。

複数回答設定: ストリームに定義された複数の回答セットは、ファイルがエクスポートされると自動的に保存されます。「フィルター」タブで、ノードの複数の回答セットを表示および編集することができます。詳しくは、トピック 161 ページの『複数回答セット編集』を参照してください。

Analytic Server エクスポート・ノード

Analytic Server エクスポート・ノードにより、分析結果のデータを既存の Analytic Server データ・ソースに書き込むことができます。Hadoop 分散ファイル・システム (HDFS) 上のテキスト・ファイルやデータベースなどを使用できます。

通常、Analytic Server エクスポート・ノードを持つストリームは、始点として Analytic Server ソース・ノードも持ち、Analytic Server に送信されて HDFS で実行されます。あるいは、「ローカル」データ・ソースを持つストリームが Analytic Server エクスポート・ノードで終了し、比較的小さな (100,000 レコード以下の) データ・セットをアップロードして、Analytic Server で使用できるようにすることができます。

管理者が定義したデフォルト接続の代わりに独自の Analytic Server 接続を使用する場合は、「デフォルトの **Analytic Server** を使用 **(Use default Analytic Server)**」の選択を解除して、接続を選択します。複数の Analytic Server 接続のセットアップ方法について詳しくは、Analytic Server に接続中を参照してください。

データ・ソース: 使用するデータを含むデータ・ソースを選択します。データ・ソースは、そのソースに関連付けられたファイルおよびメタデータを含みます。「選択」をクリックすると、使用可能なデータ・ソースのリストが表示されます。詳しくは、トピック 14 ページの『データ ソースの選択』を参照してください。

新しいデータ・ソースを作成するか既存のデータ・ソースを編集する必要がある場合は、「データ・ソース・エディターの起動...」をクリックします。

モード: 「追加」を選択すると、既存のデータ・ソースに追加されます。「上書き」を選択すると、データ・ソースの内容が置き換えられます。

このデータのインポート・ノードを生成します。指定したデータ・ソースにエクスポートしたように、ソース・ノードを生成する場合に選択します。ストリーム領域にこのノードが追加されます。

なお、複数の Analytic Server 接続を使用すると、データの流れの制御に役立ちます。例えば、Analytic Server 入力ノードおよびエクスポート・ノードを使用している場合、あるストリームのそれぞれのブランチ内に異なる Analytic Server 接続を使用することで、各ブランチの実行時にブランチがそれ自体の Analytic Server を使用して IBM SPSS Modeler Server にデータが引き出されないようにすることができます。1 つのブランチに複数の Analytic Server 接続が含まれている場合は、Analytic Server から IBM SPSS Modeler Server にデータが引き出されることに注意してください。制限などの詳細については、Analytic Server のストリームのプロパティーを参照してください。

IBM Cognos エクスポート・ノード

IBM Cognos エクスポート・ノードを使用して、IBM SPSS Modeler ストリームから Cognos Analytics にデータを UTF-8 形式でエクスポートできます。こうすることによって、Cognos は IBM SPSS Modeler からの変換データまたはスコアリング・データを使用できます。例えば、Cognos Report Authoring を使用して、予測値や確信度値など、エクスポートされたデータに基づいてレポートを作成できます。レポートは Cognos サーバーに保存し、Cognos ユーザーに配布できます。

注: 関連データだけをエクスポートし、OLAP データはエクスポートしません。

データを Cognos にエクスポートするには、次を指定する必要があります。

- Cognos の接続 - Cognos Analytics サーバーへの接続 (バージョン 11 以降がサポート対象)
- ODBC 接続 - Cognos サーバーが使用する Cognos データ・サーバーへの接続

Cognos の接続内では、使用する Cognos データソースを指定します。このデータソースは、ODBC データソースと同じログインを使用する必要があります。

実際のストリーム・データをデータ・サーバーにエクスポートし、パッケージ メタデータを Cognos サーバーにエクスポートします。

その他のエクスポート・ノードと同様に、ノード・ダイアログ・ボックスの「公開」タブを使用し、IBM SPSS Modeler Solution Publisher で展開するストリームを公開できます。

注: Cognos ソース・ノードでは、Cognos CQM パッケージのみがサポートされます。DQM パッケージは、サポートされません。

Cognos 接続

これは、エクスポートに使用する Cognos Analytics サーバー (バージョン 11 以降がサポート対象) への接続を指定する場所です。この手順では、メタデータを Cognos サーバーの新しいパッケージにメタデータをエクスポートし、ストリーム・データは Cognos データ・サーバーにエクスポートされます。

接続。「編集」ボタンをクリックするとダイアログ・ボックスが表示され、データのエクスポート先となる Cognos サーバーの URL などの詳細情報を定義することができます。IBM SPSS Modeler 経由ですでに Cognos サーバーにログオンしている場合、現在の接続の詳細を編集することもできます。詳しくは、 43 ページの『Cognos の接続』を参照してください。

データ・ソース。データをエクスポートしている Cognos データ・ソース (通常はデータベース) の名前。ドロップダウン・リストには、現在の接続でアクセスできる Cognos データ・ソースがすべて表示されます。リストを更新するには、「リフレッシュ」 ボタンをクリックします。

フォルダー。 エクスポート・パッケージを作成する Cognos サーバーのフォルダーのパスと名前。

パッケージ名。エクスポートされたメタデータを含む指定フォルダー内のパッケージの名前。単一のクエリ・サブジェクトの新しいパッケージでなければなりません。既存のパッケージへはエクスポートできません。

モード: エクスポートの実行方法を指定します。

- パッケージを今すぐ公開。(デフォルト) 「実行」 をクリックするとすぐにエクスポート操作が実行されます。

- アクション・スクリプトをエクスポート。エクスポートを後で実行する XML スクリプトを作成します (Framework Manager を使用するなど)。「ファイル」 フィールドにスクリプトのパスおよびファイル名を入力するか、「編集」 ボタンを使用して、スクリプト・ファイルの名前および場所を指定します。

データのインポート ノードを生成: 指定したデータ・ソースとテーブルにエクスポートしたように、データの入力ノードを生成する場合に選択します。「実行」をクリックすると、ストリーム・キャンバスにこのノードが追加されます。

ODBC 接続

ここで、ストリーム・データをエクスポートする Cognos データ・サーバー (データベース) への接続を指定します。

注: ここで指定するデータ・ソースが、「**Cognos** 接続」 パネルで指定したものと同じデータソースを示さなければなりません。また、Cognos 接続データソースが ODBC データソースと同じログインを使用していることを確認する必要があります。

データ・ソース: 選択したデータ・ソースが表示されます。名前を入力するか、ドロップダウン・リストから選択します。リスト中に目的のデータベースが見つからない場合は、「新規データベース接続の追加」を選択して、「データベース接続」ダイアログ・ボックスから目的のデータベースを検索してください。詳しくは、トピック 20 ページの『データベース接続の追加』を参照してください。

テーブル名: データの送信先のテーブル名を入力します。「テーブルへ挿入」 オプションを選択した場合、「選択」 ボタンをクリックして、データベース中の既存のテーブルを選択することができます。

テーブルの作成: 新規データベース・テーブルを作成する場合、または既存のデータベース・テーブルを上書きする場合に選択します。

テーブルへ挿入: 既存のデータベース・テーブルにデータを新しい行として挿入する場合に選択します。

テーブルを結合: (可能な場合) 選択したデータベースの列を、該当する入力データフィールドの値で更新します。このオプションを選択すると 「結合」 ボタンが有効となり、入力データフィールドをデータベース列に関連付けできるダイアログが表示されます。

既存のテーブルを削除: 新しいテーブルの作成時に、同じ名前を持つ既存のテーブルを削除する場合に選択します。

既存の行を削除: テーブルへの挿入時に、エクスポート前に既存の行をテーブルから削除する場合に選択します。

注: 上記の 2 種類のオプションのどちらも選択されていない場合、ノードの実行時に「上書き警告」というメッセージが表示されます。この警告メッセージを表示しない場合は、「ユーザー オプション」ダイアログ・ボックスの「通知」タブにある、「ノードがデータベース テーブルを上書きする時に警告」 を選択します。

デフォルトの文字列サイズ : 上流にあるデータ型ノードでデータ型不明とされたフィールドは、データベースに文字列フィールドとして書き込まれます。ここには、データ型不明フィールドに使う文字列のサイズを指定します。

「スキーマ」 をクリックしてダイアログ・ボックスを開きます。ここでさまざまなエクスポート・オプション (この機能をサポートするデータベース) を設定、SQL データ型をフィールドに設定し、データベー

ス・インデックス用の第 1 入力フィールド・キーを指定します。詳しくは、トピック 373 ページの『データベース・エクスポートのスキーマのオプション』を参照してください。

「インデックス」 をクリックして、エクスポートされたテーブルのインデックスを生成するオプションを指定してデータベースのパフォーマンスを高めます。詳しくは、トピック 375 ページの『データベース・エクスポートのインデックス・オプション』を参照してください。

バルク・ロードおよびデータベースのコミットに関するオプションを指定するには、「詳細」 をクリックします。詳しくは、トピック 377 ページの『データベース・エクスポートの拡張オプション』を参照してください。

表および列名を引用符で囲む: データベースに CREATE TABLE ステートメントを送信するときに使用するオプションを選択します。スペースまたは非標準文字を含むテーブルや列名は引用符で囲む必要があります。

- 必要に応じて: 引用符が必要かどうかを IBM SPSS Modeler が個別に判断して、自動的に引用符で囲む場合に選択します。
- 常時: テーブル名と列名を常に引用符で囲む場合に選択します。
- しない: 引用符を使用しない場合に選択します。

データのインポート ノードを生成: 指定したデータ・ソースとテーブルにエクスポートしたように、データの入力ノードを生成する場合に選択します。「実行」をクリックすると、ストリーム・キャンバスにこのノードが追加されます。

IBM Cognos TM1 エクスポート・ノード

IBM Cognos エクスポート・ノードを使用して、SPSS Modeler ストリームから Cognos TM1 にデータをエクスポートできます。こうすることによって、Cognos Analytics は SPSS Modeler からの変換データまたはスコアリング・データを使用できます。

注: エクスポートできるのは、指標だけです。コンテキスト ディメンション データをエクスポートすることはできません。または、新しい要素をキューブに追加することもできます。

データを Cognos Analytics (バージョン 11 以降がサポート対象) にエクスポートするには、次を指定する必要があります。

- Cognos TM1 サーバーへの接続。
- データのエクスポート先のキューブ。
- SPSS データ名から、対応する TM1 ディメンションと指標へのマッピング。

注: TM1 ユーザーは、次の権限が必要です: キューブの書き込み権限、ディメンションの読み取り権限、ディメンション要素の書き込み権限。また、SPSS Modeler が Cognos TM1 データのインポートとエクスポートを行うには、IBM Cognos TM1 10.2 フィックスパック 3 以降が必要です。以前のバージョンに基づいていた既存のストリームは、依然として機能します。

このノードに対して管理者の資格情報は不要です。しかし、17.1 より前の古いレガシー TM1 ノードを引き続き使用している場合、管理者の資格情報は今までどおり必要です。

その他のエクスポート・ノードと同様に、ノード・ダイアログ・ボックスの「公開」タブを使用し、IBM SPSS Modeler Solution Publisher で展開するストリームを公開できます。

注: SPSS Modeler で TM1 ソース ノードまたは TM1 エクスポート ノードを使用するには、事前に tm1s.cfg ファイル内の一部の設定を確認する必要があります。このファイルは、TM1 サーバーのルート ディレクトリーにある TM1 サーバー構成ファイルです。

- HTTPPortNumber - 有効なポート番号を設定します。通常は、1 から 65535 です。これは、後で接続時にノードで指定するポート番号ではなく、TM1 で使用される内部ポート (デフォルトでは無効) であることに注意してください。必要な場合は、TM1 管理者に問い合わせて、このポートの有効な設定を確認してください。
- UseSSL - これを *True* に設定すると、HTTPS がトランスポート プロトコルとして使用されます。この場合、TM1 認証を SPSS Modeler Server JRE にインポートする必要があります。

データをエクスポートする目的での IBM Cognos TM1 キューブへの接続

IBM Cognos TM1 データベースにデータをエクスポートするための最初のステップは、IBM Cognos TM1 ダイアログ ボックスの「接続」タブで、該当する TM1 管理ホストと、関連するサーバーおよびキューブを選択することです。

注: TM1 にデータをエクスポートするときは、実際の「ヌル」値のみが破棄されます。ゼロ (0) の値は有効な値としてエクスポートされます。また、「マッピング」タブでディメンションにマップできるのは、ストレージ タイプが文字列 であるフィールドに限られることにも注意してください。TM1 にエクスポートする前に、IBM SPSS Modeler クライアントを使用して、文字列以外のデータ型を文字列に変換する必要があります。

管理ホスト: 接続先の TM1 サーバーがインストールされている管理ホストの URL を入力します。管理ホストは、すべての TM1 サーバーに対する単一の URL として定義されます。この URL から、ご使用の環境にインストールされ、稼働しているすべての IBM Cognos TM1 サーバーをディスカバーし、これらのサーバーにアクセスすることができます。

TM1 サーバー: 管理ホストへの接続を確立したら、インポートするデータが存在するサーバーを選択して「ログイン」をクリックします。このサーバーに以前に接続していない場合は、「ユーザー名」と「パスワード」の入力を要求するプロンプトが出されます。または、以前に入力し、「保管されている資格情報」として保存したログイン詳細を検索できます。

エクスポート対象の **TM1** キューブの選択: データのエクスポート先の TM1 サーバーに存在するキューブの名前が表示されます。

エクスポートするデータを選択するには、キューブを選択して右矢印をクリックし、キューブを「キューブへのエクスポート」フィールドに移動します。キューブを選択したら、「マッピング」タブを使用して、TM1 ディメンションおよび測定を、該当する SPSS フィールドまたは固定値 (「選択」操作) にマップします。

エクスポート用の IBM Cognos TM1 データのマップ

TM1 管理ホストと、関連する TM1 サーバーおよびキューブを選択したら、IBM Cognos TM1 の「エクスポート」ダイアログ ボックスの「マッピング」タブを使用して、TM1 ディメンションおよび測定を SPSS フィールドにマップするか、TM1 ディメンションを固定値に設定します。

注: ディメンションにマップできるのは、ストレージ タイプが文字列 であるフィールドに限られます。TM1 にエクスポートする前に、IBM SPSS Modeler クライアントを使用して、文字列以外のデータ型を文字列に変換する必要があります。

フィールド: エクスポートに使用可能な SPSS データ ファイルのデータ フィールド名がリストされます。

TM1 ディメンジョン: 「接続」タブで選択した TM1 キューブと、標準ディメンション、数値データ ディメンション、選択した数値データ ディメンションの要素が表示されます。SPSS データ フィールドにマップする TM1 ディメンジョンまたは測定の名前を選択します。

「マッピング」タブでは以下のオプションを使用できます。

数値データ ディメンションの選択 **(Select a measure dimension)**: 選択したキューブのディメンションのリストから、数値データ ディメンションにするものを 1 つ選択します。

数値データ ディメンション以外のディメンションを選択し、「選択」をクリックすると、ダイアログが表示され、選択したディメンションの葉要素が表示されます。選択できるのは、葉要素のみです。選択した要素には、**S**の ラベルが付けられます。

マップ: 選択した SPSS データ フィールドを、選択した TM1 ディメンションまたは測定 (標準ディメンション、または数値データ ディメンションの特定の測定または要素) にマップします。マップしたフィールドには **M** のラベルが付きます。

マップ解除: 選択した SPSS データ フィールドを、選択した TM1 ディメンションまたは測定からマップ解除します。一度にマップ解除できるマッピングは 1 つだけであることに注意してください。マップ解除した SPSS データ フィールドは左側の列に戻されます。

新規作成: TM1 数値データ ディメンションに新しい測定を作成します。ダイアログが表示されるので、そこで新しい「**TM1** 測定名」を入力します。このオプションは数値データ ディメンションの場合にのみ使用できます。標準ディメンションの場合は使用できません。

TM1 について詳しくは、IBM Cognos TM1 の資料 (http://www-01.ibm.com/support/knowledgecenter/SS9RXT_10.2.2/com.ibm.swg.ba.cognos.ctm1.doc/welcome.html) を参照してください。

SAS エクスポート・ノード

この機能は SPSS Modeler Professional および SPSS Modeler Premium で使用できます。

SAS エクスポート・ノードを使用すると、データを SAS 形式で書き込み、SAS 互換または SAS 互換のソフトウェア・パッケージに読み込むことができます。次の 3 種類の SAS ファイル形式でエクスポートできます。SAS for Windows/OS2、SAS for UNIX、または SAS。

SAS エクスポート・ノード、「エクスポート」タブ

エクスポート・ファイル **:** ファイルの名前を指定します。ファイル名を入力するか、またはファイル選択ボタンをクリックしてファイルの場所を指定します。

エクスポート**:** エクスポートするファイル・フォーマットを指定します。「**SAS for Windows/OS2**」、「**SAS for UNIX**」 または、「**SAS** バージョン **7/8/9**」 を指定することができます。

フィールド名のエクスポート**:** SAS で使用するために、IBM SPSS Modeler からフィールド名とラベルをエクスポートするオプションを選択します。

- 名前と変数ラベル**:** IBM SPSS Modeler のフィールド名とフィールド・ラベルの両方をエクスポートする場合に選択します。名前は SAS の変数名としてエクスポートされ、ラベルは SAS の変数ラベルとしてエクスポートされます。

- 変数ラベルとして使用**:** IBM SPSS Modeler のフィールド名を SAS の変数ラベルとして使用する場合に選択します。IBM SPSS Modeler は SAS 変数名では無効であるフィールド名の文字を使用できます。SAS で無効な名前が作成されることを防止するには、代わりに 「名前と変数ラベル」 を選択します。

データのインポート ノードを生成**:** エクスポートされたデータ・ファイルを読み込む SAS ソース・ノードを自動的に生成するには、このオプションを選択します。詳しくは、トピック 47 ページの『SAS ソース・ノード』を参照してください。

注**:** 文字列の最大長は 255 バイトです。文字列が 255 バイトを超える場合は、エクスポート時に切り捨てられます。

Excel エクスポート・ノード

Excel エクスポート・ノードでは、データを Microsoft Excel .xlsx 形式で出力します。オプションで、ノードが実行されるときに自動的に Excel が起動し、エクスポートするファイルを開けるように選択できます。

Excel ノードの「エクスポート」タブ

ファイル名。 ファイル名を入力するか、またはファイル選択ボタンをクリックしてファイルの場所を指定します。デフォルトのファイル名は *excelexp.xlsx* です。

ファイルの種類。Excel の .xlsx ファイル形式がサポートされています。

ファイルの新規作成。新しい Excel ファイルを作成します。

既存ファイルに挿入。「セルで開始」 フィールドで指定されたセル以降の内容が置き換えられます。スプレッドシートの他のセルは、元の内容が残されます。

フィールド名を含める。フィールド名がワークシートの最初の行に表示されるかどうかを指定します。

セルの開始点。最初のエクスポート・レコードに使用されるセルの場所 (または、「フィールド名を含める」 がオンの場合、最初のフィールド名)。データは最初のセルの右側から下に向かって入力されます。

ワークシートを選択。データをエクスポートするワークシートを指定します。インデックスまたは名前のどちらかで、ワークシートを指定します。

- インデックスによる。ファイルを新規作成する場合、エクスポートするワークシートを示す 0 ～ 9 の値を指定します。最初のワークシートは 0、2 番目のワークシートは 1 というように指定します。ワークシートがすでにこの位置にある場合にのみ、10 以上の値を使用できます。
- 名前順。インポートするワークシートの名前を指定します。新しいファイルを作成している場合、ワークシートに使用される名前を指定します。既存のファイルに挿入している場合、ワークシートがあればデータはそのワークシートに挿入され、ない場合はこの名前を持つ新しいワークシートが作成されます。

Excel を起動する。ノードが実行されるときに、Excel が自動的に起動され、エクスポート・ファイルが開くようにするかどうかを指定します。IBM SPSS Modeler Server に対して分散モードで実行中の場合は、この出力はサーバーのファイル・システムに保存され、クライアント上で Excel はエクスポートされたファイルのコピーで起動されることに注意してください。

データのインポート ノードを生成**:** エクスポートされたデータ・ファイルを読み込む Excel 入力ノードを、自動的に生成する場合に選択します。 詳しくは、トピック 48 ページの『Excel ソース・ノード』を参照してください。

拡張エクスポート・ノード

拡張エクスポート・ノードを使用すると、R スクリプトまたは Python for Spark スクリプトを実行して、データをエクスポートできます。

拡張エクスポート・ノード - 「シンタックス」タブ

シンタックスのタイプ (**R** または **Python for Spark**) を選択します。詳しくは、以下のセクションを参照してください。シンタックスの準備ができたら、「実行」をクリックして、拡張エクスポート・ノードを実行できます。

R シンタックス

「**R** シンタックス」。データ分析用のカスタムの R スクリプト・シンタックスを、このフィールドに入力するか、貼り付けることができます。

「フラグ型フィールドの変換」。フラグ型フィールドの処理方法を指定します。「文字列を因子に、整数および実数を倍精度に」および「論理値 **(**真、偽**)**」の 2 つのオプションがあります。「論理値 **(**真、偽**)**」を選択した場合、フラグ型フィールドの元の値は失われます。例えば、フィールドに「Male」および「Female」という値がある場合、これらの値は「真」および「偽」に変更されます。

「欠損値を **R** の欠損値 **(NA)** に変換」。選択すると、欠損値はすべて、R の NA 値に変換されます。R では、欠損値の識別に値 NA が使用されます。使用する R 関数によっては、データに NA が含まれていた場合の関数の動作を制御するために使用される引数が含まれている場合があります。例えば、関数によって NA を含むレコードを自動的に除外することを選択できる場合があります。このオプションが選択されない場合、すべての欠損値はそのまま R に渡されます。これらの欠損値は R スクリプトの実行時にエラーの原因となる可能性があります。

「時間帯を考慮した特殊な制御で日時フィールドを **R** のクラスに変換」。選択すると、日付形式または日付/時刻形式の変数が R の日付/時刻形式に変換されます。次のいずれかのオプションを選択する必要があります。

- 「**R POSIXct**」。日付形式または日付/時刻形式の変数が R の POSIXctオブジェクトに変換されます。
- 「**R POSIXlt (**リスト**)**」。日付形式または日付/時刻形式の変数が R の POSIXltオブジェクトに変換されます。

注**:** POSIX 形式は、拡張オプションです。これらのオプションは、ご使用の R スクリプトで、これらの形式を必要とする方法で日付/時刻フィールドを処理するように指定している場合にのみ使用してください。POSIX 形式は、時刻形式の変数には適用されません。

Python シンタックス

「**Python** シンタックス」。データ分析用のカスタムの Python スクリプト・シンタックスを、このフィールドに入力するか、貼り付けることができます。Python for Spark について詳しくは、Python for Spark およびPython for Spark を使用したスクリプトを参照してください。

拡張エクスポート・ノード - 「コンソール出力」タブ

「コンソール出力」タブには、「シンタックス」タブの R スクリプトまたは Python for Spark スクリプトが実行されたときに受信するすべての出力が含まれます (例えば、R スクリプトを使用する場合、「シンタックス」タブの「**R** シンタックス」フィールドにある R スクリプトが実行されたときに R コンソールから受信する出力が表示されます)。この出力には、R スクリプトまたは Python スクリプトの実行時に生成される R または Python のエラー・メッセージや警告が含まれる場合があります。出力は、主にスクリプトをデバッグするために使用できます。「コンソール出力」タブには、「**R** シンタックス」フィールドまたは「**Python** シンタックス」フィールドのスクリプトも表示されます。

拡張エクスポート・スクリプトが実行されるたびに、R コンソールまたは Python for Spark から受信した出力で「コンソール出力」タブの内容が上書きされます。出力を編集することはできません。

XML エクスポート・ノード

XML エクスポート・ノードを使用して、UTF-8 エンコードを使用し、データを XML 形式で出力できます。オプションで、エクスポートしたデータをストリームに読み込む XML ソース・ノードを作成できます。

XML エクスポート・ファイル **:** データをエクスポートする XML ファイルの完全パスとファイル名です。

XML スキーマを使用： スキーマまたは DTD を使用して、エクスポートするデータの構造を制御します。制御することによって 「マップ」 ボタンが有効になります。

スキーマまたは DTD を使用しない場合、次のデフォルト構造がエクスポート・データに使用されます。

```
<records>
     <record>
          <fieldname1>value</fieldname1>
          <fieldname2>value</fieldname2>
      :
          <fieldnameN>value</fieldnameN>
     </record>
     <record>
      :
      :
     </record>
      :
      :
</records>
```

フィールド名のスペースはアンダースコアーに置き換えられます。例えば、「My Field」は「<My_Field>」になります。

マップ。XML スキーマの使用を選択した場合、このボタンを選択すると、XML 構造のどの部分を使用して新しいレコードを開始するかを指定するためのダイアログが表示されます。 詳しくは、トピック 398 ページの『XML マッピングのレコード・オプション』を参照してください。

マップしたフィールド： マップされたフィールド数を示します。

データのインポート ノードを生成**:** エクスポートされたデータ・ファイルをストリームに読み込む XML ソース・ノードを自動的に生成するには、このオプションを選択します。詳しくは、トピック 49 ページの『XML ソース・ノード』を参照してください。

XML データの作成

XML 要素が指定されると、フィールド値が要素タグ内に入力されます。

<element>value</element>

属性がマップされると、フィールド値が属性の値として指定されます。

<element attribute="value">

フィールドが <records> 要素の上の要素にマップされる場合、そのフィールドは一度だけ記述され、すべてのレコードで定数になります。この要素の値は最初のレコードに由来します。

Null 値が書きむ場合、空白の内容を指定します。要素の場合は、次のようになります。

<element></element>

属性の場合は、次のようになります。

<element attribute="">

XML マッピングのレコード・オプション

「レコード」タブを使用して、各新規レコードの開始に使用する XML 構造の部分を指定できます。スキーマに正しくマッピングするために、レコード区切り文字を指定する必要があります。

XML 構造: 前の画面で指定された XML スキーマの構造を示す階層ツリー。

レコード **(XPath 式):** レコード区切り文字を設定するには、XML 構造の要素を選択し、右方向矢印ボタンをクリックします。入力データにこの要素が出現するごとに、新しいレコードが出力ファイルに作成されます。

注 : XML 構造のルート要素を選択すると、レコードを 1 つだけ書き込むことができ、他のすべてのレコードがスキップされます。

XML マッピングのフィールド・オプション

スキーマ ファイルを使用する場合、「フィールド」タブを使用して、データ・セットのフィールドを XML 構造の要素または属性にマッピングします。

要素名または属性名と一致するフィールド名は、要素名または属性名が一意である限り、自動的にマッピングされます。そのため、要素および属性の名前がいずれも field1 である場合、自動マッピングは行われません。field1 という名前の項目が構造に 1 つだけある場合、ストリーム内に同じ名前を持つフィールドが自動的にマッピングされます。

フィールド: モデル内にあるフィールドのリスト。1 つまたは複数のフィールドをマッピングのソース部分として選択します。リストの一番下のボタンを使用してすべてのフィールドを選択したり、特定の測定の尺度であるすべてのフィールド選択できます。

XML 構造: XML 構造の要素をマップ対象として選択します。マッピングを作成するには、「マップ」をクリックします。その後、マッピングが表示されます。この方法でマッピングされたフィールド数は、このリストの下に表示されます。

マッピングを削除するには、XML 構造リストの項目を選択し、「マップ解除」をクリックします。

表示属性: XML 構造の XML 要素の属性があれば、表示または非表示にします。

XML マッピングのプレビュー

「プレビュー」タブで、「更新」 をクリックすると、書き込まれる XML のプレビューが表示されます。

マッピングが不適切である場合、「レコード」タブまたは「フィールド」タブに戻ってエラーを修正し、「更新」 をもう一度クリックして結果を表示します。

エクスポート・ノードの共通タブ

次のオプションは、すべてのエクスポート・ノードで、該当するタブをクリックすることで指定できます。

- 「公開」タブ: ストリームの結果を公開するために使用します。
- 「注釈」タブ: すべてのノードで使用されるこのタブには、ノード名の変更、カスタム・ツールヒントの提供、および長い注釈の保存などのオプションが用意されています。

ストリームの公開

ストリームの公開は、標準エクスポート・ノード (データベース、フラット ファイル、Statistics エクスポート、拡張エクスポート、Data Collection エクスポート、SAS エクスポート、Excel、および XML エクスポートの各ノード) のいずれかを使用して、IBM SPSS Modeler から直接行われます。エクスポート・ノードのタイプにより、IBM SPSS Modeler Solution Publisher Runtime または外部アプリケーションを使用して発行されたストリームが実行されるたびに、書き込まれる結果の形式が決定されます。例えば、発行されたストリームが実行されるたびに結果をデータベースに書き込む場合は、データベース・エクスポート・ノードを使用します。

ストリームを公開するには

1. ストリームを通常の方法で開くか構築し、エクスポート・ノードを端に接続します。
2. エクスポート・ノードの「公開」タブで、公開するファイルのルート名 (.pim、.par、および .xml の拡張子が追加されるファイル名) を指定します。
3. 「公開」をクリックして、ストリームを公開するか、「ストリームの公開」を選択して、ノードが実行されるたびにストリームを自動的に公開するようにします。

名前: 公開画像およびパラメーター・ファイルのルート名を指定します。

- 画像ファイル (*.pim) には、Runtime が、発行されたストリームをエクスポートの時点とまったく同じように実行するために必要なすべての情報があります。ストリームの設定 (入力データ・ソースや出力データ・ファイルなど) を変更する必要がないことが明らかな場合は、画像ファイルだけを展開できます。
- パラメーター・ファイル (*.par) には、データ・ソース、出力ファイル、および実行オプションに関する設定可能な情報が含まれます。ストリームを再発行せずにストリームの入力または出力を制御するには、パラメーター・ファイルおよび画像ファイルが必要です。
- メタデータ・ファイル (*.xml) は、イメージやそのデータ・モデルの入力および出力を記述します。ランタイム ライブラリーを組み込み、入力データおよび出力データの構造を認識する必要があるアプリケーションによって使用するために設計されています。

  注: このファイルは、「メタデータを公開する」オプションを選択している場合にのみ、作成されます。

パラメーターを公開する: 必要に応じて、*.par ファイルにストリーム・パラメーターを含めることができます。イメージを実行する場合、*.par を編集して、またはランタイム API によってこれらのストリーム・パラメーター値を変更することができます。

このオプションを選択すると、「パラメータ」ボタンを使用できるようになります。このボタンをクリックすると、「パラメーター公開」ダイアログ・ボックスが表示されます。

「公開」列の関連するオプションを選択して、公開画像に含めるパラメータを選択します。

ストリームの実行時: ノードが実行されたときにストリームが自動的に公開されるかどうかを指定します。

- データのエクスポート: ストリームを公開せずに、標準の方法でエクスポート・ノードを実行します。(基本的に、ノードは IBM SPSS Modeler で IBM SPSS Modeler Solution Publisher が使用できない場合と同じ方法で実行します。)このオプションを選択した場合は、「エクスポート ノード」ダイアログ・ボックスで「公開」を明示的にクリックしない限り、ストリームは公開されません。また、ツール・バーの「公開」ツールを使用するか、スクリプトを使用することで、現在のストリームを発行できます。
- ストリームの公開**:** IBM SPSS Modeler Solution Publisher を使用して、展開用にストリームを公開します。実行するたびにストリームを自動的に発行する場合は、このオプションを選択します。

注**:**

- 発行済みのストリームを新規または更新されたデータと組み合わせて実行することを計画している場合、入力フィールド内のフィールドの順序は、公開済みのストリームで指定された入力ノードの入力ファイルと同じであることが必要です。
- 外部アプリケーションに公開する場合は、無関係なフィールドをフィルタリングしたり、入力条件に準拠したフィールド名に変更することを検討してください。両方とも、エクスポート・ノードの前にフィルター・ノードを使用することで達成できます。

第 8 章 IBM SPSS Statistics ノード

IBM SPSS Statistics ノードの概要

IBM SPSS Modeler およびそのデータ・マイニング機能を補完するために、IBM SPSS Statistics では、詳細な統計分析とデータ管理を実行する機能を提供します。

互換性があり、ライセンスされた IBM SPSS Statistics がインストールされている場合、IBM SPSS Modeler から接続し、IBM SPSS Modeler でサポートされていない複雑で、多段階に及ぶデータ操作および分析を実行できます。高度なユーザーの場合、コマンド・シンタックスを使用して分析を詳細に変更できるオプションもあります。バージョンの互換性に関する詳細については、リリース・ノートを参照してください。

可能な場合、ノード・パレットの指定された部分に IBM SPSS Statistics ノードが表示されます。

注: IBM SPSS Statistics の変換ノード、モデル・ノード、出力ノードを使用する前に、データ型ノードでデータをインスタンス化することをお勧めします。また、これは AUTORECODE シンタックス・コマンドを使用する場合に必要です。

IBM SPSS Statistics パレットには次のノードがあります。

Statistics ファイル ノードは、IBM SPSS Statistics で使用される *.sav* または *.zsav* ファイル形式のデータおよび IBM SPSS Modeler に保存されたキャッシュ ファイル (同じ形式を使用する) を読み込みます。

Statistics 変換ノードは、IBM SPSS Modeler のデータ・ソースに対する IBM SPSS Statistics シンタックス・コマンドの選択を行います。このノードは、ライセンスが与えられた IBM SPSS Statistics のコピーが必要です。

Statistics モデル・ノードを使用すると、PMML を作成する IBM SPSS Statistics 手続きを実行してデータを分析および使用することができます。このノードは、ライセンスが与えられた IBM SPSS Statistics のコピーが必要です。

Statistics 出力ノードを使用すると、IBM SPSS Statistics 手続きを呼び出し、IBM SPSS Modeler データを分析することができます。さまざまな IBM SPSS Statistics 分析手続きにアクセスできます。このノードは、ライセンスが与えられた IBM SPSS Statistics のコピーが必要です。

Statistics エクスポート・ノードでは、IBM SPSS Statistics *.sav* または *.zsav* フォーマットでデータを出力します。*.sav* または *.zsav* ファイルは、IBM SPSS Statistics Base およびその他の製品で読み込むことができます。この形式は、IBM SPSS Modeler のキャッシュ・ファイルでも使用されます。

注**:** SPSS Statistics のライセンスがシングル・ユーザーのみで、複数のブランチでストリームを実行し、それぞれのブランチに SPSS Statistics ノードがある場合、ライセンス・エラーが発生する場合があります。このエラーは、一方のブランチの SPSS Statistics セッションがもう一方のセッションが開始しようとする前に終了する場合に発生します。可能な場合、SPSS Statistics ノードを持つ複数のブランチが並行して一考されないようストリームを再度設計する必要があります。

Statistics ファイル・ノード

Statistics ファイル・ノードを使用すると、保存された IBM SPSS Statistics ファイル (*.sav* または *.zsav*) からデータを直接読み込むことができます。この形式は、旧バージョンの IBM SPSS Modeler からキャッシュ・ファイルを置換するために使用されます。保存されているキャッシュ・ファイルをインポートする場合は、IBM SPSS Statistics ファイル・ノードを使用します。

インポート・ファイル。ファイルの名前を指定します。ファイル名を入力するか、省略符号ボタン (「**...**」) をクリックしてファイルを選択できます。ファイルを選択すると、ファイル・パスが表示されます。

ファイルはパスワードで暗号化されています。ファイルがパスワード保護されていることがわかっている場合に、このボックスを選択します。「パスワード」に入力を要求するプロンプトが出されます。ファイルがパスワード保護されている場合は、パスワードを入力しないと、別のタブへの移動時、データのリフレッシュ時、ノード内容のプレビュー時、またはノードを含むストリームの実行時に警告メッセージが表示されます。

注**:** パスワード保護されたファイルを開くことができるのは、IBM SPSS Modeler バージョン 16 以降のみです。

変数名。IBM SPSS Statistics の *.sav* または *.zsav* ファイルからインポートする場合の変数名とラベルの処理方法を選択します。ここで指定したメタデータは、IBM SPSS Modeler での作業中は保存され、IBM SPSS Statistics で使用するために再びエクスポートすることができます。

- 名前とラベルを読み込む。変数名とラベルの両方を IBM SPSS Modeler に読み込むために選択します。デフォルトでは、このオプションが選択され、変数名がデータ型ノードに表示されます。ラベルは、「ストリームのプロパティー」ダイアログ・ボックスで指定したオプションに応じて、グラフやモデル・ブラウザー、その他のタイプの出力中に表示できます。デフォルトでは、出力中のラベル表示は無効になっています。
- ラベルを名前として読み取る。短いフィールド名ではなく、IBM SPSS Statistics の *.sav* または *.zsav* ファイルから詳細な変数ラベルを読み取り、そのラベルを IBM SPSS Modeler で変数名として使用する場合に選択します。

値。IBM SPSS Statistics の *.sav* または *.zsav* ファイルからインポートする場合の値とラベルの処理方法を選択します。ここで指定したメタデータは、IBM SPSS Modeler での作業中は保存され、IBM SPSS Statistics で使用するために再びエクスポートすることができます。

- データとラベルを読み込む。実際の値と値ラベルの両方を IBM SPSS Modeler に読み込むために選択します。デフォルトでは、このオプションが選択され、値自体がデータ型ノードに表示されます。値ラ

ベルは、「ストリームのプロパティー」ダイアログ・ボックスで指定したオプションに応じて、式ビルダー、グラフ、モデル・ブラウザー、その他の種類の出力中に表示できます。

- ラベルをデータとして読み込み。値を表すために使用される数値コードまたはシンボリック・コードではなく、*.sav* ファイルまたは *.zsav* ファイルの値ラベルを使用する場合に選択します。例えば、「1」と「2」の値が実際には「男性」と「女性」を表す性別フィールドを持つデータについてこのオプションを選択すると、性別フィールドが文字列に変換され、「男性」と「女性」が実際の値としてインポートされます。

  このオプションを選択する前に、IBM SPSS Statistics データ中の欠損値を検討しておくことが大切です。例えば、数値フィールドで欠損値についてのみラベルを使用している場合 (0=回答なし、99=不明)、このオプションを選択すると、「回答なし」と「不明」という値ラベルだけがインポートされ、数値フィールドが文字列に変換されます。このような場合は、値自体をインポートして、データ型ノードに欠損値を設定する必要があります。

フィールド形式情報を使用して、ストレージを指定します。このボックスを選択解除すると、 *.sav* ファイル内で整数形式のフィールド値 (例えば、IBM SPSS Statistics の変数ビューで F*n*.0 として指定されているフィールド) は、整数ストレージを使用してインポートされます。文字列を除くすべてのフィールド値は、実数としてインポートされます。

このボックスを選択すると (デフォルト)、文字列以外のフィールド値はすべて、 *.sav* ファイル内で整数形式であるかないかに関係なく、実数としてインポートされます。

複数回答設定。ストリームに定義された複数の回答セットは、IBM SPSS Statistics ファイルがエクスポートされると自動的に保存されます。「フィルター」タブで、ノードの複数の回答セットを表示および編集することができます。 詳しくは、トピック 161 ページの『複数回答セット編集』を参照してください。

Statistics 変換ノード

Statistics 変換ノードで IBM SPSS Statistics コマンド・シンタックスを使用して、データ変換を実行することができます。これにより、IBM SPSS Modeler がサポートしていない多くの変換を完了することができ、さらに単一ノードからの数多くのフィールド作成など、複雑で他段階におよぶ変換を自動化することができます。それは、さらなる分析のためにデータが IBM SPSS Modeler に返されることを除いて Statistics 出力ノードに似ている一方、出力ノードではデータはグラフやテーブルなど要求された出力オブジェクトとして返されます。

このノードを使用するには、互換性のあるバージョンの IBM SPSS Statistics をインストールし、ライセンス認証する必要があります。詳しくは、トピック 366 ページの『IBM SPSS Statistics ヘルパー アプリケーション』を参照してください。 互換性に関する詳細については、リリース・ノートを参照してください。

必要に応じて、「フィルター」タブを使用して、フィールドの名前をフィルターに掛けたり、IBM SPSS Statistics の命名規格に適合するようにフィールドの名前を変更したりできます。詳しくは、トピック 388 ページの『IBM SPSS Statistics 用のフィールドの名前変更またはフィルタリング』を参照してください。

シンタックス・リファレンス。 IBM SPSS Statistics の具体的な手続きについて詳しくは、IBM SPSS Statistics ソフトウェアのコピーに付属している「*IBM SPSS Statistics* コマンド・シンタックス・リファレンス」ガイドを参照してください。「シンタックス」タブでガイドを表示するには、「シンタックス エディター」オプションを選択し、「IBM SPSS Statistics シンタックスのヘルプを起動」ボタンをクリックします。

注：このノードではすべての IBM SPSS Statistics シンタックスをサポートしているわけではありません。 詳しくは、トピック『利用可能なシンタックス』を参照してください。

Statistics 変換ノードの「シンタックス」タブ

IBM SPSS Statistics ダイアログ・オプション

IBM SPSS Statistics シンタックスについてよく知らない場合、IBM SPSS Modeler でシンタックスを作成する最も簡単な方法は、「**IBM SPSS Statistics** ダイアログ」 オプションを選択し、手順のダイアログ・ボックスを選択、ダイアログボックスを入力して「OK」をクリックします。その後、IBM SPSS Modeler で使用している IBM SPSS Statistics ノードの「シンタックス」タブにシンタックスが配置されます。その後、ストリームを実行して手順より出力を取得することができます。

IBM SPSS Statistics シンタックス・エディターのオプション

検査：ダイアログ・ボックスの上部にシンタックス・コマンドを入力し、このボタンを使用して入力内容を有効にします。いかなる無効なシンタックスも、ダイアログ・ボックスの下部で識別されます。

チェックのプロセスに時間がかからないようにするために、シンタックスを有効にする時、データ・セット全体ではなくデータの代表的なサンプルをチェックして入力が有効であることを確認します。

利用可能なシンタックス

IBM SPSS Statistics の古いシンタックスが数多くある場合や、IBM SPSS Statistics のデータ準備機能を理解している場合、Statistics 変換ノードを利用して、多くの既存の変換を実行することができます。原則的に、ノードを使用すると、予測可能な方法でデータを変換することができます。例えば、ループ・コマンドの実行や、データの追加、ソート、フィルタリング、選択によって変換することができます。

実行可能なコマンドの例

- 二系分布に従った乱数の計算

  ```
  COMPUTE newvar = RV.BINOM(10000,0.1)
  ```

- ある変数を新しい変数に再コード

  ```
  RECODE Age (Lowest thru 30=1) (30 thru 50=2) (50 thru Highest=3) INTO AgeRecoded
  ```

- 欠損値を置換

```
RMV Age_1=SMEAN(Age)
```

Statistics 変換ノードでサポートされる IBM SPSS Statistics シンタックスを以下に示します。

コマンド名

```
ADD VALUE LABELS
APPLY DICTIONARY
AUTORECODE
BREAK
CD
CLEAR MODEL PROGRAMS
CLEAR TIME PROGRAM
CLEAR TRANSFORMATIONS
COMPUTE
```

コマンド名

COUNT

CREATE

DATE

DEFINE-!ENDDEFINE

DELETE VARIABLES

DO IF

DO REPEAT

ELSE

ELSE IF

END CASE

END FILE

END IF

END INPUT PROGRAM

END LOOP

END REPEAT

EXECUTE

FILE HANDLE

FILE LABEL

FILE TYPE-END FILE TYPE

FILTER

FORMATS

IF

INCLUDE

INPUT PROGRAM-END INPUT PROGRAM

INSERT

LEAVE

LOOP-END LOOP

MATRIX-END MATRIX

MISSING VALUES

N OF CASES

NUMERIC

PERMISSIONS

PRESERVE

RANK

RECODE

RENAME VARIABLES

RESTORE

RMV

SAMPLE

SELECT IF

SET

コマンド名

SORT CASES

STRING

SUBTITLE

TEMPORARY

TITLE

UPDATE

V2C

VALIDATEDATA

VALUE LABELS

VARIABLE ATTRIBUTE

VARSTOCASES

VECTOR

Statistics モデル・ノード

Statistics モデル・ノードを使用すると、PMML を作成する IBM SPSS Statistics 手続きを実行してデータを分析および使用することができます。作成したモデル・ナゲットを IBM SPSS Modeler ストリーム内で通常の方法で使用し、スコアリングを行うことができます。

このノードを使用するには、互換性のあるバージョンの IBM SPSS Statistics をインストールし、ライセンス認証する必要があります。詳しくは、トピック 366 ページの『IBM SPSS Statistics ヘルパー アプリケーション』を参照してください。 互換性に関する詳細については、リリース・ノートを参照してください。

利用可能な IBM SPSS Statistics 分析手続きは、ライセンスの種類によって異なります。

Statistics モデル・ノードの「モデル」タブ

モデル名**:** 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

ダイアログの選択 **:** 選択し、実行できる IBM SPSS Statistics 手続きのリストを表示します。PMML を生成し、ライセンス付与されるこれらの手続きのみが表示され、ユーザー指定の手続きは含まれません。

1. 該当する手続きをクリックすると、関連する IBM SPSS Statistics ダイアログが表示されます。
2. 「IBM SPSS Statistics」ダイアログで、手続きの詳細を入力します。
3. 「**OK**」 をクリックして Statistics モデル・ノードに戻ります。IBM SPSS Statistics シンタックスが「モデル」タブに表示されます。
4. クエリーを変更する場合など、「IBM SPSS Statistics」ダイアログにいつでも戻るには、手続き選択ボタンの右側の「IBM SPSS Statistics ダイアログ表示」ボタンをクリックします。

Statistics モデル・ノード - モデル・ナゲットの要約

Statistics モデル・ノードを実行する場合、関連する IBM SPSS Statistics 手続きを実行し、スコアリングするために IBM SPSS Modeler ストリームで使用できるモデル・ナゲットを生成できます。

モデル・ナゲットの「要約」タブには、フィールド、構築の設定、およびモデル推定プロセスについての情報が表示されます。結果は、特定の項目をクリックすると開いたり閉じたりできるツリーで表示されます。

「モデルの表示」 をクリックすると、IBM SPSS Statistics 出力ビューアーの修正された形式で結果が表示されます。このビューアーの詳細については、IBM SPSS Statistics のマニュアルを参照してください。

通常のエクスポート、および印刷オプションは、「ファイル」メニューから行うことができます。 詳しくは、トピック 323 ページの『出力を表示』を参照してください。

Statistics 出力ノード

Statistics 出力ノードを使用すると、IBM SPSS Statistics プロシージャーを呼び出して IBM SPSS Modeler データを分析することができます。結果はブラウザー・ウィンドウに表示したり、IBM SPSS Statistics 出力ファイル形式で保存することができます。IBM SPSS Modeler からさまざまな IBM SPSS Statistics 分析手続きにアクセスできます。

このノードを使用するには、互換性のあるバージョンの IBM SPSS Statistics をインストールし、ライセンス認証する必要があります。詳しくは、トピック 366 ページの『IBM SPSS Statistics ヘルパー アプリケーション』を参照してください。 互換性に関する詳細については、リリース・ノートを参照してください。

必要に応じて、「フィルター」タブを使用して、フィールドの名前をフィルターに掛けたり、IBM SPSS Statistics の命名規格に適合するようにフィールドの名前を変更したりできます。詳しくは、トピック 388 ページの『IBM SPSS Statistics 用のフィールドの名前変更またはフィルタリング』を参照してください。

シンタックス・リファレンス。 IBM SPSS Statistics の具体的な手続きについて詳しくは、IBM SPSS Statistics ソフトウェアのコピーに付属している「*IBM SPSS Statistics* コマンド・シンタックス・リファレンス」ガイドを参照してください。「シンタックス」タブでガイドを表示するには、「シンタックス エディター」オプションを選択し、「IBM SPSS Statistics シンタックスのヘルプを起動」ボタンをクリックします。

Statistics 出力ノードの「シンタックス」タブ

このタブを使用して、データの分析に使用する SPSS Statistics 手続きのシンタックスを作成します。シンタックスは、次の 2 つのパートで構成されています。文とそれに関連するオプションです。文には、実行する分析または操作と使用するフィールドを指定します。オプションには、表示する統計量、保存する生成フィールドなど、その他のすべてを指定します。

SPSS Statistics ダイアログ・オプション

IBM SPSS Statistics シンタックスについてよく知らない場合、IBM SPSS Modeler でシンタックスを作成する最も簡単な方法は、「**IBM SPSS Statistics** ダイアログ」 オプションを選択し、手順のダイアログ・ボックスを選択、ダイアログボックスを入力して「OK」をクリックします。その後、IBM SPSS Modeler で使用している IBM SPSS Statistics ノードの「シンタックス」タブにシンタックスが配置されます。その後、ストリームを実行して手順より出力を取得することができます。

オプションで、Statistics ファイル・ソース・ノードを生成し、データをインポートすることができます。これは例えば、出力を表示するほか、スコアなどのフィールドをアクティブ データ・セットに書き込む場合に役立ちます。

注:

- 英語以外の言語で出力を生成するときは、シンタックスでその言語を指定することをお勧めします。
- Statistics 出力ノードでは、「出力スタイル」オプションはサポートされていません。

シンタックスを作成するには、以下を実行します。

1. 「ダイアログを選択」 ボタンをクリックします。
2. 次のいずれかのオプションを選択してください。
   - 分析: SPSS Statistics の「分析」メニューの内容がリストされます。使用する手続きを選択します。
   - その他: このオプションが表示される場合は、SPSS Statistics のカスタム・ダイアログ・ビルダーで作成されたダイアログと、「分析」メニューに表示されず、ライセンスが付与されているその他の SPSS Statistics ダイアログがリスト表示されます。該当するダイアログがない場合、このオプションは表示されません。

注: 「自動データ準備」ダイアログは表示されません。

新しいフィールドを作成する SPSS Statistics カスタム・ダイアログがある場合、Statistics 出力ノードはターミナル・ノードであるため、これらのフィールドを SPSS Modeler で使用することはできません。

任意で 「結果のデータのインポート ノードを生成する」 ボックスをオンにして、生成データを別のストリームにインポートできる Statistics ファイル・ソース・ノードを作成します。このノードは画面の領域上に表示され、データは 「ファイル」フィールドで指定された .sav ファイルに含まれます (デフォルトの場所は SPSS Modeler インストール・ディレクトリーです)。

シンタックス・エディターのオプション

度数を使用する手続きに作成したシンタックスを保存する手順は、次のとおりです。

1. ツールバーの最初のボタン「ファイル オプション」ボタンをクリックします。
2. メニューから 「保存」 または 「名前を付けて保存」 を選択します。
3. ファイルを ..sps ファイルの形式で保存します。

以前作成したシンタックス・ファイルを使用するには、シンタックス・エディターの現在の内容を置き換えます。

1. ツールバーの最初のボタン「ファイル オプション」ボタンをクリックします。
2. メニューから 「開く」 を選択します。
3. .sps ファイルを選択すると、その内容が出力ノードの「シンタックス」タブに貼り付けられます。

現在の内容を置き換えずに以前保存したシンタックスを挿入する手順は、次のとおりです。

1. ツールバーの最初のボタン「ファイル オプション」ボタンをクリックします。
2. メニューから 「挿入」 を選択します。
3. .sps ファイルを選択すると、その内容が出力ノードのカーソルで指定されたポイントに貼り付けられます。

任意で 「結果のデータのインポート ノードを生成する」 ボックスをオンにして、生成データを別のストリームにインポートできる Statistics ファイル・ソース・ノードを作成します。このノードは画面の領域上に表示され、データは 「ファイル」フィールドで指定された .sav ファイルに含まれます (デフォルトの場所は SPSS Modeler インストール・ディレクトリーです)。

「実行」 をクリックすると、結果が SPSS Statistics 出力ビューアーに表示されます。このビューアーの詳細については、SPSS Statistics のマニュアルを参照してください。

注: 次の項目 (およびこれらの項目に対応する SPSS Statistics のダイアログ・ボックスのオプション) の構文はサポートされません。これらは出力に影響を与えません。

- OUTPUT ACTIVATE
- OUTPUT CLOSE
- OUTPUT DISPLAY
- OUTPUT EXPORT
- OUTPUT MODIFY
- OUTPUT NAME
- OUTPUT NEW
- OUTPUT OPEN
- OUTPUT SAVE

Statistics 出力ノードの「出力」タブ

「出力」タブを使用すると、出力形式や位置を指定することができます。結果をスクリーンに表示、または結果を使用可能なファイル形式の 1 つに送ることができます。

出力名。ノードの実行時に生成される出力の名前を指定します。「自動」 は、出力を生成するノードの名前に基づいて名前を選択します。「ユーザー設定」 で別の名前を指定することもできます。

画面に出力 (デフォルト)。オンラインで表示するための出力を作成します。出力ノードを実行すると、出力オブジェクトはマネージャー・ウィンドウの「出力」タブに表示されます。

ファイルに出力。ノードの実行時に、出力をファイルに保存します。このオプションを選択した場合は、「ファイル名」 フィールドにファイル名を入力して (または、ファイル選択ボタンを使用してディレクトリーを参照し、ファイル名を指定して)、ファイル形式を選択してください。

ファイルの種類。出力を送信するファイルの種類を選択します。

- **HTML** ドキュメント **(*.html)**。HTML 形式で出力を書き込みます。
- **IBM SPSS Statistics** ビューアー・ファイル **(*.spv)**： 出力を IBM SPSS Statistics 出力ビューアーで読み取れる形式で書き込みます。
- **IBM SPSS Statistics Web** レポート・ファイル **(*.spw):** IBM SPSS Statistics Web レポートの形式で出力を書き込みます。この出力は、IBM SPSS Collaboration and Deployment Services のリポジトリーに公開してから、Web ブラウザーで表示することができます。詳しくは、トピック 323 ページの『Web に公開』を参照してください。

注: 「画面に出力」を選択すると、IBM SPSS Statistics OMS ディレクティブの VIEWER=NO が無効になります。また、スクリプト API (*Basic* および *Python SpssClient* モジュール) を IBM SPSS Modeler で使用することもできません。

Statistics エクスポート・ノード

Statistics エクスポート・ノードを使用すると、IBM SPSS Statistics の *.sav* 形式でデータをエクスポートすることができます。IBM SPSS Statistics *.sav* ファイルは、IBM SPSS Statistics Base およびその他のモジュールで読み込むことができます。この形式は、IBM SPSS Modeler キャッシュ・ファイルでも使用されます。

IBM SPSS Statistics の変数名は 64 文字までに制限されており、スペース、ドル記号 ($)、ダッシュ (-) など一部の文字を使用できないため、IBM SPSS Modeler のフィールド名を IBM SPSS Statistics の変数名にマップするとエラーが発生することがあります。この問題に対処するには、次の 2 通りの方法があります。

- 「フィルター」タブをクリックして、IBM SPSS Statistics 変数名の要件に準拠したフィールド名に変更することができます。 詳しくは、トピック 388 ページの『IBM SPSS Statistics 用のフィールドの名前変更またはフィルタリング』を参照してください。
- IBM SPSS Modeler でフィールド名とラベルをエクスポートします。

注: IBM SPSS Modeler は、*.sav* ファイルを Unicode の UTF-8 形式で書き込みます。リリース 16.0 以降の IBM SPSS Statistics でサポートしているのは Unicode UTF-8 形式だけです。データの破損の可能性を回避するために、Unicode エンコードで保存した *.sav* ファイルはリリース 16.0 以前の IBM SPSS Statistics で使用することはできません。詳細は、IBM SPSS Statistics のヘルプを参照してください。

複数回答設定: ストリームに定義された複数の回答セットは、ファイルがエクスポートされると自動的に保存されます。「フィルター」タブで、ノードの複数の回答セットを表示および編集することができます。詳しくは、トピック 161 ページの『複数回答セット編集』を参照してください。

Statistics エクスポート・ノードの「エクスポート」タブ

エクスポート・ファイル: ファイルの名前を指定します。ファイル名を入力するか、またはファイル選択ボタンをクリックしてファイルの場所を指定します。

ファイルの種類: ファイルを通常の *.sav* または圧縮の *.zsav* のいずれの形式で保存するかを選択します。

パスワードでファイルを暗号化: パスワードを使用してファイルを保護するには、このボックスを選択します。別のダイアログ・ボックスで「パスワード」の入力および確認を要求するプロンプトが出されます。

注: パスワード保護されたファイルを開くことができるのは、SPSS Modeler バージョン 16 以降、または SPSS Statistics バージョン 21 以降のみです。

フィールド名のエクスポート: SPSS Modeler から SPSS Statistics の *.sav* または *.zsav* ファイルにエクスポートする場合の変数名とラベルの処理方法を指定します。

- 「名前と変数ラベル」 SPSS Modeler のフィールド名とフィールド・ラベルの両方をエクスポートする場合に選択します。名前は SPSS Statistics の変数名としてエクスポートされ、ラベルは SPSS Statistics の変数ラベルとしてエクスポートされます。
- 「変数ラベルとして使用」 SPSS Modeler のフィールド名を SPSS Statistics の変数ラベルとして使用する場合に選択します。SPSS Modeler は SPSS Statistics 変数名では無効であるフィールド名の文字を使用できます。無効な SPSS Statistics 名を作成しないように、「変数ラベルとして使用」 を選択するか、フィールド名を調整するための「フィルター」タブを利用します。

アプリケーションの起動: SPSS Statistics がコンピューターにインストールされている場合、このオプションを選択することにより、保存したデータ・ファイルに対してこのアプリケーションを直接起動できます。

アプリケーションを起動するためのオプションは、「ヘルパー アプリケーション」ダイアログ・ボックスで指定する必要があります。詳しくは、トピック 366 ページの『IBM SPSS Statistics ヘルパー アプリケーション』を参照してください。外部プログラムで起動しないで、単に SPSS Statistics *.sav* または *.zsav* ファイルを作成する場合は、このオプションの選択を解除してください。

注**:** SPSS Modeler と SPSS Statistics をサーバー (分散) モードで一緒に実行すると、データを書き込んで SPSS Statistics セッションを起動しても、SPSS Statistics クライアントが自動的に開かないため、アクティブ・データ・セットに読み込まれるデータ・セットが表示されません。この問題を回避するには、SPSS Statistics クライアントを起動して、手動でデータ・ファイルを開きます。

データのインポート ノードを生成**:** エクスポートされたデータ・ファイルを読み込む、Statistics ファイル・ソース・ノードを自動的に生成する場合に選択します。詳しくは、トピック 34 ページの『Statistics ファイル・ノード』を参照してください。

IBM SPSS Statistics 用のフィールドの名前変更またはフィルタリング

IBM SPSS Modeler から IBM SPSS Statistics などの外部アプリケーションにデータをエクスポートまたは展開する前に、フィールド名を変更したり、調節しなければならないこともあります。Statistics 変換、Statistics 出力、および Statistics エクスポート ダイアログ・ボックスには、「フィルター」タブが用意されており、ここからこの処理を簡単に行うことができます。

「フィルター」タブの機能に関する基本的な説明は、他の場所で説明されています。詳しくは、 159 ページの『フィルタリング・オプションの設定』を参照してください。ここでは、IBM SPSS Statistics にデータを読み込む場合のヒントについて説明します。

IBM SPSS Statistics の命名規則に準ずるよう、ファイル名を調整する手順は次のとおりです。

1. 「フィルター」タブで、「フィルター・オプション・メニュー」ツールバー・ボタン (ツールバーの最初のボタン) をクリックします。
2. 「IBM SPSS Statistics 用に名前変更」を選択します。
3. 「IBM SPSS Statistics 用に名前変更」ダイアログで、ファイル名の無効な文字を「ハッシュ **(#)**」文字または「下線 **(_)**」のいずれに置き換えるかを選択することができます。

複数回答セットの名前を変更: Statistics ファイル入力ノードを使用して、IBM SPSS Modeler にインポートできる複数回答セットの名前を変更する場合、このオプションを選択します。調査の回答など、ケースごとに複数の値があるデータを記録するのに使用されます。

第 9 章 スーパーノード

スーパーノードの概要

IBM SPSS Modeler の視覚的なプログラミング インターフェースは非常に簡単に習得できますが、その理由の 1 つとして、各ノードの機能が明確に定義されている点が挙げられます。しかし、複雑な処理を行うには、長い一連のノードが必要となることがあります。この結果、ストリーム領域が複雑になってストリーム・ダイアグラムの追跡が困難になることがあります。ストリームが長く複雑になるのを防ぐには、次の 2 つの方法があります。

- 処理シーケンスを複数のストリームに分割して、1 つのストリームが別のストリームに送られるようにします。例えば、1 番目のストリームがデータ・ファイルを作成し、それを 2 番目のストリームで入力ファイルとして使用します。次に 2 番目のストリームがデータ・ファイルを作成し、3 番目のストリームがそれを入力ファイルとして使用します。これらの複数のスクリプトをプロジェクトに保存して管理することができます。プロジェクトは、複数のストリームや出力の編成手段を提供しています。しかし、プロジェクト・ファイルには、それに含まれるオブジェクトの参照だけが格納されているため、依然として複数のストリーム・ファイルを管理する必要があります。
- 複雑なストリームで作業を行う際のより効率的な方法として、スーパーノードを作成することできます。

スーパーノードが、データ・ストリームのセクションをカプセル化し、複数のノードを 1 つのノードにグループ化します。この機能を使用することで、データ・マイニングには多くの利点が生まれます。

- ストリームが簡潔で管理しやすくなります。
- ノードを組み合わせて、そのビジネスに固有のスーパーノードを作成できます。
- 複数のデータ・マイニング・プロジェクトで再利用するため、スーパーノードをライブラリーにエクスポートすることができます。

スーパーノードの種類

スーパーノードは、データ・ストリームでは星形のアイコンで表されます。アイコンの影によって、スーパーノードの種類とそのストリームの入出力方向が示されます。

スーパーノードには、次の 3 種類があります。

- 入力スーパーノード
- プロセス スーパーノード
- ターミナル・スーパーノード

入力スーパーノード

入力スーパーノードには、通常のソース・ノードと同じようなデータ・ソースが含まれており、通常のソース・ノードを使用できる任意の場所で使用できます。入力スーパーノードのアイコンには左側に影が付けられています。これは、左側が「閉じられている」ことを表しており、データはスーパーノードから 常に下流に流れることを示しています。

入力スーパーノードの右側には 1 つだけ接続があり、スーパーノードからデータが出力されストリームに流れていくことを表しています。

プロセス スーパーノード

プロセス・スーパーノードに含めることができるのはプロセス・ノードだけで、影は表示されません。これは、このタイプのスーパーノードに対して、データが入力 と出力 の両方向に流れることができることを示しています。

プロセス スーパーノードには、左側と右側の両方に接続点があり、データはスーパーノードに入力された後、出力されて下流のストリームに流れることを表しています。スーパーノードにはストリーム・フラグメントを追加したり、余分なストリームを入れることもできますが、両方の接続点はインとアウトを接続する単一のパス上になければなりません。

注 : プロセス スーパーノードは、「操作スーパーノード」と呼ばれることもあります。

ターミナル・スーパーノード

ターミナル・スーパーノードには、1 つ以上のターミナル・ノード (散布図、テーブルなど) を入れることができ、ターミナル・ノードと同じように使用することができます。ターミナル・スーパーノードの右側には影が表示されます。これは、ノードの右側が「閉じて」いて、ターミナル・スーパーノードへのデータの入力 だけが可能であることを示しています。

ターミナル・スーパーノードの左側には 1 つだけ接続点があり、データはストリームからスーパーノードに入力され、スーパーノード内で処理が終了することを表しています。

ターミナル・スーパーノードには、スーパーノード内のターミナル・ノードの実行順序を指定するためのスクリプトを入れることもできます。 詳しくは、トピック 420 ページの『スーパーノードとスクリプト』を参照してください。

スーパーノードの作成

スーパーノードを作成すると、単一のノード中に複数のノードがカプセル化されるため、データ・ストリームが「縮小」されます。ストリーム領域上にストリームを作成またはロードしたら、さまざまな方法でスーパーノードを作成することができます。

複数選択

スーパーノードを作成するもっとも簡単な方法は、カプセル化するすべてのノードを選択することです。

1. マウスを使用してストリーム領域上の複数のノードを選択します。Shift キーを押しながらクリックして、ストリームまたはストリームの一部を選択することもできます。

   注**:** 選択する複数のノードは連続したストリームまたは分岐したストリームでなければなりません。隣接しないまたは接続されていない複数のノードを選択することはできません。
2. 次のいずれかの方法で、選択したノードをカプセル化します。
   - ツールバーのスーパーノード・アイコン (星形) をクリックします。
   - スーパーノードを右クリックして表示されるコンテキスト・メニューから、次の各項目を選択します。

     「スーパーノードの作成」 > 「選択項目から」
   - 「スーパーノード」メニューから次の各項目を選択します。

     「スーパーノードの作成」 > 「選択項目から」

これらの 3 種類のオプションはすべて、ノードをスーパーノードにカプセル化して、その内容に応じた種類 (入力、プロセス、またはターミナル) を表す影を付けます。

単一選択

単一のノードを選択して、メニュー・オプションからスーパーノードの始点と終点を指定するか、または選択したノードの下流にあるすべてのノードをカプセル化して、スーパーノードを作成することもできます。

1. カプセル化を開始する始点となるノードをクリックします。
2. 「スーパーノード」メニューから次の各項目を選択します。

   「スーパーノードの作成」 > 「ここから実行」

カプセル化するストリーム中のノードの始点と終点を選択して、より対話的にスーパーノードを作成することもできます。

1. スーパーノードに入れる最初または最後のノードをクリックします。
2. 「スーパーノード」メニューから次の各項目を選択します。

   「スーパーノードの作成」 > 「選択...」
3. 代わりに、目的のノードを右クリックして表示されるコンテキスト・メニューを使用することもできます。
4. カーソルがスーパーノード・アイコンの形に変わります。これは、ストリーム中のもう一方のノードを選択する必要があることを示します。スーパーノード・フラグメントの「もう一方の端」(上流または下流) に移動して、適切なノードをクリックします。選択した 2 つのノード間にある、すべてのノードがスーパーノードの星形アイコンに置き換わります。

注: 選択する複数のノードは連続したストリームまたは分岐したストリームでなければなりません。隣接しないまたは接続されていない複数のノードを選択することはできません。

スーパーノードのネスト

スーパーノードを他のスーパーノードに入れる (ネストする) ことができます。ネストされたスーパーノードには、それぞれの種類のスーパーノード (入力、プロセス、およびターミナル) と同じ規則が適用されます。例えば、ネスティングしているプロセス スーパーノード中のすべてのスーパーノードを連続してデータが流れていないと、そのスーパーノードはプロセス スーパーノードを維持できません。ネストされているスーパーノードのいずれかがターミナル・スーパーノードだった場合、データがその階層中を連続して流れることができません。

ターミナル・スーパーノードと入力スーパーノードは、他の種類のスーパーノードをネストすることができます。ただし、同じようにスーパーノードを作成する際の基本的な規則が適用されます。

スーパーノードのロック

スーパーノードを作成すると、パスワードを使用してスーパーノードをロックし、修正されるのを防ぎます。例えば、IBM SPSS Modeler の質問を設定する経験があまりない組織内のユーザーが使用できるよう、固定値のテンプレートとしてストリームまたはストリームの一部を作成する場合に、スーパーノードを作成します。

スーパーノードがロックされている場合も、定義されたパラメーターの「パラメーター」タブに値を入力したり、パスワードを入力せずにロックされたスーパーノードを実行することができます。

注**:** ロックおよびロックの解除は、スクリプトを使用して実行することはできません。

スーパーノードのロックとロック解除

注**:** 消失したパスワードを復元することはできません。

3 つのタブのうちどのタブからもスーパーノードをロックまたはロック解除できます。

1. 「ノードのロック」 をクリックします。
2. パスワードを入力して確定します。
3. 「**OK**」をクリックします。

パスワード保護されたスーパーノードは、ストリーム領域のスーパーノードのアイコンの左上に小さい南京錠のシンボルで示されます。

スーパーノードのロック解除

1. パスワード保護を永続的に解除するには、「ノードのロック解除」 をクリックします。パスワードの入力を求めるプロンプトが出されます。
2. パスワードを入力して 「**OK**」 をクリックします。スーパーノードはパスワード保護されず、ストリーム内のアイコンの隣に南京錠のシンボルは表示されなくなります。

SPSS Modeler バージョン 16 から 17.0 の間に保存された、スーパーノードを含むストリームの場合、IBM SPSS Collaboration and Deployment Services や Mac などの SPSS Modeler によりインストールされた JRE が異なる別の環境でそのストリームを開くには、そのストリームが最後に保存された古い環境で、バージョン 17.1 以降を使用してストリームを開き、ロックを解除し、保存し直す必要があります。

バージョン 18 よりも古いストリーム内のスーパーノードのロックを解除する際に、パスワードの誤りエラーが表示される場合があります。これを回避するには、そのストリームが最後に保存されたときと同じシステム・ローカル設定で、同じプラットフォーム上で、同一の IBM SPSS Modeler バージョン (またはより新しいバージョン) を使用して、ノードを再オープンし、ロックを解除します。その後、バージョン 18 以降でノードを開き、ノードをロックして、ストリームを再度保存します。

ロックされたスーパーノードの編集

パラメーターを定義するか、ズーム・インしてロックされたスーパーノードを表示する場合、パスワード入力を要求するプロンプトが表示されます。

パスワードを入力して 「**OK**」 をクリックします。

スーパーノードのあるストリームが終了するまで必要に応じて、パラメーター定義を編集したりズーム・インやズーム・アウトしたりできます。

これでパスワード保護を解除したわけではなく、スーパーノードを使用できるようアクセスできるにすぎません。 詳しくは、トピック『スーパーノードのロックとロック解除』を参照してください。

スーパーノードの編集

スーパーノードを作成すると、ズーム・インすることによってより近くでスーパーノードを検証することができます。スーパーノードがロックされている場合は、パスワードを要求するプロンプトが表示されます。詳しくは、トピック『ロックされたスーパーノードの編集』を参照してください。

スーパーノードの内容を表示するには、IBM SPSS Modeler ツールバーのズーム・イン・アイコンを使用するか、または次の方法を利用します。

1. スーパーノードを右クリックします。
2. コンテキスト・メニューから、「ズーム・イン」を選択します。

選択したスーパーノードの内容が、ストリームまたはストリーム・フラグメントを流れるデータを表すコネクタとともに、少し異なる IBM SPSS Modeler 環境に表示されます。ストリーム・キャンバスのこのレベルでは、以下のような作業を行うことができます。

- スーパーノードの種類 (入力、プロセス、またはターミナル) の変更。
- パラメーターの作成またはパラメーターの値の編集。パラメーターはスクリプトや CLEM 式中で使用されます。
- スーパーノードおよびそのサブ ノードのキャッシュ・オプションの指定。
- スーパーノード スクリプトの作成と変更 (ターミナル・スーパーノードの場合)。

スーパーノードの種類の変更

状況によっては、スーパーノードの種類を変更する方が良いこともあります。このオプションは、スーパーノードにズーム・インしている場合にだけ利用でき、このレベルのスーパーノードにだけ適用されます。3 種類のスーパーノードについて次の表で説明します。

表 *51.* スーパーノードの種類：

スーパーノードの種類	説明
ソース・スーパーノード	1 つの出力接続があります。
プロセス・スーパーノード	2 個の接続 :入力接続と出力接続の 2 つの接続があります。
ターミナル・スーパーノード	1 つの入力接続があります。

スーパーノードの種類を変更するには、以下を実行します。

1. スーパーノードにズーム・インしていることを確認してください。
2. 「スーパーノード」メニューの 「スーパーノードの種類」を選択し、次に適切な種類を選択します。

スーパーノードの注釈付けと名前の変更

ストリームに表示されるスーパーノードの名前を変更したり、プロジェクトやレポートで使われる注釈を付けることができます。これらの作業を行うには、次の手順に従ってください。

- スーパーノードを右クリックして (ズーム・アウトされる)、「名前の変更と注釈」を選択します。
- 代わりに、「スーパーノード」メニューの 「名前の変更と注釈」を選択することもできます。このオプションは、ズーム・インおよびズーム・アウトの両方のモードで利用できます。

どちらの場合でも、「注釈」タブが選択された状態でダイアログ・ボックスが表示されます。このタブのオプションを使用して、ストリーム領域上に表示される名前を変更したり、スーパーノードの操作に関する説明や情報などを入力します。

スーパーノードによるコメントの使用

コメントが付いているノードまたはナゲットからスーパーノードを作成する場合、スーパーノードにコメントを表示するには、スーパーノードを作成するための選択内容にそのコメントを含める必要があります。選択内容でコメントを省略すると、スーパーノードの作成時にコメントがストリーム内に残ったままになります。

コメントを含むスーパーノードを拡張する場合、スーパーノードが作成される前にコメントを元の場所に戻します。

コメント付きオブジェクトを含むスーパーノードを拡張したにもかかわらず、スーパーノードにコメントが含まれない場合、オブジェクトは元の場所に戻されますが、コメントは再接続されません。

スーパーノードのパラメーター

IBM SPSS Modeler では、Minvalue のようなユーザー独自の変数を設定することができます。変数の値は、スクリプトまたは CLEM 式で指定できます。これらの変数は、パラメーターと呼ばれます。パラメーターは、ストリーム、セッション、およびスーパーノードに対して設定することができます。スーパーノードに対して設定されたパラメーターは、スーパーノードまたはスーパーノードにネストされているノードで CLEM 式を作成する際に利用できます。ネストされたスーパーノードに対して設定されたパラメーターを、その親スーパーノードで利用することはできません。

スーパーノードのパラメーターを作成して設定するには、次の 2 つのステップがあります。

1. スーパーノードのパラメーターを定義します。
2. 次に、スーパーノードの各パラメーターの値を設定します。

これらのパラメーターは、カプセル化された任意のノードの CLEM 式中で利用できます。

スーパーノードのパラメーターの定義

スーパーノードのパラメーターは、ズーム・イン・モードとズーム・アウト・モードのどちらでも定義することができます。定義されたパラメーターは、カプセル化されたすべてのノードに適用されます。スーパーノードのパラメーターを定義するには、まず「スーパーノード」ダイアログ・ボックスの「パラメーター」タブに移動する必要があります。次のいずれかの方法で、ダイアログ・ボックスを表示してください。

- ストリーム中のスーパーノードをダブルクリックする。
- 「スーパーノード」メニューから「パラメーター設定」を選択する。
- スーパーノードにズーム・インしている場合に、コンテキスト・メニューから「パラメーター設定」を選択する。

ダイアログ・ボックスを開くと、今までに定義されているパラメーターが「パラメーター」タブに表示されます。

新しいパラメーターを定義するには

「パラメーターの定義」ボタンをクリックして、ダイアログ・ボックスを開きます。

「名前」。ここにはパラメーター名が表示されます。新しくパラメーターを作成するには、このフィールドに名前を入力します。例えば、最低気温を表すパラメーターを作成する場合に、minvalue と入力することができます。CLEM 式内でパラメーターを示す接頭辞の $P- を付けないようにしてください。ここで指定した名前は、CLEM 式ビルダーにも表示されます。

ロング ネーム **:** 作成したパラメーターを説明する名前が表示されます。

ストレージ。 リストからストレージ・タイプを選択します。ストレージで、データ値がパラメーター内にどのように格納されるかを示します。例えば、「008」のように先頭に 0 がある値を扱う場合に、その 0 を保持する必要があるならば、ストレージ・タイプとして 「文字列」 を選択する必要があります。選択し

ないと、値から 0 が除去されます。ストレージ・タイプとしては、文字列、整数、実数、時間、日付、またはタイムスタンプを利用できます。日付のパラメーターには、次の段落で示す ISO 規格の表記を使用して値を指定する必要があります。

値: 各パラメーターの現在の値が表示されます。必要に応じてパラメーターを調整してください。日付のパラメーターには、ISO 規格の表記 (つまり、YYYY-MM-DD) を使用して値を指定する必要があります。他の形式で指定された日付は受け入れられません。

データ型 **(**オプション**) :** ストリームを外部アプリケーションに展開する場合は、使用する測定の尺度をリストから選択します。それ以外の場合は、データ型の欄はそのままにしておくことをお勧めします。数値範囲の上限および下限など、パラメーターに値の制約を指定したい場合、リストから 「指定」 を選択します。

ロング ネーム、ストレージ、およびデータ型のオプションは、ユーザー・インターフェースを通じてだけ、パラメーターに設定できます。これらのオプションは、スクリプトを使用して設定できません。

右にある矢印をクリックして、選択したパラメーターを使用可能なパラメーターのリストの上または下に移動することができます。選択したパラメーターを削除するには、削除ボタン (*X* マーク) を使用します。

スーパーノード・パラメーターの値の設定

スーパーノードのパラメーターを定義したら、CLEM 式やスクリプトを使用してパラメーターに値を指定することができます。

スーパーノードのパラメーターを指定するには

1. スーパーノードのアイコンをダブルクリックして、「スーパーノード」ダイアログ・ボックスを開きます。
2. 「スーパーノード」メニューから 「パラメーター設定」を選択することもできます。
3. 「パラメーター」タブをクリックします。注 : このダイアログ・ボックス中のフィールドは、このタブの「パラメーターの定義」ボタンをクリックして定義されたフィールドです。
4. 作成した各パラメーターの値を、テキスト・ボックスに入力します。例えば、 *minvalue* に特定のしきい値を設定することができます。値を設定したパラメーターは、例えばこのしきい値以上または以下のレコードを選択するなど、さまざまな操作に利用することができます。

スーパーノード・パラメーターを使ったノードのプロパティーへのアクセス

スーパーノード・パラメーターは、カプセル化されているノードのプロパティー (スロット・パラメーターとも呼ばれます) を定義するために使用することもできます。例えば、無作為のデータ・サンプルを使用して、カプセル化されたニューラル・ノードを一定時間に渡って学習するようにスーパーノードを設定する場合を考えてみましょう。パラメーターを使用して、学習時間の長さやサンプルの割合を指定することができます。

サンプルのスーパーノードに、*Sample* というサンプル・ノードと、*Train* というニューラル・ネットワーク・ノードが含まれているとします。ノードのダイアログ・ボックスを使用して、サンプル・ノードの 「サンプル」を「無作為 **%**」 に、ニューラル・ノードの 「停止条件」を 「時間」に設定します。これらのオプションを指定したら、パラメーターを使用してノードのプロパティーにアクセスし、スーパーノードに関する特定の値を指定することができます。「スーパーノード」ダイアログ・ボックスで「パラメーターの定義」をクリックし、次の表に示すパラメーターを作成します。

表 52. 作成するパラメーター

パラメーター	値	長形式名
Train.time	5	学習時間 (分)
Sample.random	10	無作為サンプルのパーセンテージ

注 : *Sample.random* のようなパラメーター名は、ノードのプロパティーを参照するために正しいシンタックスを使用しています。ここで、*Sample* はノード名を、*random* はノードのプロパティーを表しています。

これらのパラメーターを定義したら、各ノードのダイアログ・ボックスを開くことなく、サンプル・ノードとニューラル・ノードのプロパティーの値を簡単に変更することができます。代わりに、単に「スーパーノード」メニューの 「パラメーター設定」を選択して、「スーパーノード」ダイアログ・ボックスの「パラメーター」タブを表示し、「無作為 **%**] と 「時間] に新しい値を指定することもできます。この方法は、モデル構築を何回も繰り返してデータを探索しているような場合に役立ちます。

スーパーノードとキャッシュ

スーパーノードないでは、ターミナル・ノードを除くすべてのノードでキャッシュを設定することができます。キャッシュを制御するには、ノードを右クリックして「キャッシュ」コンテキスト・メニューから、適切なオプションを選択します。このメニュー・オプションは、スーパーノード外から、およびスーパーノード内のカプセル化されたノードで利用することができます。

次に、スーバーノードのキャッシュに関するガイドラインをいくつか示します。

- スーパーノードでカプセル化されたノードのキャッシュが有効になっている場合は、そのスーパーノードもキャッシュが有効になります。
- スーパーノードのキャッシュを無効にすると、カプセル化されているすべての ノードのキャッシュが無効になります。
- スーパーノードのキャッシュを有効にすると、キャッシュ可能な最後のサブノードを実際にキャッシュできるようになります。つまり、最後のサブノードが条件抽出ノードならば、その条件抽出ノードのキャッシュが有効になります。また、最後のサブノードが (キャッシュできない) ターミナル・ノードの場合は、キャッシュをサポートする次の上流ノードのキャッシュが有効になります。
- スーパーノードのサブノードにキャッシュを設定すると、キャッシュされるノードからの上流操作 (ノードの追加や編集など) により、キャッシュが取り消されます。

スーパーノードとスクリプト

SPSS Modeler のスクリプト言語を使用すると、ターミナル・スーパーノードの内容を操作、実行する簡単なプログラムを作成できます。例えば、複雑なストリームの実行順序を指定することができます。散布図ノードの前に実行する必要があるグローバル・ノードがスーパーノード中にある場合は、グローバル・ノードを先に実行するスクリプトを作成することができます。このノードが計算する平均や標準偏差などの値は、散布図ノードを実行するときに使用します。

「スーパーノード」ダイアログ・ボックスの「スクリプト」タブは、ターミナル・スーパーノードの場合にだけ利用することができます。

ターミナル・スーパーノードのスクリプト ダイアログ・ボックスを表示するには

- スーパーノード領域を右クリックして、「スーパーノード スクリプト」を選択します。

- 代わりに、ズーム・イン・モードおよびズーム・アウト・モードの両方で、「スーパーノード」メニューの 「スーパーノード スクリプト」を選択することもできます。

注: スーパーノード スクリプトは、ダイアログ ボックスで「このスクリプトを実行」を選択した場合にだけ、ストリームまたはスーパーノードとともに実行されます。

スクリプト処理に固有のオプションと SPSS Modeler 内での使用については、「スクリプトと自動化のガイド」(製品ダウンロードの一部として PDF ファイルで入手可能) を参照してください。

スーパーノードの保存とロード

スーパーノードの利点の 1 つとして、保存して他のストリームで再使用できることがあげられます。スーパーノードの保存やロードには、拡張子 `.slb` を使用することに注意してください。

スーパーノードを保存するには、以下を実行します。

1. スーパーノードをズーム・インします。
2. 「スーパーノード」メニューから「スーパーノードの保存」を選択します。
3. ダイアログ・ボックスで、フィールド名とディレクトリーを指定します。
4. 保存したスーパーノードを、現在のプロジェクトに追加するかどうかを選択します。
5. 「保存」をクリックします。

スーパーノードをロードするには、以下を実行します。

1. IBM SPSS Modeler ウィンドウの「挿入」メニューから、「スーパーノード」を選択します。
2. 現在のディレクトリーからスーパーノード・ファイル (`.slb`) を選択するか、別のディレクトリーのファイルを指定します。
3. 「ロード」をクリックします。

注: インポートしたスーパーノードのパラメーターには、すべてデフォルト値があります。パラメーターを変更する場合は、ストリーム領域でスーパーノードをダブルクリックしてください。

特記事項

本書は米国 IBM が提供する製品およびサービスについて作成したものです。この資料は、IBM から他の言語でも提供されている可能性があります。ただし、これを入手するには、本製品または当該言語版製品を所有している必要がある場合があります。

本書に記載の製品、サービス、または機能が日本においては提供されていない場合があります。日本で利用可能な製品、サービス、および機能については、日本 IBM の営業担当員にお尋ねください。本書で IBM 製品、プログラム、またはサービスに言及していても、その IBM 製品、プログラム、またはサービスのみが使用可能であることを意味するものではありません。これらに代えて、IBM の知的所有権を侵害することのない、機能的に同等の製品、プログラム、またはサービスを使用することができます。ただし、IBM 以外の製品とプログラムの操作またはサービスの評価および検証は、お客様の責任で行っていただきます。

IBM は、本書に記載されている内容に関して特許権 (特許出願中のものを含む) を保有している場合があります。本書の提供は、お客様にこれらの特許権について実施権を許諾することを意味するものではありません。実施権についてのお問い合わせは、書面にて下記宛先にお送りください。

〒*103-8510*
東京都中央区日本橋箱崎町*19*番*21*号
日本アイ・ビー・エム株式会社
法務・知的財産
知的財産権ライセンス渉外

IBM およびその直接または間接の子会社は、本書を特定物として現存するままの状態で提供し、商品性の保証、特定目的適合性の保証および法律上の瑕疵担保責任を含むすべての明示もしくは黙示の保証責任を負わないものとします。国または地域によっては、法律の強行規定により、保証責任の制限が禁じられる場合、強行規定の制限を受けるものとします。

この情報には、技術的に不適切な記述や誤植を含む場合があります。本書は定期的に見直され、必要な変更は本書の次版に組み込まれます。IBM は予告なしに、随時、この文書に記載されている製品またはプログラムに対して、改良または変更を行うことがあります。

本書において IBM 以外の Web サイトに言及している場合がありますが、便宜のため記載しただけであり、決してそれらの Web サイトを推奨するものではありません。それらの Web サイトにある資料は、この IBM 製品の資料の一部ではありません。それらの Web サイトは、お客様の責任でご使用ください。

IBM は、お客様が提供するいかなる情報も、お客様に対してなんら義務も負うことのない、自ら適切と信ずる方法で、使用もしくは配布することができるものとします。

本プログラムのライセンス保持者で、(i) 独自に作成したプログラムとその他のプログラム (本プログラムを含む) との間での情報交換、および (ii) 交換された情報の相互利用を可能にすることを目的として、本プログラムに関する情報を必要とする方は、下記に連絡してください。

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

本プログラムに関する上記の情報は、適切な使用条件の下で使用することができますが、有償の場合もあります。

本書で説明されているライセンス・プログラムまたはその他のライセンス資料は、IBM 所定のプログラム契約の契約条項、IBM プログラムのご使用条件、またはそれと同等の条項に基づいて、IBM より提供されます。

記載されている性能データとお客様事例は、例として示す目的でのみ提供されています。実際の結果は特定の構成や稼働条件によって異なります。

IBM 以外の製品に関する情報は、その製品の供給者、出版物、もしくはその他の公に利用可能なソースから入手したものです。IBM は、それらの製品のテストは行っておりません。したがって、他社製品に関する実行性、互換性、またはその他の要求については確証できません。 IBM 以外の製品の性能に関する質問は、それらの製品の供給者にお願いします。

IBM の将来の方向性および指針に関する記述は、予告なく変更または撤回される場合があります。これらは目標および目的を提示するものにすぎません。

本書には、日常の業務処理で用いられるデータや報告書の例が含まれています。より具体性を与えるために、それらの例には、個人、企業、ブランド、あるいは製品などの名前が含まれている場合があります。これらの名称はすべて架空のものであり、類似する個人や企業が実在しているとしても、それは偶然にすぎません。

商標

IBM、IBM ロゴおよび ibm.com は、世界の多くの国で登録された International Business Machines Corporation の商標です。他の製品名およびサービス名等は、それぞれ IBM または各社の商標である場合があります。現時点での IBM の商標リストについては、http://www.ibm.com/legal/copytrade.shtml をご覧ください。

Adobe、Adobe ロゴ、PostScript、PostScript ロゴは、Adobe Systems Incorporated の米国およびその他の国における登録商標または商標です。

インテル、Intel、Intel ロゴ、Intel Inside、Intel Inside ロゴ、Centrino、Intel Centrino ロゴ、Celeron、Xeon、Intel SpeedStep、Itanium、および Pentium は、Intel Corporation または子会社の米国およびその他の国における商標または登録商標です。

Linux は、Linus Torvalds の米国およびその他の国における登録商標です。

Microsoft、Windows、Windows NT および Windows ロゴは、Microsoft Corporation の米国およびその他の国における商標です。

UNIX は The Open Group の米国およびその他の国における登録商標です。

Java およびすべての Java 関連の商標およびロゴは Oracle やその関連会社の米国およびその他の国における商標または登録商標です。

製品資料に関するご使用条件

これらの資料は、以下のご使用条件に同意していただける場合に限りご使用いただけます。

適用条件

IBM Web サイトの「ご利用条件」に加えて、以下のご使用条件が適用されます。

個人的使用

これらの資料は、すべての著作権表示その他の所有権表示をしていただくことを条件に、非商業的な個人による使用目的に限り複製することができます。ただし、IBM の明示的な承諾をえずに、これらの資料またはその一部について、二次的著作物を作成したり、配布 (頒布、送信を含む) または表示 (上映を含む) することはできません。

商業的使用

これらの資料は、すべての著作権表示その他の所有権表示をしていただくことを条件に、お客様の企業内に限り、複製、配布、および表示することができます。ただし、IBM の明示的な承諾をえずにこれらの資料の二次的著作物を作成したり、お客様の企業外で資料またはその一部を複製、配布、または表示することはできません。

権利

ここで明示的に許可されているもの以外に、資料や資料内に含まれる情報、データ、ソフトウェア、またはその他の知的所有権に対するいかなる許可、ライセンス、または権利を明示的にも黙示的にも付与するものではありません。

資料の使用が IBM の利益を損なうと判断された場合や、上記の条件が適切に守られていないと判断された場合、IBM はいつでも自らの判断により、ここで与えた許可を撤回できるものとさせていただきます。

お客様がこの情報をダウンロード、輸出、または再輸出する際には、米国のすべての輸出入 関連法規を含む、すべての関連法規を遵守するものとします。

IBM は、これらの資料の内容についていかなる保証もしません。これらの資料は、特定物として現存するままの状態で提供され、商品性の保証、特定目的適合性の保証および法律上の瑕疵担保責任を含むすべての明示もしくは黙示の保証責任なしで提供されます。

用語集

C

Covariance (共分散). 2 つの変数の間の、標準化されていない関連度。偏差の積和を N-1 で割った値に等しくなります。

K

Kurtosis (尖度). 外れ値が存在する度合いの指標。正規分布の場合、尖度の統計値は 0 です。尖度が正の場合、そのデータの極端な外れ値は正規分布よりも多いことを示します。尖度が負の場合、そのデータの極端な外れ値は正規分布よりも少ないことを示します。

M

Maximum (最大). 数値変数の最大値。

Mean (平均). 中心傾向の指標。算術平均 (合計をケース数で割った値) です。

Median (中央値). この値より上と下それぞれにケースの半数ずつが該当することになる値。50 パーセンタイル。ケース数が偶数の場合の中央値は、昇順または降順にソートしたときに中央に来る 2 つのケースの平均です。中央値は、外れ値に対して敏感でない、中心傾向の指標です。それに対して平均値は、少数の極端に大きいまたは小さい値に影響されることがあります。

Minimum (最小値). 数値変数の最小値。

Mode (最頻値). 最も多く出現する値。複数の値が最高の頻度で出現し、その頻度が同じである場合は、それぞれが最頻値となります。

R

Range (*OK* (ファイルオープン時のオプション)). 数値変数の最大値と最小値の差。最大値から最小値を引いた値。

S

Skewness (歪度). 分布の非対称性の指標。正規分布は対称であり、歪度の値は 0 です。歪度が正の大きな値である分布は、右側の裾が長くなります。歪度が負で絶対値が大きい分布は、左側の裾が長くなります。目安として、歪度が標準誤差の 2 倍より大きい場合は、対称分布からずれていると解釈します。

standard deviation (標準偏差). 平均の周りの散らばりの指標。分散の平方根に等しくなります。標準偏差は元の変数と同じ単位で表します。

Standard Deviation (標準偏差). 平均値の周りの散らばりの指標。正規分布では、平均から 1 標準偏差以内にケースの 68% が含まれ、2 標準偏差以内にケースの 95% が含まれます。例えば平均年齢が 45 で標準偏差が 10 である場合、正規分布ではケースの 95% が 25 と 65 の間に含まれます。

Standard Error (標準誤差). サンプル間で検定統計量の値がどの程度ばらついているかの指標。統計量のサンプル分布の標準偏差です。例えば、平均値の標準誤差はサンプル平均の標準偏差です。

Standard Error of Kurtosis (尖度の標準誤差). 標準誤差に対する尖度の比率は、正規性の検定として使用することができます (比率が -2 より小さいか +2 より大きい場合は、正規性を棄却することができます)。尖度が大きな正の値である場合は、分布の裾が正規分布の裾より長いことを示します。尖度が負の値である場合は、裾が短いことを示します (箱形の一様分布に似た形になります)。

Standard Error of Mean (平均値の標準誤差). 同じ分布から抽出したサンプルの間で平均値がどの程度異なるかを示す指標。観測した平均と仮説による値をおおまかに比較するために使用することができます (差と標準誤差の比率が -2 より小さいか +2 より大きい場合は、2 つの値が異なっていると結論付けることができます)。

Standard Error of Skewness (歪度の標準誤差). 標準誤差に対する歪度の比率は、正規性の検定として使用することができます (比率が -2 より小さいか +2 より大きい場合は、正規性を棄却することができます)。歪度が大きな正の値である場合は、右側の裾が長いことを示します。極端な負の値の場合は、左側の裾が長いことを示します。

Sum (合計). 欠損値でない値を持つすべてのケースにわたる値の和 (合計)。

U

Unique (固有). あらゆる種類の他のすべての効果に適合するように各効果を調整することによって、すべての効果を同時に評価します。

V

Valid (有効). ユーザー欠損として定義された値もシステム欠損値も持たない有効なケース。

Variance (分散 (信頼性分析)). 平均値の周りの値の散らばりの指標。平均値からの偏差の平方和を、ケース数より 1 少ない値で割ったものに等しくなります。分散の測定単位は、変数自体の単位の 2 乗です。

索引

日本語, 数字, 英字, 特殊文字の順に配列されています。なお, 濁音と半濁音は清音と同等に扱われています。

［ア行］

［カ行］

[サ行]

[タ行]

[ナ行]

[ハ行]

[マ行]

[ヤ行]

[ラ行]

［ワ行］

［数字］

A

B

C

D

E

F

G

H

I

L

M

N

O

P

Q

R

S

T

U

V

W

X

[特殊文字]

Printed in Japan

日本アイ・ビー・エム株式会社
〒103-8510 東京都中央区日本橋箱崎町19-21