IBM SPSS Modeler 15 Quellen- , Prozess- und Ausgabeknoten

Hinweis: Lesen Sie zunächst die allgemeinen Informationen unter Hinweise auf S. 539, bevor Sie dieses Informationsmaterial sowie das zugehörige Produkt verwenden.

Diese Ausgabe bezieht sich auf IBM SPSS Modeler 15 und alle nachfolgenden Versionen sowie Anpassungen, sofern dies in neuen Ausgaben nicht anders angegeben ist.

Screenshots von Adobe-Produkten werden mit Genehmigung von Adobe Systems Incorporated abgedruckt.

Screenshots von Microsoft-Produkten werden mit Genehmigung der Microsoft Corporation abgedruckt.

Vorwort

IBM® SPSS® Modeler ist die auf Unternehmensebene einsetzbare Data-Mining-Workbench von IBM Corp.. Mit SPSS Modeler können Unternehmen und Organisationen die Beziehungen zu ihren Kunden bzw. zu den Bürgern durch ein tief greifendes Verständnis der Daten verbessern. Organisationen benutzen die mithilfe von SPSS Modeler gewonnenen Erkenntnisse zur Bindung profitabler Kunden, zur Ermittlung von Cross-Selling-Möglichkeiten, zur Gewinnung neuer Kunden, zur Ermittlung von Betrugsfällen, zur Reduzierung von Risiken und zur Verbesserung der Verfügbarkeit öffentlicher Dienstleistungen.

Die visuelle Benutzeroberfläche von SPSS Modeler erleichtert die Anwendung des spezifischen Geschäftswissens der Benutzer, was zu leistungsstärkeren Vorhersagemodellen führt und die Zeit bis zur Lösungserstellung verkürzt. SPSS Modeler bietet zahlreiche Modellierungsverfahren, beispielsweise Algorithmen für Vorhersage, Klassifizierung, Segmentierung und Assoziationserkennung. Nach der Modellerstellung ermöglicht IBM® SPSS® Modeler Solution Publisher die unternehmensweite Bereitstellung für Entscheidungsträger oder in einer Datenbank.

Über IBM Business Analytics

IBM Business Analytics-Software bietet vollständige, einheitliche und genaue Informationen, auf die Entscheidungsträger vertrauen, um die Unternehmensleistung zu steigern. Ein umfassendes Portfolio von Anwendungen für Unternehmensinformationen, Vorhersageanalysen, Verwaltung der Finanzleistung und Strategie sowie Analysen bietet sofort klare und umsetzbare Einblicke in die aktuelle Leistung und ermöglicht die Vorhersage zukünftiger Ergebnisse. In Kombination mit umfassenden Branchenlösungen, bewährten Vorgehensweisen und professionellen Dienstleistungen können Unternehmen jeder Größe optimale Produktivität erreichen, die Entscheidungsfindung zuverlässig automatisieren und bessere Ergebnisse erzielen.

Als Teil dieses Portfolios unterstützt die IBM SPSS Predictive Analytics-Software Unternehmen dabei, zukünftige Ereignisse vorherzusagen und aktiv auf diese Erkenntnisse zu reagieren, um bessere Geschäftsergebnisse zu erzielen. Kunden aus den Bereichen Wirtschaft, Behörden und Bildung aus aller Welt verlassen sich auf die IBM SPSS-Technologie. Sie bringt Ihnen beim Gewinnen, Halten und Ausbauen neuer Kundenbeziehungen einen Wettbewerbsvorteil und verringert gleichzeitig das Betrugs- sowie andere Risiken. Durch Integration der IBM SPSS-Software in den täglichen Betrieb können diese Unternehmen qualifizierte Vorhersagen treffen und dadurch die Entscheidungsfindung so ausrichten und automatisieren, dass Geschäftsziele erreicht werden und ein messbarer Wettbewerbsvorteil entsteht. Wenn Sie weitere Informationen wünschen oder einen Mitarbeiter kontaktieren möchten, ist dies unter *http://www.ibm.com/spss* möglich.

Technischer Support

Kunden mit Wartungsvertrag können den technischen Support in Anspruch nehmen. Kunden können sich an den technischen Support wenden, wenn sie Hilfe bei der Arbeit mit IBM Corp.-Produkten oder bei der Installation in einer der unterstützten Hardware-Umgebungen benötigen. Die Kontaktdaten des Technischen Supports finden Sie auf der IBM Corp.-Website

unter *http://www.ibm.com/support*. Sie müssen bei der Kontaktaufnahme Ihren Namen, Ihre Organisation und Ihre Supportvereinbarung angeben.

Inhalt

Anhang

Kapitel 1

Informationen zu IBM SPSS Modeler

IBM® SPSS® Modeler ist ein Set von Data Mining-Tools, mit dem Sie auf der Grundlage Ihres Geschäftswissens schnell und einfach Vorhersagemodelle erstellen und zur Erleichterung der Entscheidungsfindung in die Betriebsabläufe einbinden können. SPSS Modeler, das auf der Grundlage des den Industrienormen entsprechenden Modells CRISP-DM entwickelt wurde, unterstützt den gesamten Data Mining-Prozess, von den Daten bis hin zu besseren Geschäftsergebnissen.

SPSS Modeler bietet eine Vielzahl von Modellbildungsmethoden, die aus dem maschinellen Lernen, der künstlichen Intelligenz und der Statistik stammen. Mit den in der Modellierungspalette verfügbaren Methoden können Sie aus Ihren Daten neue Informationen ableiten und Vorhersagemodelle erstellen. Jede Methode besitzt ihre Stärken und eignet sich besonders für bestimmte Problemtypen.

SPSS Modeler kann als Standalone-Produkt oder als Client in Verbindung mit SPSS Modeler Server erworben werden. Außerdem ist eine Reihe von Zusatzoptionen verfügbar, die in den folgenden Abschnitten kurz dargelegt werden. Weitere Informationen finden Sie unter *http://www.ibm.com/software/analytics/spss/products/modeler/*.

IBM SPSS Modeler-Produkte

Zur IBM® SPSS® Modeler-Produktfamilie und der zugehörigen Software gehören folgende Elemente.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server-Adapter für IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler ist eine funktionell in sich abgeschlossene Produktversion, die Sie auf Ihrem PC installieren und ausführen können. Sie können SPSS Modeler im lokalen Modus als Standalone-Produkt oder im verteilten Modus zusammen mit IBM® SPSS® Modeler Server verwenden, um bei Daten-Sets die Leistung zu verbessern.

Mit SPSS Modeler können Sie schnell und intuitiv genaue Vorhersagemodelle erstellen, und das ohne Programmierung. Mithilfe der speziellen visuellen Benutzeroberfläche können Sie ganz einfach den Data Mining-Prozess visualisieren. Mit der Unterstützung der in das Produkt

eingebetteten erweiterten Analyseprozesse können Sie zuvor verborgene Muster und Trends in Ihren Daten aufdecken. Sie können Ergebnisse modellieren und Einblick in die Faktoren gewinnen, die Einfluss auf diese Ergebnisse haben, wodurch Sie in die Lage versetzt werden, Geschäftschancen zu nutzen und Risiken abzuschwächen.

SPSS Modeler ist in zwei Editionen erhältlich: SPSS Modeler Professional und SPSS Modeler Premium. Für weitere Informationen siehe Thema IBM SPSS Modeler-Editionen auf S. 3.

IBM SPSS Modeler Server

SPSS Modeler verwendet eine Client/Server-Architektur zur Verteilung von Anforderungen für ressourcenintensive Vorgänge an leistungsstarke Serversoftware, wodurch bei größeren Daten-Sets eine schnellere Leistung erzielt werden kann.

SPSS Modeler Server ist ein separat lizenziertes Produkt, das durchgehend im verteilten Analysemodus auf einem Server-Host in Verbindung mit einer oder mehreren IBM® SPSS® Modeler-Installationen ausgeführt wird. Auf diese Weise bietet SPSS Modeler Server eine herausragende Leistung bei großen Daten-Sets, da speicherintensive Vorgänge auf dem Server ausgeführt werden können, ohne Daten auf den Client-Computer herunterladen zu müssen. IBM® SPSS® Modeler Server bietet außerdem Unterstützung für SQL-Optimierung sowie Möglichkeiten zur Modellierung innerhalb der Datenbank, was weitere Vorteile hinsichtlich Leistung und Automatisierung mit sich bringt.

IBM SPSS Modeler Administration Console

Die Modeler Administration Console ist eine grafische Anwendung zur Verwaltung einer Vielzahl der SPSS Modeler Server-Konfigurationsoptionen, die auch mithilfe einer Optionsdatei konfiguriert werden können. Die Anwendung bietet eine Konsolen-Benutzeroberfläche zur Überwachung und Konfiguration der SPSS Modeler Server-Installationen und steht aktuellen SPSS Modeler Server-Kunden kostenlos zur Verfügung. Die Anwendung kann nur unter Windows installiert werden. Der von ihr verwaltete Server kann jedoch auf einer beliebigen unterstützten Plattform installiert sein.

IBM SPSS Modeler Batch

Data Mining ist zwar für gewöhnlich ein interaktiver Vorgang, es ist jedoch auch möglich, SPSS Modeler über eine Befehlszeile auszuführen, ohne dass die grafische Benutzeroberfläche verwendet werden muss. Beispielsweise kann es sinnvoll sein, langwierige oder repetitive Aufgaben ohne Eingreifen des Benutzers durchzuführen. SPSS Modeler Batch ist eine spezielle Version des Produkts, die die vollständigen Analysefunktionen von SPSS Modeler ohne Zugriff auf die reguläre Benutzeroberfläche bietet. Zur Verwendung von SPSS Modeler Batch ist eine SPSS Modeler Server-Lizenz erforderlich.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher ist ein Tool, mit dem Sie eine gepackte Version eines SPSS Modeler-Streams erstellen können, der durch eine externe Runtime-Engine ausgeführt oder in eine externe Anwendung eingebettet werden kann. Auf diese Weise können Sie vollständige SPSS Modeler-Streams für die Verwendung in Umgebungen veröffentlichen und bereitstellen, in denen SPSS Modeler nicht installiert ist. SPSS Modeler Solution Publisher wird als Teil des IBM SPSS Collaboration and Deployment Services - Scoring-Diensts verteilt, für den eine separate Lizenz erforderlich ist. Mit dieser Lizenz erhalten Sie SPSS Modeler Solution Publisher Runtime, womit Sie die veröffentlichen Streams ausführen können.

IBM SPSS Modeler Server-Adapter für IBM SPSS Collaboration and Deployment Services

Es ist eine Reihe von Adaptern für IBM® SPSS® Collaboration and Deployment Services verfügbar, mit denen SPSS Modeler und SPSS Modeler Server mit einem IBM SPSS Collaboration and Deployment Services-Repository interagieren können. Auf diese Weise kann ein im Repository bereitgestellter SPSS Modeler-Stream von mehreren Benutzern gemeinsam verwendet werden. Auch der Zugriff über die Thin-Client-Anwendung IBM SPSS Modeler Advantage ist möglich. Sie installieren den Adapter auf dem System, das als Host für das Repository fungiert.

IBM SPSS Modeler-Editionen

SPSS Modeler ist in den folgenden Editionen erhältlich.

SPSS Modeler Professional

SPSS Modeler Professional bietet sämtliche Tools, die Sie für die Arbeit mit den meisten Typen von strukturierten Daten benötigen, beispielsweise in CRM-Systemen erfasste Verhaltensweisen und Interaktionen, demografische Daten, Kaufverhalten und Umsatzdaten.

SPSS Modeler Premium

SPSS Modeler Premium ist ein separat lizenziertes Produkt, das SPSS Modeler Professional für die Arbeit mit spezialisierten Daten erweitert, wie beispielsweise den Daten, die für Entitätsanalysen oder soziale Netzwerke verwendet werden, sowie für die Arbeit mit unstrukturierten Textdaten. SPSS Modeler Premium umfasst die folgenden Komponenten.

IBM® SPSS® Modeler Entity Analytics fügt eine völlig neue Dimension zu den IBM® SPSS® Modeler-Vorhersageanalysen hinzu. Während bei Vorhersageanalysen versucht wird, zukünftiges Verhalten aus früheren Daten vorherzusagen, liegt der Schwerpunkt bei der Entitätsanalyse auf der Verbesserung von Kohärenz und Konsistenz der aktuellen Daten, indem Identitätskonflikte innerhalb der Datensätze selbst aufgelöst werden. Bei der Identität kann es sich um die Identität einer Person, einer Organisation, eines Objekts oder einer anderen Entität handelt, bei der Unklarheiten bestehen könnten. Die Identitätsauflösung kann in einer Reihe von Bereichen

entscheidend sein, darunter Customer Relationship Management, Betrugserkennung, Bekämpfung der Geldwäsche sowie nationale und internationale Sicherheit.

IBM SPSS Modeler Social Network Analysis transformiert Informationen zu Beziehungen in Felder, die das Sozialverhalten von Einzelpersonen und Gruppen charakterisieren. Durch die Verwendung von Daten, die die Beziehungen beschreiben, die sozialen Netzwerken zugrunde liegen, ermittelt IBM® SPSS® Modeler Social Network Analysis Führer in sozialen Netzwerken, die das Verhalten anderer Personen im Netzwerk beeinflussen. Außerdem können Sie feststellen, welche Personen am meisten durch andere Teilnehmer im Netzwerk beeinflusst werden. Durch die Kombination dieser Ergebnisse mit anderen Maßzahlen können Sie aussagekräftige Profile für Einzelpersonen, die Sie als Grundlage für Ihre Vorhersagemodelle verwenden können. Modelle, die diese sozialen Informationen berücksichtigen, sind leistungsstärker als Modelle, die dies nicht tun.

Text Analytics for IBM® SPSS® Modeler verwendet hoch entwickelte linguistische Technologien und die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP), um eine schnelle Verarbeitung einer großen Vielfalt an unstrukturierten Textdaten zu ermöglichen, um die Schlüsselkonzepte zu extrahieren und zu ordnen und um diese Konzepte in Kategorien zusammenzufassen. Extrahierte Konzepte und Kategorien können mit bestehenden strukturierten Daten, beispielsweise demografischen Informationen, kombiniert und mithilfe der vollständigen Suite der Data-Mining-Tools von SPSS Modeler auf die Modellierung angewendet werden, um bessere und fokussiertere Entscheidungen zu ermöglichen.

IBM SPSS Modeler-Dokumentation

Dokumentation im Online-Hilfe-Format finden Sie im Hilfe-Menü von SPSS Modeler. Dazu gehören die Dokumentation für SPSS Modeler, SPSS Modeler Server und SPSS Modeler Solution Publisher sowie das Anwendungshandbuch und weiteres Material zur Unterstützung.

Die vollständige Dokumentation für die einzelnen Produkte (einschließlich Installationsanweisungen) steht im PDF-Format im Ordner *\Documentation* auf der jeweiligen Produkt-DVD zur Verfügung. Installationsdokumente können auch aus dem Internet unter *http://www-01.ibm.com/support/docview.wss?uid=swg27023172* heruntergeladen werden:

Dokumentation in beiden Formaten steht auch im SPSS Modeler Information Center unter *http://publib.boulder.ibm.com/infocenter/spssmodl/v15r0m0/* zur Verfügung.

SPSS Modeler Professional-Dokumentation

Die SPSS Modeler Professional-Dokumentationssuite (ohne Installationsanweisungen) umfasst folgende Dokumente:

- **IBM SPSS Modeler-Benutzerhandbuch.** Allgemeine Einführung in die Verwendung von SPSS Modeler, in der u. a. die Erstellung von Daten-Streams, der Umgang mit fehlenden Werten, die Erstellung von CLEM-Ausdrücken, die Arbeit mit Projekten und Berichten sowie das Packen von Streams für das Deployment in IBM SPSS Collaboration and Deployment Services, Predictive Applications (Prognoseanwendungen) oder IBM SPSS Modeler Advantage beschrieben werden.

- **Quellen-, Prozess- und Ausgabeknoten in IBM SPSS Modeler.** Beschreibung aller Knoten, die zum Lesen, zum Verarbeiten und zur Ausgabe von Daten in verschiedenen Formaten verwendet werden. Im Grunde sind sie alle Knoten, mit Ausnahme der Modellierungsknoten.
- **IBM SPSS Modeler Modellierungsknoten.** Beschreibungen sämtlicher für die Erstellung von Data Mining-Modellen verwendeter Knoten. IBM® SPSS® Modeler bietet eine Vielzahl von Modellbildungsmethoden, die aus dem maschinellen Lernen, der künstlichen Intelligenz und der Statistik stammen.
- **IBM SPSS Modeler-Algorithmushandbuch.** Beschreibung der mathematischen Grundlagen der in SPSS Modeler verwendeten Modellierungsmethoden. Dieses Handbuch steht nur im PDF-Format zur Verfügung.
- **IBM SPSS Modeler-Anwendungshandbuch.** Die Beispiele in diesem Handbuch bieten eine kurze, gezielte Einführung in bestimmte Modellierungsmethoden und -verfahren. Eine Online-Version dieses Handbuchs kann auch über das Hilfe-Menü aufgerufen werden. Für weitere Informationen siehe Thema Anwendungsbeispiele auf S. 6.
- **Skripterstellung und Automatisierung in IBM SPSS Modeler.** Informationen zur Automatisierung des Systems über Skripterstellung, einschließlich der Eigenschaften, die zur Bearbeitung von Knoten und Streams verwendet werden können.
- **IBM SPSS Modeler Deployment-Handbuch.** Informationen zum Ausführen von SPSS Modeler-Streams und -Szenarien als Schritte bei der Verarbeitung von Jobs im IBM® SPSS® Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler CLEF-Entwicklerhandbuch.** CLEF bietet die Möglichkeit, Drittanbieterprogramme, wie Datenverarbeitungsroutinen oder Modellierungsalgorithmen, als Knoten in SPSS Modeler zu integrieren.
- **In-Database Mining-Handbuch für IBM SPSS Modeler.** Informationen darüber, wie Sie Ihre Datenbank dazu einsetzen, die Leistung zu verbessern, und wie Sie die Palette der Analysefunktionen über Drittanbieteralgorithmen erweitern.
- **IBM SPSS Modeler Server-Verwaltungs- und -Leistungshandbuch.** Informationen zur Konfiguration und Verwaltung von IBM® SPSS® Modeler Server.
- **IBM SPSS Modeler Administration Console – Benutzerhandbuch.** Informationen zur Installation und Nutzung der Konsolen-Benutzeroberfläche zur Überwachung und Konfiguration von SPSS Modeler Server. Die Konsole ist als Plugin für die Deployment Manager-Anwendung implementiert.
- **IBM SPSS Modeler Solution Publisher-Handbuch.** SPSS Modeler Solution Publisher ist eine Zusatzkomponente, mit der Unternehmen Streams zur Verwendung außerhalb der SPSS Modeler-Standardumgebung veröffentlichen können.
- **IBM SPSS Modeler-Handbuch zu CRISP-DM.** Schritt-für-Schritt-Anleitung für das Data Mining mit SPSS Modeler unter Verwendung der CRISP-DM-Methode.
- **IBM SPSS Modeler Batch-Benutzerhandbuch.** Vollständiges Handbuch für die Verwendung von IBM SPSS Modeler im Batch-Modus, einschließlich Details zur Ausführung des Batch-Modus und zu Befehlszeilenargumenten. Dieses Handbuch steht nur im PDF-Format zur Verfügung.

SPSS Modeler Premium-Dokumentation

Die SPSS Modeler Premium-Dokumentationssuite (ohne Installationsanweisungen) umfasst folgende Dokumente:

- **IBM SPSS Modeler Entity Analytics – Benutzerhandbuch.** Information zur Verwendung von Entitätsanalysen mit SPSS Modeler, unter Behandlung von Repository-Installation und -Konfiguration, Entity Analytics-Knoten und Verwaltungsaufgaben.
- **IBM SPSS Modeler Social Network Analysis – Benutzerhandbuch.** Ein Handbuch zur Durchführung sozialer Netzwerkanalyse mit SPSS Modeler, einschließlich Gruppenanalyse und Diffusionsanalyse.
- **Text Analytics for SPSS Modeler – Benutzerhandbuch.** Informationen zur Verwendung von Textanalysen mit SPSS Modeler, unter Behandlung der Text Mining-Knoten, der interaktiven Workbench sowie von Vorlagen und anderen Ressourcen.
- **Text Analytics for IBM SPSS Modeler Administration Console – Benutzerhandbuch.** Informationen zur Installation und Nutzung der Konsolen-Benutzeroberfläche zur Überwachung und Konfiguration von IBM® SPSS® Modeler Server für die Verwendung mit Text Analytics for SPSS Modeler. Die Konsole ist als Plugin für die Deployment Manager-Anwendung implementiert.

Anwendungsbeispiele

Mit den Data-Mining-Tools in SPSS Modeler kann eine große Bandbreite an geschäfts- und unternehmensbezogenen Problemen gelöst werden; die Anwendungsbeispiele dagegen bieten jeweils eine kurze, gezielte Einführung in spezielle Modellierungsmethoden und -verfahren. Die hier verwendeten Daten-Sets sind viel kleiner als die riesigen Datenbestände, die von einigen Data-Mining-Experten verwaltet werden müssen, die zugrunde liegenden Konzepte und Methoden sollten sich jedoch auch auf reale Anwendungen übertragen lassen.

Sie können auf die Beispiele zugreifen, indem Sie im Menü “Hilfe” in SPSS Modeler auf die Option Anwendungsbeispiele klicken. Die Datendateien und Beispiel-Streams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnisses, installiert. Für weitere Informationen siehe Thema Ordner “Demos” auf S. 6.

Beispiele für die Datenbank-Modellierung. Die Beispiele finden Sie im *IBM SPSS Modeler In-Database Mining-Handbuch*.

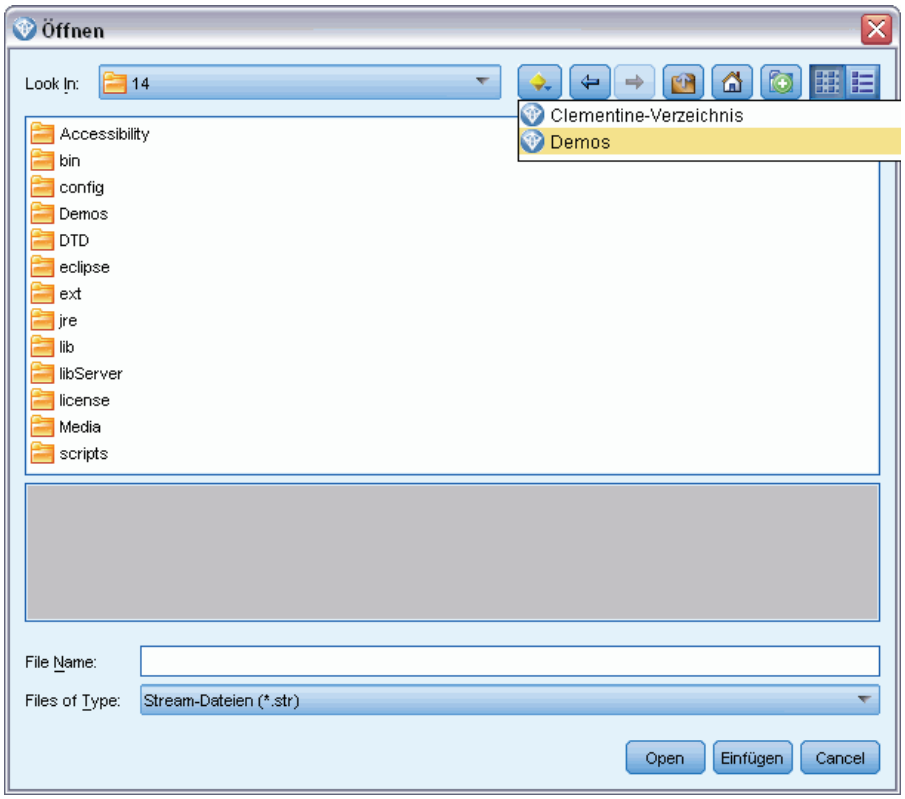
Skriptbeispiele. Die Beispiele finden Sie im *IBM SPSS Modeler Handbuch für die Skripterstellung und Automatisierung*.

Ordner “Demos”

Die in den Anwendungsbeispielen verwendeten Datendateien und Beispiel-Streams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnisses, installiert. Auf diesen Ordner können Sie auch über die Programmgruppe IBM SPSS Modeler 15 im Windows-Startmenü

oder durch Klicken auf *Demos* in der Liste der zuletzt angezeigten Verzeichnisse im Dialogfeld "Datei öffnen" zugreifen.

Abbildung 1-1
Auswahl des Ordners "Demos" in der Liste der zuletzt angezeigten Verzeichnisse



Kapitel
2

Quellenknoten

Übersicht

Mithilfe von Quellenknoten können Sie Daten importieren, die in einer Reihe von Formaten gespeichert sind, darunter Textdateien, IBM® SPSS® Statistics (*.sav*), SAS, Microsoft Excel und ODBC-kompatible relationale Datenbanken. Mit dem Benutzereingabeknoten können Sie außerdem künstliche Daten generieren.

Die Palette "Datenquellen" enthält folgende Knoten:

Der Enterprise-Ansichts-Knoten erstellt eine Verbindung mit einem IBM® SPSS® Collaboration and Deployment Services Repository, was es Ihnen ermöglicht, Enterprise-Ansichts-Daten in einen Stream einzulesen und ein Modell in ein Szenario zu packen, auf das andere Benutzer über das Repository zugreifen können. Für weitere Informationen siehe Thema Enterprise-Ansichts-Knoten auf S. 9.

Mit dem Datenbankknoten lassen sich Daten aus einer Reihe von anderen Paketen importieren, die ODBC (Open Database Connectivity) verwenden, darunter u. a. Microsoft SQL Server, DB2 und Oracle. Für weitere Informationen siehe Thema Datenbankquellenknoten auf S. 15.

Der Knoten für variable Dateien liest Daten aus Textdateien mit freien Feldern, also aus Dateien, deren Datensätze eine konstante Anzahl von Feldern, aber eine variable Anzahl von Zeichen enthalten. Dieser Knoten ist außerdem nützlich für Dateien mit fester Länge, Überschriftentext und bestimmten Anmerkungen. Für weitere Informationen siehe Thema Knoten "Datei (var.)" auf S. 26.

Der Knoten des Typs "Datei (fest)" importiert Daten aus Textdateien mit festen Feldern, also aus Dateien, deren Felder nicht begrenzt sind, sondern an derselben Position beginnen und eine feste Länge haben. Maschinell erzeugte Daten oder Legacydaten werden häufig im Format mit festen Feldern gespeichert. Für weitere Informationen siehe Thema Knoten "Datei (fest)" auf S. 29.

Der Statistikdateiknoten liest Daten aus dem Dateiformat *.sav* ein, das von SPSS Statistics verwendet wird, sowie in IBM® SPSS® Modeler gespeicherte Cache-Dateien, die ebenfalls dasselbe Format verwenden. Für weitere Informationen siehe Thema Statistikdateiknoten in Kapitel 8 auf S. 499.

Der IBM® SPSS® Data Collection-Knoten importiert Daten aus zahlreichen in der Marktforschungs-Software verwendeten Formaten und passt sie dem Data Collection-Datenmodell an. Um diesen Knoten verwenden zu können, muss die Data Collection Developer Library installiert sein. Für weitere Informationen siehe Thema Data Collection Knoten auf S. 35.

Der IBM Cognos BI-Quellenknoten importiert Daten aus Cognos BI-Datenbanken. Für weitere Informationen siehe Thema Importieren von Cognos-Daten auf S. 44.

Der SAS-Dateiknoten importiert SAS-Daten in SPSS Modeler. Für weitere Informationen siehe Thema SAS-Quellenknoten auf S. 50.

Der Excel--Knoten importiert Daten aus einer beliebigen Version von Microsoft Excel. Es ist keine ODBC-Datenquelle erforderlich. Für weitere Informationen siehe Thema Excel-Quellenknoten auf S. 52.

Der XML-Quellenknoten importiert Daten im XML-Format in den Stream. Sie können eine einzelne Datei oder alle Dateien in einem Verzeichnis importieren. Optional können Sie eine Schemadatei angeben, aus der die XML-Struktur gelesen werden soll. Für weitere Informationen siehe Thema XML-Quellenknoten auf S. 53.

Der Benutzereingabeknoten bietet eine einfache Möglichkeit, künstliche Daten zu erstellen. Dazu können entweder neue Daten ohne Vorlage erstellt oder vorhandene Daten geändert werden. Diese Funktion ist nützlich, wenn Sie z. B. ein Test-Daten-Set für die Modellierung erstellen möchten. Für weitere Informationen siehe Thema Benutzereingabeknoten auf S. 57.

Um mit dem Erstellen eines Streams zu beginnen, fügen Sie einen Quellenknoten zum Stream-Zeichenbereich hinzu. Doppelklicken Sie dann auf den Knoten, um das zugehörige Dialogfeld zu öffnen. Auf den einzelnen Registerkarten im Dialogfeld können Sie Daten einlesen, Felder und Werte anzeigen und eine Vielzahl von Optionen festlegen, wie Filter, Datentypen, Feldrolle und die Überprüfung fehlender Werte.

Enterprise-Ansichts-Knoten

Mit dem Enterprise-Ansichts-Knoten können Sie eine Verbindung zwischen einer IBM® SPSS® Modeler-Sitzung und einer Enterprise-Ansicht in einem freigegebenen IBM® SPSS® Collaboration and Deployment Services Repository herstellen und aufrechterhalten. Dadurch können Sie Daten aus einer Enterprise-Ansicht in einen SPSS Modeler-Stream einlesen und ein SPSS Modeler-Modell in ein Szenario packen, auf das andere Benutzer des gemeinsam genutzten Repository Zugriff haben.

Ein **Szenario** ist eine Datei, die einen SPSS Modeler-Stream mit bestimmten Knoten, Modellen und weiteren Eigenschaften enthält, die es möglich machen, ein Deployment des Streams in einem IBM SPSS Collaboration and Deployment Services Repository vorzunehmen, um ein Scoring oder eine automatisierte Modellaktualisierung durchzuführen. Die Verwendung von Enterprise-Ansichts-Knoten mit Szenarios gewährleistet, dass bei einer Konstellation mit mehreren Benutzern alle Benutzer auf der Grundlage derselben Daten arbeiten. Eine **Verbindung** ist eine Verknüpfung von einer SPSS Modeler-Sitzung zu einer Enterprise-Ansicht im IBM SPSS Collaboration and Deployment Services Repository.

Die **Enterprise-Ansicht** ist die Gesamtmenge der Daten, die zu einer Organisation gehören, unabhängig davon, wo sich diese Daten physisch befinden. Jede Verbindung besteht aus einer Auswahl einer einzelnen **Application-Ansicht** (Untermenge der Enterprise-Ansicht, die auf eine bestimmte Anwendung zugeschnitten ist), einer **Daten-Provider-Definition** (DPD – stellt eine Verknüpfung zwischen den logischen Tabellen und Spalten der Application-Ansicht und einer physischen Datenquelle her) und einer **Umgebung** (gibt an, welche Spalten jeweils den definierten

Geschäftssegmenten zugeordnet werden sollen). Enterprise-Ansicht, Application-Ansicht und PDP-Definitionen sind im Repository gespeichert (dort findet auch die Versionsverwaltung statt). Die tatsächlichen Daten befinden sich jedoch in einer oder mehreren Datenbanken oder in anderen externen Quellen.

Sobald eine Verbindung hergestellt wurde, geben Sie eine Application-Ansichts-**Tabelle** für die Arbeit in SPSS Modeler an. In einer Application-Ansicht ist eine Tabelle eine logische Ansicht, die aus einigen oder allen Spalten aus einer oder mehreren physischen Tabellen in einer oder mehreren physischen Datenbanken besteht. So ermöglicht der Enterprise-Ansichts-Knoten, dass Datensätze aus mehreren Datenbanktabellen in SPSS Modeler als eine einzige Tabelle angezeigt werden.

Voraussetzungen

- Um den Enterprise-Ansichts-Knoten verwenden zu können, muss zuvor IBM SPSS Collaboration and Deployment Services Repository an Ihrem Standort installiert und konfiguriert sein. Dabei müssen bereits eine Enterprise-Ansicht, Application-Ansichten und DPDs definiert sein.

Hinweis: Für den Zugriff auf ein IBM® SPSS® Collaboration and Deployment Services-Repository ist eine separate Lizenz erforderlich. Weitere Informationen finden Sie im Dokument *http://www.ibm.com/software/analytics/spss/products/deployment/cds/*

- Außerdem muss auf jedem Computer, der zur Bearbeitung oder Ausführung des Streams verwendet wird, der IBM® SPSS® Collaboration and Deployment Services Enterprise View Driver installiert sein. Unter Windows installieren Sie den Treiber einfach auf dem Computer, auf dem IBM® SPSS® Modeler bzw. IBM® SPSS® Modeler Server installiert ist. Es ist keine weitere Konfiguration des Treibers erforderlich. Unter UNIX muss ein Verweis auf das Skript *pev.sh* zum Startskript hinzugefügt werden. Details zur Installation des IBM SPSS Collaboration and Deployment Services Enterprise View Driver erhalten Sie bei Ihrem lokalen Administrator.
- Eine DPD wird anhand einer bestimmten ODBC-Datenquelle definiert. Um eine DPD aus SPSS Modeler zu verwenden, muss eine ODBC-Datenquelle auf dem SPSS Modeler Server-Host definiert sein, der denselben Namen trägt und der eine Verbindung zum selben Datenspeicher herstellt wie die in der DPD referenzierte Datenquelle.

Festlegen der Optionen für den Enterprise-Ansichts-Knoten

Mit den Optionen auf der Registerkarte “Daten” des Dialogfelds “Enterprise-Ansicht” haben Sie folgende Möglichkeiten:

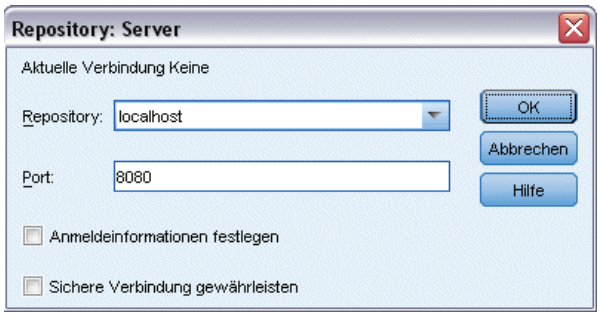
- Auswahl einer bestehenden Repository-Verbindung
- Bearbeiten einer bestehenden Repository-Verbindung
- Erstellen einer neuen Repository-Verbindung
- Auswahl einer Application-Ansichts-Tabelle

Einzelheiten zur Arbeit mit Repositories finden Sie im *IBM® SPSS® Collaboration and Deployment Services-Administratorhandbuch*.

Abbildung 2-1
Hinzufügen einer Verbindung zu einem IBM SPSS Collaboration and Deployment Services Repository

Verbindung. Die Dropdown-Liste bietet Optionen zur Auswahl einer bestehenden Repository-Verbindung, zum Bearbeiten einer bestehenden Repository-Verbindung bzw. zum Hinzufügen einer Verbindung. Wenn Sie bereits über IBM® SPSS® Modeler bei einem Repository angemeldet sind, wird durch Auswahl der Option Verbindung hinzufügen/bearbeiten das Dialogfeld "Enterprise-Ansichts-Verbindungen" angezeigt. In diesem Dialogfeld können Sie die erforderlichen Details für die aktuelle Verbindung definieren bzw. bearbeiten. Wenn Sie nicht angemeldet sind, zeigt diese Option das Anmeldedialogfeld für das Repository an.

Abbildung 2-2
Anmelden bei einem Repository



Informationen zur Anmeldung im Repository finden Sie im *SPSS Modeler-Benutzerhandbuch*.

Sobald eine Verbindung zu einem Repository hergestellt wurde, bleibt diese Verbindung erhalten, bis Sie SPSS Modeler beenden. Eine Verbindung kann für andere Knoten innerhalb desselben Streams freigegeben werden; Sie müssen jedoch für jeden neuen Stream eine neue Verbindung erstellen.

Bei erfolgreicher Anmeldung wird das Dialogfeld “Enterprise-Ansichts-Verbindungen” angezeigt.

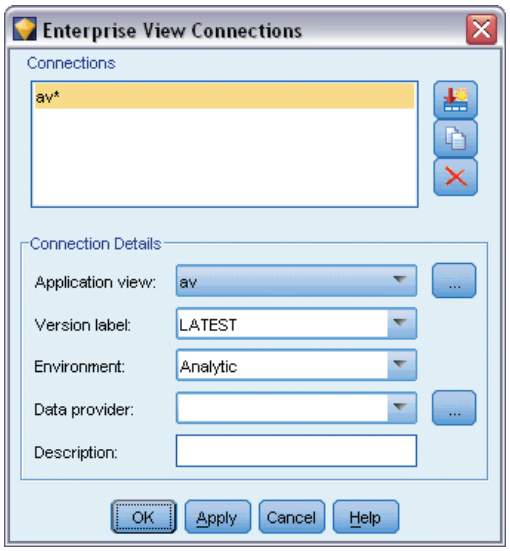
Tabellenname. Dieses Feld ist ursprünglich leer und kann erst nach der Herstellung einer Verbindung mit Daten versehen werden. Wenn Ihnen der Name der Application-Ansichts-Tabelle, auf die Sie zugreifen möchten, bekannt ist, geben Sie ihn in das Feld “Tabellenname” ein. Klicken Sie andernfalls auf die Schaltfläche Auswählen, um ein Dialogfeld mit einer Liste der verfügbaren Application-Ansichts-Tabellen zu öffnen.

Enterprise-Ansichts-Verbindungen

In diesem Dialogfeld können Sie die erforderlichen Details für die Repository-Verbindung definieren bzw. bearbeiten. Sie können folgende Elemente angeben:

- Application-Ansicht und Version
- Umgebung
- Daten-Provider-Definition (DPD)
- Verbindungsbeschreibung

Abbildung 2-3
Auswählen einer Application-Ansicht



Verbindungen. Listet bestehende Repository-Verbindungen auf.

- **Neue Verbindung hinzufügen.** Zeigt das Dialogfeld “Objekt abrufen” an, in dem Sie nach einer Application-Ansicht aus dem Repository suchen und diese auswählen können.
- **Ausgewählte Verbindung kopieren.** Erstellt eine Kopie einer ausgewählten Verbindung, sodass Sie nicht erneut zu derselben Application-Ansicht blättern müssen.
- **Ausgewählte Verbindung löschen.** Löscht die ausgewählte Verbindung aus der Liste.

Verbindungsdetails. Zeigt für die aktuell im Fenster "Verbindungen" ausgewählte Verbindung die Application-Ansicht, die Versionsbeschriftung, die Umgebung, DPD sowie einen beschreibenden Text an.

- **Application-Ansicht.** In der Dropdown-Liste wird ggf. die ausgewählte Application-Ansicht angezeigt. Wenn in der aktuellen Sitzung Verbindungen zu anderen Application-Ansichten hergestellt wurden, werden diese ebenfalls in der Dropdown-Liste angezeigt. Klicken Sie auf die angrenzende Schaltfläche "Durchsuchen", um nach anderen Application-Ansichten im Repository zu suchen.
- **Versionsbezeichnung.** Im Dropdown-Feld werden alle definierten Versionsbeschriftungen für die angegebene Application-Ansicht aufgeführt. Die Versionsbeschriftungen erleichtern die Kennzeichnung bestimmter Versionen von Repository-Objekten. Beispielsweise kann es zwei Versionen einer bestimmten Application-Ansicht geben. Bei Verwendung von Beschriftungen können Sie beispielsweise die Beschriftung TEST für die Version angeben, die in der Entwicklungsumgebung verwendet wird, und die Beschriftung PRODUKTION für die in der Produktionsumgebung verwendete Version. Wählen Sie eine geeignete Beschriftung aus.

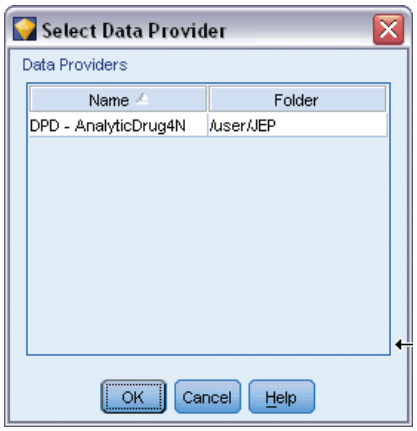
Hinweis: Bezeichnungen sollten das Zeichen "[" nicht enthalten, da sonst der Tabellenname nicht auf der Registerkarte "Daten" im Dialogfeld "Enterprise-Ansicht" angezeigt wird.

- **Umgebung.** Im Dropdown-Feld werden alle gültigen Umgebungen aufgelistet. Die Umgebungseinstellung bestimmt, welche DPDs verfügbar sind, gibt also an, welche Spalten definierten Geschäftssegmenten zugeordnet werden sollen. Bei Auswahl von Analytisch beispielsweise, werden nur die Spalten der Application-Ansicht zurückgegeben, die als Analytisch definiert sind. Die Standardumgebung lautet Analytisch; Sie können jedoch auch Betrieb auswählen.
- **Daten-Provider.** In der Dropdown-Liste werden die Namen von bis zu zehn Daten-Provider-Definitionen für die ausgewählte Application-Ansicht aufgeführt. Nur DPDs, die auf die ausgewählte Application-Ansicht verweisen, werden angezeigt. Klicken Sie auf die angrenzende Schaltfläche "Durchsuchen", um Namen und Pfad aller DPDs anzuzeigen, die sich auf die aktuelle Application-Ansicht beziehen.
- **Beschreibung.** Beschreibender Text zur Repository-Verbindung. Dieser Text wird als Verbindungsname verwendet. Beim Klicken auf OK wird der Text in der Dropdown-Liste "Verbindung" und in der Titelleiste des Dialogfelds "Enterprise-Ansicht" sowie als Beschriftung des Enterprise-Ansichts-Knotens im Zeichenbereich angezeigt.

Auswählen der DPD

Im Dialogfeld "Daten-Provider auswählen" werden Name und Pfad aller DPDs angezeigt, die auf die aktuelle Application-Ansicht verweisen.

Abbildung 2-4
Auswählen einer DPD



Application-Ansichten können mehrere DPDs aufweisen, um die verschiedenen Phasen eines Projekts zu unterstützen. So können beispielsweise die zur Modellerstellung verwendeten historischen Daten aus einer bestimmten Datenbank stammen, die operativen Daten jedoch aus einer anderen.

Eine DPD wird anhand einer bestimmten ODBC-Datenquelle definiert. Um eine DPD aus IBM® SPSS® Modeler zu verwenden, muss eine ODBC-Datenquelle auf dem SPSS Modeler Server-Host definiert sein, der denselben Namen trägt und der eine Verbindung zum selben Datenspeicher herstellt wie die in der DPD referenzierte Datenquelle.

- Um die zu verwendende DPD auszuwählen, markieren Sie ihren Namen auf der Liste und klicken Sie auf OK.

Auswählen der Tabelle

Im Dialogfeld “Tabelle auswählen” werden alle Tabellen aufgelistet, die in der aktuellen Application-Ansicht referenziert werden. Das Dialogfeld ist leer, wenn keine Verbindung zu einem IBM SPSS Collaboration and Deployment Services Repository hergestellt wurde.

Abbildung 2-5
Auswählen einer Tabelle

▶ Um die zu verwendende Tabelle auszuwählen, markieren Sie ihren Namen auf der Liste und klicken Sie auf OK.

Datenbankquellenknoten

Mit dem Datenbankquellenknoten lassen sich Daten aus einer Reihe von anderen Paketen importieren, die ODBC (Open Database Connectivity) verwenden, darunter u. a. Microsoft SQL Server, DB2 und Oracle.

Um in einer Datenbank zu lesen oder in ihr zu schreiben, muss eine ODBC-Datenquelle für die entsprechende Datenbank mit den erforderlichen Lese- und Schreibberechtigungen installiert und konfiguriert sein. Das IBM® SPSS® Data Access Pack umfasst eine Reihe von ODBC-Treibern, die zu diesem Zweck verwendet werden können. Diese Treiber stehen auf dem IBM SPSS Data Access Pack DVD oder auf der Download-Website zur Verfügung. Wenn Sie Fragen zur Erstellung oder Einstellung von Berechtigungen für ODBC-Datenquellen haben, wenden Sie sich an Ihren Datenbankadministrator.

Die Datenbankunterstützung in IBM® SPSS® Modeler wird in drei Stufen eingeteilt, wobei jede Stufe je nach Datenbankanbieter für einen unterschiedlichen Unterstützungsgrad für SQL-Pushback und -Optimierung steht. Die unterschiedlichen Unterstützungsebenen werden durch eine Reihe von Systemeinstellungen implementiert, die als Teil einer Dienstleistungsabsprache angepasst werden können.

Die drei Stufen der Datenbankunterstützung sind:

Tabelle 2-1
Stufen der Datenbankunterstützung

Unter-stützungs-stufe	Beschreibung
Stufe 1	Vollständiger SQL-Pushback verfügbar, mit datenbankspezifischer SQL-Optimierung.
Stufe 2	Teilweiser SQL-Pushback verfügbar, mit datenbankspezifischer SQL-Optimierung.
Stufe 3	Kein SQL-Pushback oder -Optimierung, Daten können nur von der Datenbank gelesen oder in die Datenbank geschrieben werden.

Unterstützte ODBC-Treiber

Neueste Informationen zu Datenbanken und ODBC-Treibern, die für die Verwendung mit SPSS Modeler 15 getestet wurden und unterstützt werden, finden Sie in den Produktkompatibilitätsdiagrammen auf der unternehmenseigenen Support-Site unter *http://www.ibm.com/support*.

Installationsort der Treiber

Beachten Sie, dass die ODBC-Treiber auf jedem Computer installiert und konfiguriert werden müssen, auf dem eine Verarbeitung erfolgt.

- Wenn Sie IBM® SPSS® Modeler im lokalen (Standalone-) Modus ausführen, müssen die Treiber auf dem lokalen Computer installiert sein.

- Wenn Sie SPSS Modeler im verteilten Modus mit einem Remote-IBM® SPSS® Modeler Server ausführen, müssen die ODBC-Treiber auf dem Computer installiert sein, auf dem SPSS Modeler Server installiert ist. Beachten Sie bei SPSS Modeler Server auf UNIX-Systemen auch "Konfiguration von ODBC-Treibern auf UNIX-Systemen" weiter hinten in diesem Abschnitt.
- Wenn Sie von SPSS Modeler und SPSS Modeler Server auf die gleichen Datenquellen zugreifen müssen, müssen die ODBC-Treiber auf beiden Computern installiert sein.
- Wenn Sie SPSS Modeler über Terminaldienste ausführen, müssen die ODBC-Treiber auf dem Terminaldienste-Server installiert sein, auf dem Sie SPSS Modeler installiert haben.
- Wenn Sie IBM® SPSS® Modeler Solution Publisher Runtime verwenden, um veröffentlichte Streams auf einem separaten Computer auszuführen, müssen Sie die ODBC-Treiber auch auf diesem Computer installieren und konfigurieren.

Hinweis: Wenn Sie SPSS Modeler Server unter UNIX zum Zugriff auf eine Teradata-Datenbank verwenden, müssen Sie den ODBC-Treiber-Manager verwenden, der mit dem Teradata-ODBC-Treiber installiert wurde. Um diese Änderung an SPSS Modeler Server vorzunehmen, geben Sie für ODBC_DRIVER_MANAGER_PATH einen Wert in der Nähe des oberen Bereichs des Skripts modelersrv.sh ein, wo dies durch die Kommentare angegeben wurde. Diese Umgebungsvariable muss auf den Speicherort des ODBC-Treiber-Managers eingestellt werden, der mit dem Teradata ODBC-Treiber ausgeliefert wird (/usr/odbc/lib in einer Standardinstallation eines Teradata ODBC-Treibers). Sie müssen SPSS Modeler Server neu starten, damit die Änderung wirksam wird. Weitere Informationen zu den SPSS Modeler Server-Plattformen, die Teradata-Zugriff unterstützen, sowie über die unterstützte Teradata ODBC-Treiberversion finden Sie auf der unternehmenseigenen Support-Site unter *http://www.ibm.com/support*.

Konfiguration von ODBC-Treibern auf UNIX-Systemen

Standardmäßig ist der DataDirect-Treiber-Manager nicht für SPSS Modeler Server auf UNIX-Systemen konfiguriert. Geben Sie folgende Befehle ein, um UNIX so zu konfigurieren, dass der DataDirect-Treiber-Manager geladen wird:

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

Dadurch wird die Standardverknüpfung entfernt und eine Verknüpfung zum DataDirect-Treiber-Manager erstellt.

Führen Sie die folgenden allgemeinen Schritte aus, um auf Daten einer Datenbank zuzugreifen:

▶ Installieren Sie einen ODBC-Treiber und konfigurieren Sie eine Datenquelle für die zu verwendende Datenbank.

▶ Stellen Sie im Dialogfeld des Datenbankknotens im Modus "Tabelle" oder "SQL-Abfrage" eine Verbindung zu einer Datenbank her.

▶ Wählen Sie eine Tabelle aus der Datenbank.

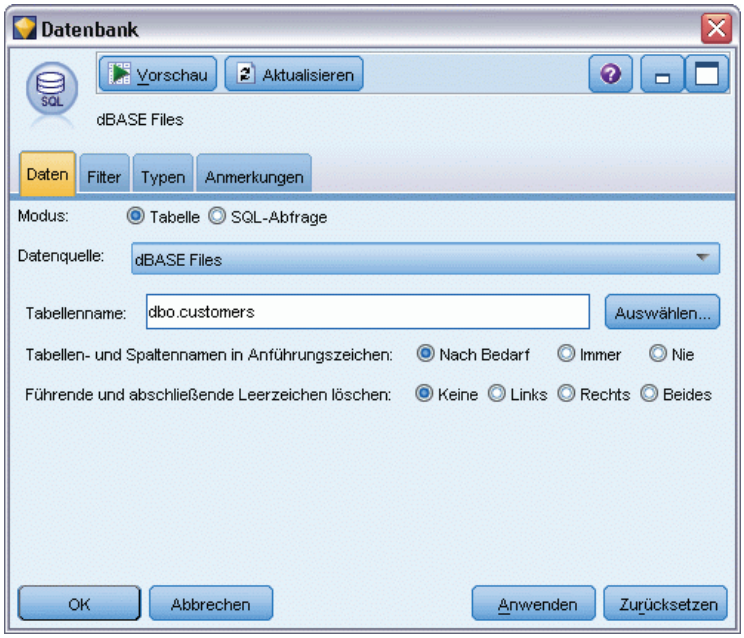
- Anhand der Registerkarten des Dialogfelds des Datenbankknotens können Sie Verwendungstypen ändern und Datenfelder filtern.

Diese Schritte werden in den nächsten Themenabschnitten ausführlicher beschrieben.

Festlegen von Optionen für Datenbankknoten

Mit den Optionen auf der Registerkarte "Daten" des Dialogfelds des Datenbankquellenknotens erhalten Sie Zugriff auf eine Datenbank und können Daten aus der ausgewählten Tabelle lesen.

Abbildung 2-6
Laden von Daten durch Auswahl einer Tabelle



Modus. Wählen Sie Tabelle, um mit den Steuerelementen des Dialogfelds eine Verbindung zu einer Tabelle herzustellen.

Wählen Sie SQL-Abfrage, um die unten ausgewählte Datenbank unter Verwendung von SQL abzufragen. Für weitere Informationen siehe Thema Abfragen der Datenbank auf S. 24.

Datenquelle. Sowohl im Modus "Tabelle" als auch im Modus "SQL-Abfrage" können Sie einen Namen in das Feld "Datenquelle" eingeben oder die Option Neue Datenbankverbindung hinzufügen in der Dropdown-Liste auswählen.

Die folgenden Optionen dienen zur Verbindung mit einer Datenbank und zur Auswahl einer Tabelle anhand des Dialogfelds:

Tabellenname. Wenn Ihnen der Name der Tabelle, auf die Sie zugreifen möchten, bekannt ist, geben Sie ihn in das Feld "Tabellenname" ein. Klicken Sie andernfalls auf die Schaltfläche Auswählen, um ein Dialogfeld mit einer Liste der verfügbaren Tabellen zu öffnen.

Tabellen- und Spaltennamen in Anführungszeichen. Legen Sie fest, ob die Tabellen- und Spaltennamen in Anführungszeichen eingeschlossen werden sollen, wenn Abfragen an die Datenbank gesendet werden (wenn sie z. B. Leerzeichen oder Satzzeichen enthalten).

- Bei Auswahl der Option Nach Bedarf werden Tabellen- und Feldnamen *nur* in Anführungszeichen gesetzt, wenn sie Nichtstandardzeichen enthalten. Nichtstandardzeichen sind Nicht-ASCII-Zeichen, Leerzeichen und alle nicht alphanumerischen Zeichen außer einem Punkt (.).
- Wählen Sie Nie, wenn Tabellen- und Feldnamen *nie* in Anführungszeichen gesetzt werden sollen.
- Wählen Sie Immer, wenn *alle* Tabellen- und Feldnamen in Anführungszeichen gesetzt werden sollen.

Führende und abschließende Leerzeichen löschen. Wählen Sie die Optionen zum Verwerfen von führenden und abschließenden Leerzeichen in Zeichenketten.

Anmerkung. Vergleiche zwischen Zeichenketten, die SQL-Pushback verwenden oder nicht, können unterschiedliche Ergebnisse generieren, wenn nachgestellte Leerzeichen vorhanden sind.

Lesen leerer Zeichenketten aus Oracle. Beim Lesen aus oder Schreiben in Oracle-Datenbanken sollten Sie darauf achten, dass Oracle im Gegensatz zu IBM® SPSS® Modeler und den meisten anderen Datenbanken leere Zeichenkettenwerte wie Nullwerte behandelt und speichert. Dies bedeutet, dass dieselben Daten sich unterschiedlich verhalten können und unterschiedliche Ergebnisse ausgeben können, je nachdem ob sie aus einer Oracle-Datenbank oder aus einer anderen Datenbank bzw. einer Datei extrahiert wurden.

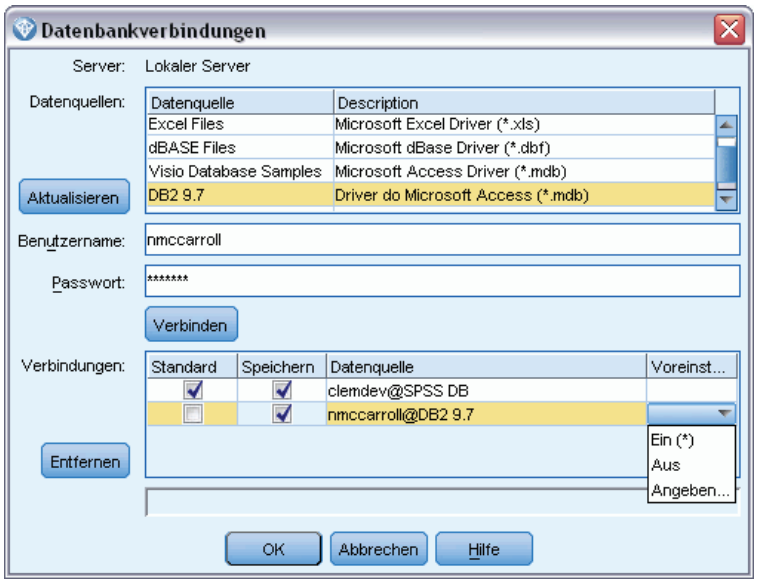
Hinzufügen einer Datenbankverbindung

Um eine Datenbank zu öffnen, müssen Sie zunächst die Datenquelle auswählen, mit der Sie sich verbinden möchten. Wählen Sie auf der Registerkarte “Daten” in der Dropdown-Liste “Datenquelle” die Option Neue Datenbankverbindung hinzufügen.

Das Dialogfeld “Datenbankverbindungen” wird geöffnet.

Hinweis: Alternativ können Sie dieses Dialogfeld über das Hauptmenü öffnen, indem Sie die folgenden Befehle wählen:
Werkzeuge > Datenbanken...

Abbildung 2-7
Datenbankverbindungen (Dialogfeld)



Datenquellen. Listet die verfügbaren Datenquellen auf. Führen Sie einen Bildlauf nach unten durch, wenn die gewünschte Datenbank nicht angezeigt wird. Nachdem Sie eine Datenquelle ausgewählt und ggf. Passwörter eingegeben haben, klicken Sie auf Verbinden. Klicken Sie zum Aktualisieren der Liste auf Aktualisieren.

Benutzername. Wenn die Datenquelle passwortgeschützt ist, geben Sie Ihren Benutzernamen ein.

Passwort. Wenn die Datenquelle passwortgeschützt ist, geben Sie Ihr Passwort ein.

Verbindungen. Zeigt die momentan verbundenen Datenbanken an.

- **Standard.** Sie können eine Verbindung optional als Standard auswählen. Diese Verbindung ist daraufhin für Datenbankquellen- und Exportknoten als Datenquelle vordefiniert. Sie können diese Einstellung jederzeit ändern.
- **Speichern.** Wählen Sie optional eine oder mehr Verbindungen aus, die in nachfolgenden Sitzungen wieder eingeblendet werden soll.
- **Datenquelle.** Die Verbindungszeichenketten für die aktuell verbundenen Datenbanken.
- **Voreingestellt.** Zeigt an (mit dem Zeichen *), ob voreingestellte Werte für die Datenbankverbindung angegeben wurden. Um voreingestellte Werte anzugeben, klicken Sie in dieser Spalte auf die der Datenbankverbindung entsprechenden Reihe und wählen Sie aus der Liste “Angeben” aus. Für weitere Informationen siehe Thema Angeben von voreingestellten Werten für eine Datenbankverbindung auf S. 20.

Zum Entfernen von Verbindungen wählen Sie eine Verbindung in der Liste aus und klicken Sie auf Entfernen.

Wenn Sie Ihre Auswahl getroffen haben, klicken Sie auf OK.

Angeben von voreingestellten Werten für eine Datenbankverbindung

Bei einigen Datenbanken können Sie verschiedene Standardeinstellungen für die Datenbankverbindung angeben. Diese Einstellungen gelten alle für den Datenbankexport.

Diese Funktion wird von folgenden Datenbanktypen unterstützt:

- IBM InfoSphere Warehouse unter DB2 9.1 oder höher. Für weitere Informationen siehe Thema Einstellungen für IBM DB2 InfoSphere Warehouse auf S. 20.
- SQL Server 2008 oder höher, Enterprise und Developer Edition. Für weitere Informationen siehe Thema Einstellungen für SQL Server auf S. 20.
- Oracle 10g und 11gR1 oder höher, Enterprise oder Personal Edition. Für weitere Informationen siehe Thema Einstellungen für Oracle auf S. 21.
- IBM Netezza, IBM DB2 unter z/OS und Teradata stellen eine Verbindung zu einer Datenbank oder zu einem Schema auf ähnliche Weise her. Für weitere Informationen siehe Thema Einstellungen für IBM Netezza, IBM DB2 unter z/OS und Teradata auf S. 22.

Wenn Sie mit einer Datenbank oder einem Schema verbunden sind, die bzw. das diese Funktion nicht unterstützt, wird die Meldung Für diese Datenbankverbindung können keine Voreinstellungen konfiguriert werden angezeigt.

Einstellungen für IBM DB2 InfoSphere Warehouse

Diese Einstellungen werden für IBM InfoSphere Warehouse unter DB2 9.1 oder höher angezeigt.

Tabellenbereich. Der Tabellenbereich, der für den Export verwendet wird. Datenbankadministratoren können Tabellenbereiche partioniert erstellen oder konfigurieren. Wir empfehlen, einen dieser Tabellenbereiche (anstelle des standardmäßig eingestellten) für den Datenbankexport zu verwenden.

Komprimierung verwenden. Wenn ausgewählt, werden Tabellen für den komprimierten Export erstellt (entspricht z. B. CREATE TABLE MYTABLE(…) COMPRESS YES; in SQL).

Do not log updates. Wenn ausgewählt, werden keine Protokolle beim Erstellen von Tabellen und Einfügen von Daten erstellt (entspricht CREATE TABLE MYTABLE(…) NOT LOGGED INITIALLY; in SQL).

Einstellungen für SQL Server

Diese Einstellungen werden für SQL Server 2008 oder höher, Enterprise und Developer Edition, angezeigt.

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Zeile.** Aktiviert Komprimierung auf der Zeilenebene (entspricht z. B. CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = ROW); in SQL).
- **Seite.** Aktiviert Komprimierung auf der Seitenebene (z. B. CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = PAGE); in SQL).

Einstellungen für Oracle

Oracle 10g-Einstellungen

Diese Einstellungen werden für Oracle 10g, Enterprise oder Personal Edition, angezeigt.

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt. Für diese Version der Datenbank steht nur einfache Komprimierung zur Verfügung (beispielsweise CREATE TABLE MYTABLE(…) COMPRESS; in SQL).

Oracle 11gR1-Einstellungen

Diese Einstellungen werden für Oracle 11g, Enterprise oder Personal Edition, angezeigt.

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. CREATE TABLE MYTABLE(…) COMPRESS; in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option Direkte Ladevorgänge.
- **Direkte Ladevorgänge.** Aktiviert Komprimierung ausschließlich für Masseneinfügevorgänge (direkter Pfad) (z. B. CREATE TABLE MYTABLE(…)COMPRESS FOR DIRECT_LOAD OPERATIONS; in SQL).
- **Alle Vorgänge.** Aktiviert Komprimierung für alle Vorgänge (z. B. CREATE TABLE MYTABLE(…)COMPRESS FOR ALL OPERATIONS; in SQL).

Oracle 11gR2-Einstellungen – Option "Basic" (Einfach)

Diese Einstellungen werden für Oracle 11g R2, Enterprise oder Personal Edition, bei Verwendung der Option "Basic" (Einfach) angezeigt.

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. CREATE TABLE MYTABLE(…) COMPRESS; in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option Einfach.
- **Einfach.** Aktiviert einfache Komprimierung (z. B. CREATE TABLE MYTABLE(…) COMPRESS BASIC; in SQL).

Oracle 11gR2-Einstellungen – Option "Advanced" (Erweitert)

Diese Einstellungen werden für Oracle 11g R2, Enterprise oder Personal Edition, bei Verwendung der Option "Advanced" (Erweitert) angezeigt.

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. CREATE TABLE MYTABLE(…) COMPRESS; in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option Einfach.
- **Einfach.** Aktiviert einfache Komprimierung (z. B. CREATE TABLE MYTABLE(…) COMPRESS BASIC; in SQL).
- **OLTP.** Aktiviert OLTP-Komprimierung (z. B. CREATE TABLE MYTABLE(…)COMPRESS FOR OLTP; in SQL).
- **Abfrage niedrig/hoch.** (nur Exadata-Server) Aktiviert Hybrid Columnar Compression für Abfrage (z. B. CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY LOW; oder CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY HIGH; in SQL). Komprimierung für Abfrage ist in Data Warehousing-Umgebungen sinnvoll; HIGH bietet ein höheres Komprimierungsverhältnis als LOW.
- **Archiv niedrig/hoch.** (nur Exadata-Server) Aktiviert Hybrid Columnar Compression für Archiv (z. B. CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE LOW; oder CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE HIGH; in SQL). Komprimierung für Archiv ist sinnvoll zur Komprimierung von Daten, die lange Zeit gespeichert werden sollen; HIGH bietet ein höheres Komprimierungsverhältnis als LOW.

Einstellungen für IBM Netezza, IBM DB2 unter z/OS und Teradata

Wenn Sie Voreinstellungen für IBM Netezza, IBM DB2 unter z/OS oder Teradata festlegen, werden Sie aufgefordert, Folgendes auszuwählen:

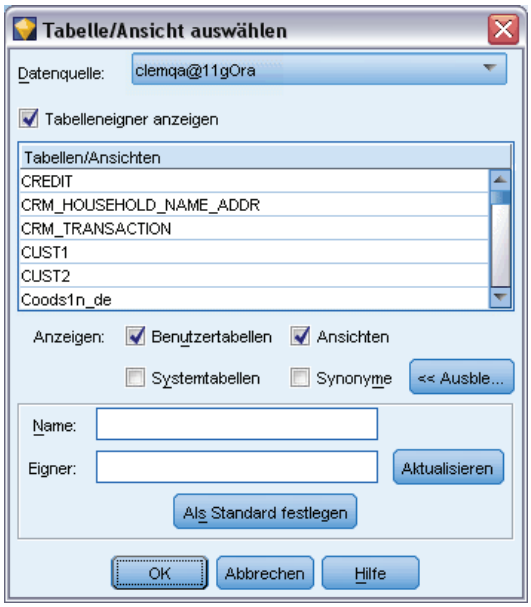
Datenbank/Schema mit Server-Scoring-Adapter verwenden. Wenn ausgewählt, wird die Option Datenbank/Schema mit Server-Scoring-Adapter aktiviert.

Datenbank/Schema mit Server-Scoring-Adapter Wählen Sie die erforderliche Verbindung aus der Dropdown-Liste aus.

Auswählen einer Datenbanktabelle

Nachdem Sie eine Verbindung zu einer Datenquelle hergestellt haben, können Sie wahlweise Felder aus einer bestimmten Tabelle oder Ansicht importieren. Auf der Registerkarte "Daten" des Dialogfelds "Datenbank" können Sie entweder den Namen einer Tabelle in das Feld "Tabellenname" eingeben oder auf Auswählen klicken, um ein Dialogfeld mit einer Liste der verfügbaren Tabellen und Ansichten zu öffnen.

Abbildung 2-8
Auswählen einer Tabelle aus der momentan verbundenen Datenbank



Tabelleneigner anzeigen. Wählen Sie diese Option, wenn für eine Datenquelle die Angabe des Tabellenbesitzers erforderlich ist, damit Sie auf die Tabelle zugreifen können. Deaktivieren Sie diese Option für Datenquellen, die die Angabe des Tabellenbesitzers nicht erfordern.

Hinweis: Für SAS- und Oracle-Datenbanken ist es in der Regel erforderlich, den Tabellenbesitzer anzuzeigen.

Tabellen/Ansichten. Wählen Sie die Tabelle oder Ansicht aus, die Sie importieren möchten.

Anzeigen. Listet die Spalten der Datenquelle auf, mit der Sie verbunden sind. Klicken Sie auf eine der folgenden Optionen, um Ihre Ansicht der verfügbaren Tabellen anzupassen:

- Klicken Sie auf Benutzertabellen, um gewöhnliche, von Datenbankbenutzern erstellte Datenbanktabellen anzuzeigen.
- Klicken Sie auf Systemtabellen, um systemeigene Datenbanktabellen anzuzeigen (dies sind z. B. Tabellen, die Informationen über die Datenbank wie Indexdetails enthalten). Mit dieser Option können Sie die in Excel-Datenbanken verwendeten Register anzeigen. (Beachten Sie, dass auch ein eigener Excel-Quellenknoten verfügbar ist. Für weitere Informationen siehe Thema Excel-Quellenknoten auf S. 52.)
- Klicken Sie auf Ansichten, um virtuelle Tabellen basierend auf einer Abfrage, die eine oder mehrere gewöhnliche Tabellen betrifft, anzuzeigen.
- Klicken Sie auf Synonyme, um Synonyme anzuzeigen, die in der Datenbank für bereits vorhandene Tabellen erstellt wurden.

Namens-/Eignerfilter. Mit diesen Feldern können Sie die Liste der angezeigten Tabellen nach Name oder Besitzer filtern. Geben Sie z. B. SYS ein, um nur Tabellen mit diesem Besitzer aufzulisten. Bei Suchvorgänge mit Platzhalterzeichen steht ein Unterstrich (_) für ein einzelnes Zeichen und ein Prozentzeichen (%) für eine Folge von null oder mehr Zeichen.

Als Standard festlegen. Diese Option speichert die aktuellen Einstellungen als Standardwerte für den aktuellen Benutzer. Diese Einstellungen werden zukünftig wiederhergestellt, wenn ein Benutzer ein neues Dialogfeld zur Auswahl einer Tabelle öffnet. Dies gilt jedoch *nur für denselben Datenquellennamen und dieselbe Benutzeranmeldung*.

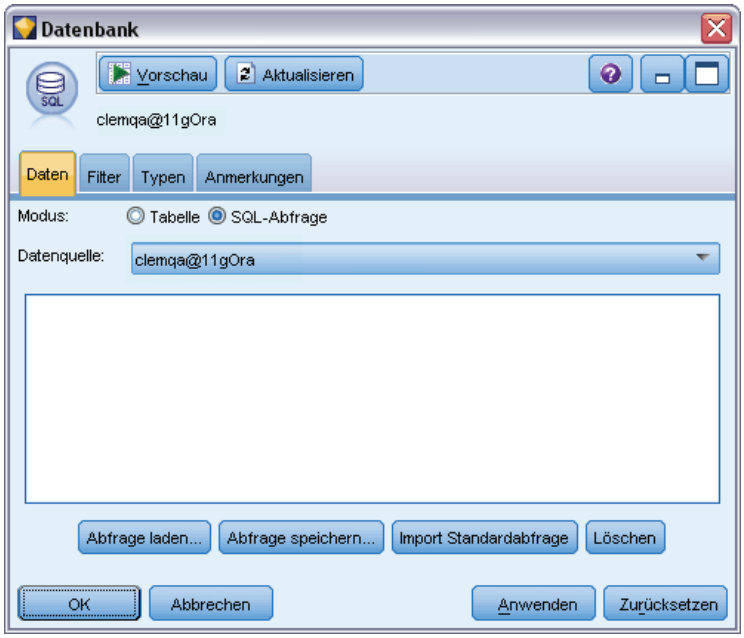
Abfragen der Datenbank

Sobald Sie eine Verbindung zu einer Datenquelle hergestellt haben, können Sie Felder anhand von SQL-Abfragen importieren. Wählen Sie im Hauptdialogfeld SQL-Abfrage als Verbindungsmodus. Dem Dialogfeld wird ein Fenster für den Abfrage-Editor hinzugefügt. Mit dem Abfrage-Editor können Sie eine oder mehrere SQL-Abfragen erstellen oder laden, deren Ergebnis-Set in den Daten-Stream eingelesen wird.

Wenn Sie mehrere SQL-Abfragen angeben, trennen Sie sie durch Strichpunkte (;) und achten Sie darauf, dass es nicht mehrere SELECT-Anweisungen gibt.

Um das Fenster des Abfrage-Editors abzubrechen und zu schließen, wählen Sie Tabelle als Verbindungsmodus.

Abbildung 2-9
Laden von Daten mit SQL-Abfragen



Sie können SPSS Modeler-Stream-Parameter (eine Art benutzerdefinierte Variable) in die SQL-Abfrage mit aufnehmen. Für weitere Informationen siehe Thema Verwenden von Stream-Parametern in einer SQL-Abfrage auf S. 25.

Abfrage laden. Klicken Sie auf diese Option, um den Dateibrowser zu öffnen, mit dem Sie eine bereits gespeicherte Abfrage laden können.

Abfrage speichern. Klicken Sie auf diese Option, um das Dialogfeld “Abfrage speichern” zu öffnen. In diesem Dialogfeld können Sie die aktuelle Abfrage speichern.

Import Standardabfrage. Klicken Sie auf diese Option, um eine SQL SELECT-Beispielanweisung zu importieren, die automatisch anhand der im Dialogfeld ausgewählten Tabelle und Spalten erstellt wird.

Löschen. Löscht den Inhalt des Arbeitsbereichs. Verwenden Sie diese Option, wenn Sie neu beginnen möchten.

Verwenden von Stream-Parametern in einer SQL-Abfrage

Beim Schreiben einer SQL-Abfrage für den Feldimport können Sie zuvor definierte SPSS Modeler-Stream-Parameter mit einschließen. Es werden sämtliche Arten von Stream-Parametern unterstützt.

In der folgenden Tabelle wird angezeigt, wie einige Beispiele für Stream-Parameter in der SQL-Abfrage interpretiert werden.

Tabelle 2-2
Beispiele für Stream-Parameter

Name des Stream-Parameters (Beispiel)	Speicher	Wert des Stream-Parameters	Interpretiert als
PString	Zeichenfolge	ss	’ss’
PInt	Ganzzahl	5	5
PReal	Reelle Zahl	5.5	5.5
PTime	Zeit	23:05:01	t{‘23:05:01’}
PDate	Datum	2011-03-02	d{‘2011-03-02’}
PTimeStamp	Zeitstempel	2011-03-02 23:05:01	ts{‘2011-03-02 23:05:01’}
PColumn	Unbekannt	IntValue	IntValue

In der SQL-Abfrage geben Sie einen Stream-Parameter auf dieselbe Weise an wie in einem CLEM-Ausdruck, nämlich anhand von '$P-<parameter_name>', wobei <parameter_name> der für den Stream-Parameter definierte Name ist.

Beim Verweisen auf ein Feld muss der Speichertyp als “Unbekannt” definiert sein und der Parameterwert muss in Anführungszeichen eingeschlossen sein, falls er benötigt wird. Wenn Sie also unter Verwendung der in der Tabelle angezeigten Beispiele folgende SQL-Abfrage eingeben:

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

würde sie ausgewertet als:

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

Wenn Sie auf das Feld IntValue mit dem Parameter PColumn verweisen, müssen Sie die Abfrage wie folgt angeben, um dasselbe Ergebnis zu erhalten:

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

Knoten "Datei (var.)"

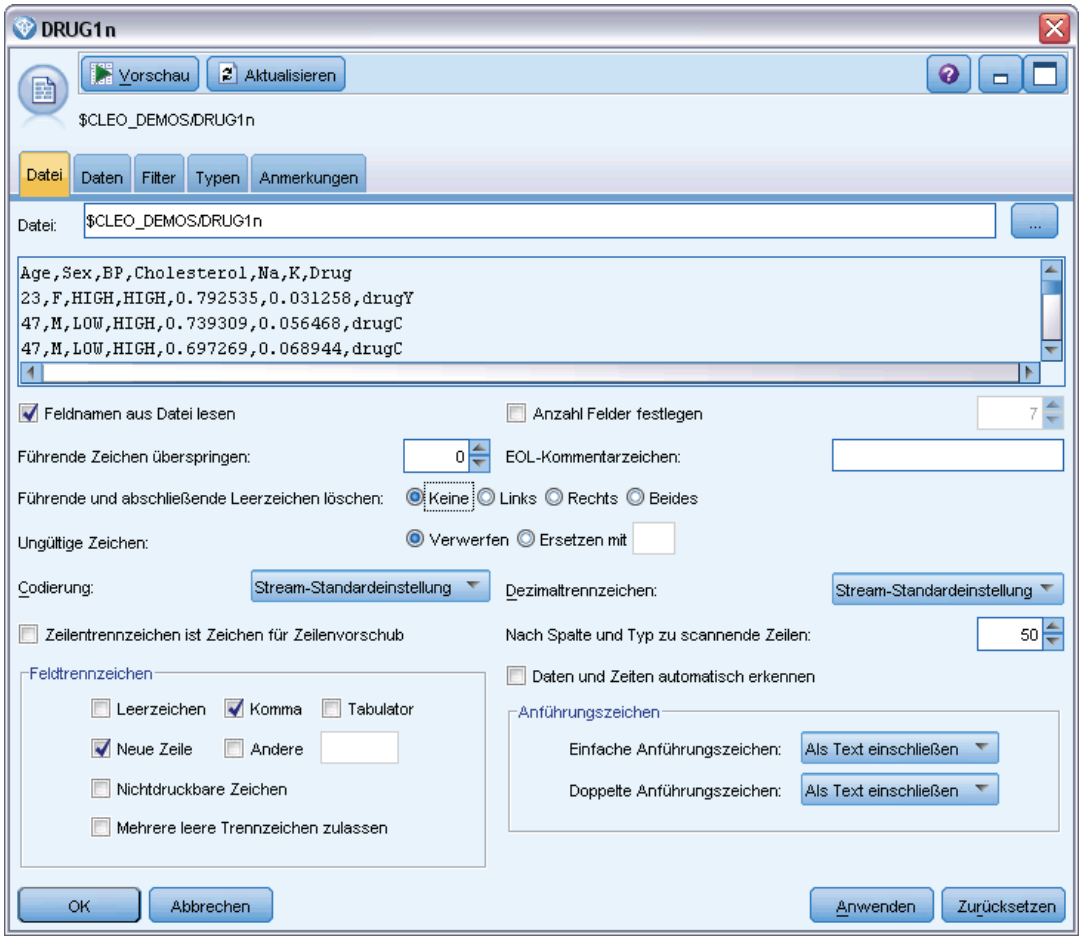
Mit Knoten des Typs "Datei (var.)" können Sie Daten aus Textdateien mit freien Feldern lesen (dies sind Dateien, deren Datensätze eine konstante Anzahl von Feldern und eine variable Anzahl von Zeichen enthalten). Diese Dateien sind auch als Textdateien mit Trennzeichen bekannt. Dieser Knotentyp ist außerdem nützlich für Dateien mit fester Länge, Überschriftentext und bestimmten Anmerkungen. Datensätze werden einzeln nacheinander eingelesen und durch den Stream geleitet, bis die gesamte Datei eingelesen ist.

Hinweise zum Einlesen von Textdaten mit Trennzeichen

- Datensätze müssen durch einen Zeilenumbruch am Ende jeder Zeile getrennt sein. Das Zeilenumbruchzeichen darf für keinen anderen Zweck verwendet werden (beispielsweise innerhalb von Feldnamen oder Werten). Führende und abschließende Leerzeichen sollten idealerweise entfernt werden, um Platz zu sparen. Dies ist jedoch nicht unbedingt erforderlich. Optional können sie auch durch den Knoten entfernt werden.
- Felder müssen durch ein Komma oder ein anderes Zeichen getrennt werden, das idealerweise ausschließlich als Trennzeichen verwendet wird, also nicht in Feldnamen oder Werten vorkommt. Wenn dies nicht möglich ist, können alle Textfelder in doppelte Anführungszeichen gesetzt werden, vorausgesetzt dass keiner der Feldnamen oder Textwerte ein doppeltes Anführungszeichen enthält. Wenn Feldnamen oder Werte doppelte Anführungszeichen enthalten, können die Textfelder alternativ in einfache Anführungszeichen gesetzt werden. Auch hier gilt natürlich wieder die Bedingung, dass einzelne Anführungszeichen nicht bereits an anderen Stellen in Werten verwendet werden. Wenn weder einfache noch doppelte Anführungszeichen verwendet werden können, müssen die Textwerte geändert werden, um entweder das Trennzeichen oder die einfachen bzw. doppelten Anführungszeichen zu entfernen bzw. zu ersetzen.
- Alle Zeilen, einschließlich der Zeile für die Überschrift, sollten die gleiche Anzahl von Feldern enthalten.
- Die erste Zeile sollte die Feldnamen enthalten. Wenn dies nicht der Fall ist, müssen Sie die Auswahl von Feldnamen aus Datei lesen aufheben, um jedem Feld einen allgemeinen Namen zu geben, wie *Feld1, Feld2* usw.
- Die zweite Zeile muss den ersten Datensatz enthalten. Leerzeilen und Kommentare sind nicht zulässig.
- Numerische Werte dürfen kein Tausendertrennzeichen oder Gruppierungssymbol enthalten— – also muss 3.000,00 beispielsweise ohne Punkt geschrieben werden. Das Dezimaltrennzeichen (Komma in Deutschland) darf nur an den entsprechenden Stellen verwendet werden.
- Datums- und Zeitangaben sollten in einem der Format vorliegen, die vom Dialogfeld für die Stream-Optionen erkannt werden, beispielsweise DD/MM/YYYY oder HH:MM:SS. Alle Datums- und Zeitfelder in der Datei sollten idealerweise dasselbe Format verwenden und alle Felder, die ein Datum enthalten, müssen für alle Werte in diesem Feld dasselbe Format verwenden.

Festlegen der Optionen für Knoten "Variable Datei"

Abbildung 2-10
Dialogfeld des Knotens für variable Dateien



Datei. Geben Sie den Namen der Datei an. Zur Auswahl einer Datei können Sie einen Dateinamen eingeben oder auf die Schaltfläche mit den Auslassungspunkten (...) klicken. Der Dateipfad wird angezeigt, sobald Sie eine Datei ausgewählt haben, und der entsprechende Inhalt wird mit Trennzeichen im Fenster darunter angezeigt.

Der von Ihrer Datenquelle angezeigte Beispieltext kann kopiert und in folgende Steuerelemente eingefügt werden: EOL-Kommentarzeichen und benutzerdefinierte Trennzeichen. Verwenden Sie zum Kopieren und Einfügen Strg-C und Strg-V.

Feldnamen aus Datei lesen. Diese standardmäßig ausgewählte Option behandelt die erste Zeile der Datendatei als Beschriftungen für die Spalte. Handelt es sich bei der ersten Zeile nicht um eine Überschrift, deaktivieren Sie die Option, damit jedes Feld im Daten-Set automatisch einen generischen Namen, wie *Feld1, Feld2*, erhält.

Anzahl Felder festlegen. Geben Sie die Anzahl der Felder in jedem Datensatz an. Die Anzahl der Felder kann automatisch ermittelt werden, sofern sich am Ende der Datensätze ein Zeilenumbruch befindet. Sie können auch manuell eine Zahl angeben.

Führende Zeichen überspringen. Legen Sie fest, wie viele Zeichen am Anfang des ersten Datensatzes ignoriert werden sollen.

EOL-Kommentarzeichen Geben Sie Zeichen wie # oder ! ein, um auf Anmerkungen in den Daten hinzuweisen. Wenn ein solches Zeichen in der Datendatei angezeigt wird, werden alle Daten bis zu diesem Zeichen, jedoch nicht einschließlich des nächsten Zeilenumbruchs, ignoriert.

Führende und abschließende Leerzeichen löschen. Wählen Sie die Optionen zum Verwerfen von führenden und abschließenden Leerzeichen in Zeichenketten beim Importieren.

Anmerkung. Vergleiche zwischen Zeichenketten, die SQL-Pushback verwenden oder nicht, können unterschiedliche Ergebnisse generieren, wenn nachgestellte Leerzeichen vorhanden sind.

Ungültige Zeichen. Wählen Sie Verwerfen, um ungültige Zeichen aus der Datenquelle zu entfernen. Wählen Sie Ersetzen mit, um ungültige Zeichen durch das angegebene Symbol (nur ein Zeichen) zu ersetzen. Ungültige Zeichen sind Null-Zeichen bzw. alle Zeichen, die nicht in der angegebenen Kodierungsmethode vorhanden sind.

Kodierung. Gibt die verwendete Textkodierungsmethode an. Sie haben die Wahl zwischen der System-Standardeinstellung, der Stream-Standardeinstellung und UTF-8.

- Die System-Standardeinstellung wird in der Windows-Systemsteuerung bzw. bei Ausführung im verteilten Modus auf dem Server-Computer angegeben.
- Der Stream-Standard wird im Dialogfeld "Stream-Eigenschaften" festgelegt.

Dezimaltrennzeichen. Wählen Sie das in Ihrer Datenquelle verwendete Dezimaltrennzeichen. Die Stream-Standardeinstellung entspricht dem auf der Registerkarte "Optionen" des Dialogfelds "Stream-Eigenschaften" ausgewählten Zeichen. Wählen Sie andernfalls entweder Punkt (.) oder Komma (,), um alle Daten dieses Dialogfelds mit dem ausgewählten Zeichen als Dezimaltrennzeichen zu lesen.

Zeilentrennzeichen ist Zeichen für Zeilenvorschub. Wählen Sie diese Option, um das Zeichen für den Zeilenvorschub als Zeilentrennzeichen anstatt als Feldtrennzeichen zu verwenden. Dies kann beispielsweise dann nützlich sein, wenn es in einer Zeile eine ungerade Anzahl von Trennzeichen gibt, die einen Zeilenumbruch bewirken. Beachten Sie: Wenn Sie diese Option auswählen, können Sie in der Liste der Trennzeichen nicht die Option Neue Zeile auswählen.

Trennzeichen. Mit den für dieses Steuerelement aufgelisteten Kontrollkästchen können Sie angeben, welche Zeichen, z. B. das Komma (,), die Feldbegrenzungen in der Datei definieren. Außerdem können Sie mehr als ein Trennzeichen angeben, z. B. ", |" für Datensätze mit mehreren Trennzeichen. Das Standardtrennzeichen ist das Komma.

Anmerkung: Wenn das Komma auch als Dezimaltrennzeichen definiert wurde, funktionieren die Standardeinstellungen in diesem Fall nicht. Wenn das Komma sowohl als Feldtrennzeichen als auch als Dezimaltrennzeichen festgelegt ist, wählen Sie in der Liste "Trennzeichen" Andere aus. Geben Sie dann manuell ein Komma in das Eingabefeld ein.

Wählen Sie Mehrere leere Trennzeichen zulassen aus, um mehrere nebeneinander liegende leere Trennzeichen als ein einzelnes Trennzeichen zu behandeln. Beispiel: Wenn auf ein Datenwert vier Leerzeichen und ein weiterer Datenwert folgen, wird diese Gruppe als zwei statt fünf Felder betrachtet.

Nach Typ zu durchsuchende Zeilen und Spalten. Legen Sie fest, wie viele Zeilen und Spalten nach angegebenen Datentypen durchsucht werden sollen.

Datum und Uhrzeit automatisch erkennen. Aktivieren Sie dieses Kontrollkästchen, damit IBM® SPSS® Modeler automatisch versucht, Dateneinträge als Datum oder Uhrzeit zu erkennen. Das bedeutet beispielsweise, dass ein Eintrag wie 07-11-1965 als Datum erkannt wird und 02:35:58 als Uhrzeit. Zweideutige Einträge wie 07111965 oder 023558 werden jedoch als Ganzzahlen angezeigt, da die Zahlen nicht durch Trennzeichen separiert sind.

Hinweis: Um mögliche Datenprobleme bei der Verwendung von Datendateien älterer Versionen von SPSS Modeler zu vermeiden, ist dieses Kontrollkästchen standardmäßig für Informationen deaktiviert, die in älteren Versionen als 13 gespeichert wurden.

Anführungszeichen. Mit den Dropdown-Listen können Sie angeben, wie einfache und doppelte Anführungszeichen beim Importieren zu behandeln sind. Sie können alle Fragezeichen Verwerfen, Als Text einschließen, indem Sie sie in den Feldwert einschließen, oder Paaren und verwerfen, um Anführungszeichenpaare zu finden und zu löschen. Kann einem Anführungszeichen kein zweites Anführungszeichen zugeordnet werden, wird eine Fehlermeldung ausgegeben. Sowohl die Option Verwerfen als auch Paaren und verwerfen speichert den Feldwert (ohne Anführungszeichen) als Zeichenkette.

Klicken Sie bei der Bearbeitung dieses Dialogfelds zu einem beliebigen Zeitpunkt auf Aktualisieren, um Daten aus der Datenquelle neu zu laden. Diese Funktion ist nützlich, wenn Sie Datenverbindungen zum Quellenknoten ändern oder wenn Sie die verschiedenen Registerkarten des Dialogfelds bearbeiten.

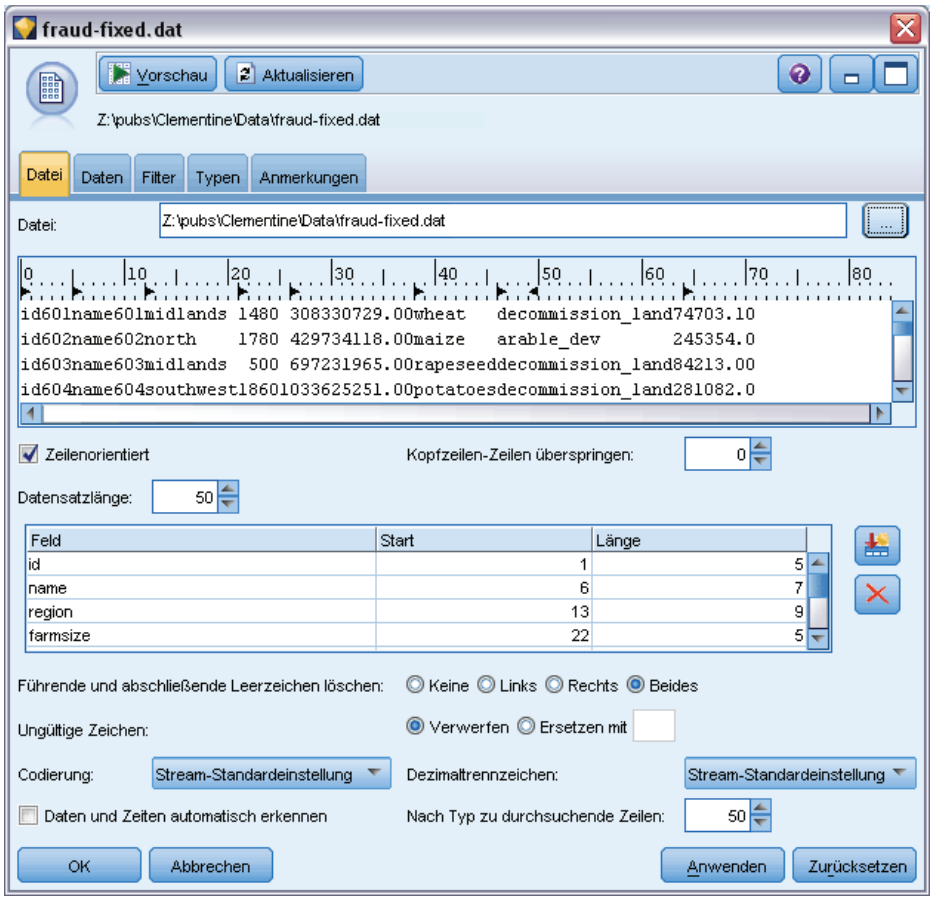
Knoten "Datei (fest)"

Mit Knoten des Typs "Datei (fest)" können Sie Daten aus Textdateien mit festen Feldern importieren (dies sind Dateien, deren Felder nicht begrenzt sind, sondern an derselben Position beginnen und eine feste Länge haben). Maschinell erzeugte Daten oder Legacydaten werden häufig im Format mit festen Feldern gespeichert. Anhand der Registerkarte "Datei" des Knotens "Datei (fest)" können Sie problemlos die Position und Länge der Spalten Ihrer Daten angeben.

Festlegen der Optionen für den Knoten "Datei (fest)"

Auf der Registerkarte "Datei" des Knotens "Datei (fest)" können Sie Daten in IBM® SPSS® Modeler importieren und die Spaltenposition und Datensatzlänge angeben. Klicken Sie im Datenvorschaufenster in der Mitte des Dialogfelds, um Pfeile hinzuzufügen, mit denen die Haltepunkte zwischen den Feldern angegeben werden.

Abbildung 2-11
Festlegen von Spalten in Daten mit festen Feldern



Datei. Geben Sie den Namen der Datei an. Zur Auswahl einer Datei können Sie einen Dateinamen eingeben oder auf die Schaltfläche mit den Auslassungspunkten (...) klicken. Sobald Sie eine Datei ausgewählt haben, wird der Dateipfad angezeigt und der entsprechende Inhalt wird mit Trennzeichen im Fenster unten angezeigt.

Im Datenvorschaufenster können Sie die Spaltenposition und Länge festlegen. Das Lineal am oberen Rand des Vorschaufensters unterstützt Sie beim Messen der Länge der Variablen und beim Festlegen des Haltepunkts zwischen den Variablen. Sie können Haltepunktlinien festlegen, indem Sie in den Linealbereich oberhalb der Felder klicken. Haltepunkte können durch Ziehen verschoben werden. Um sie zu verwerfen, ziehen Sie sie aus dem Datenvorschaubereich.

- Jede Haltepunktlinie fügt automatisch ein neues Feld zur Feldtabelle hinzu.
- Durch die Pfeile markierte Startpositionen werden automatisch zur Startspalte in der Tabelle unten hinzugefügt.

Zeilenorientiert. Wählen Sie diese Option, wenn Sie das Zeilenwechselzeichen am Ende jedes Datensatzes überspringen möchten.

Kopfzeilen überspringen. Legen Sie fest, wie viele Zeilen am Anfang des ersten Datensatzes ignoriert werden sollen. Diese Funktion ist nützlich, um Spaltenkopfzeilen zu ignorieren.

Datensatzlänge. Geben Sie die Zahl der Zeichen in jedem Datensatz an.

Feld. Alle Felder, die Sie für diese Datendatei definiert haben, werden hier aufgelistet. Es gibt zwei Methoden für das Definieren von Feldern:

- Felder interaktiv anhand des Datenvorschaufensters festlegen.
- Felder manuell durch Hinzufügen leerer Feldzeilen zur Tabelle unten festlegen. Klicken Sie auf die Schaltfläche rechts neben dem Feldfenster, um neue Felder hinzuzufügen. Geben Sie anschließend einen Feldnamen, eine Start-Position und eine Länge in das leere Feld ein. Mit diesen Optionen werden automatisch Pfeile zum Datenvorschaufenster hinzugefügt, die problemlos angepasst werden können.

Um ein bereits definiertes Feld zu löschen, wählen Sie das Feld in der Liste aus und klicken Sie auf die rote Löschschaltfläche.

Start. Legen Sie die Position des ersten Zeichens im Feld fest. Beispiel: Wenn das zweite Feld eines Datensatzes beim sechzehnten Zeichen beginnt, geben Sie 16 als Startwert ein.

Länge. Legen Sie fest, wie viele Zeichen sich im längsten Wert für jedes Feld befinden. Dadurch wird der Abbruchpunkt für das nächste Feld bestimmt.

Führende und abschließende Leerzeichen löschen. Wählen Sie diese Option aus, um führende und abschließende Leerzeichen in Zeichenketten beim Importieren zu verwerfen.

Anmerkung. Vergleiche zwischen Zeichenketten, die SQL-Pushback verwenden oder nicht, können unterschiedliche Ergebnisse generieren, wenn nachgestellte Leerzeichen vorhanden sind.

Ungültige Zeichen. Wählen Sie Verwerfen, um ungültige Zeichen aus der Dateneingabe zu entfernen. Wählen Sie Ersetzen mit, um ungültige Zeichen durch das angegebene Symbol (nur ein Zeichen) zu ersetzen. Ungültige Zeichen sind Null-Zeichen (0) bzw. alle Zeichen, die nicht in der aktuellen Kodierung vorhanden sind.

Kodierung. Gibt die verwendete Textkodierungsmethode an. Sie haben die Wahl zwischen der System-Standardeinstellung, der Stream-Standardeinstellung und UTF-8.

- Die System-Standardeinstellung wird in der Windows-Systemsteuerung bzw. bei Ausführung im verteilten Modus auf dem Server-Computer angegeben.
- Der Stream-Standard wird im Dialogfeld “Stream-Eigenschaften” festgelegt.

Dezimaltrennzeichen. Wählen Sie das in Ihrer Datenquelle verwendete Dezimaltrennzeichen. Stream-Standardeinstellung entspricht dem auf der Registerkarte “Optionen” des Dialogfelds “Stream-Eigenschaften” ausgewählten Zeichen. Wählen Sie andernfalls entweder Punkt (.) oder Komma (,), um alle Daten dieses Dialogfelds mit dem ausgewählten Zeichen als Dezimaltrennzeichen zu lesen.

Datum und Uhrzeit automatisch erkennen. Aktivieren Sie dieses Kontrollkästchen, damit SPSS Modeler automatisch versucht, Dateneinträge als Datum oder Uhrzeit zu erkennen. Das bedeutet beispielsweise, dass ein Eintrag wie 07-11-1965 als Datum erkannt wird und 02:35:58 als Uhrzeit. Zweideutige Einträge wie 07111965 oder 023558 werden jedoch als Ganzzahlen angezeigt, da die Zahlen nicht durch Trennzeichen separiert sind.

Hinweis: Um mögliche Datenprobleme bei der Verwendung von Datendateien älterer Versionen von SPSS Modeler zu vermeiden, ist dieses Kontrollkästchen standardmäßig für Informationen deaktiviert, die in älteren Versionen als 13 gespeichert wurden.

Nach Typ zu durchsuchende Zeilen. Legen Sie fest, wie viele Zeilen nach angegebenen Datentypen durchsucht werden sollen.

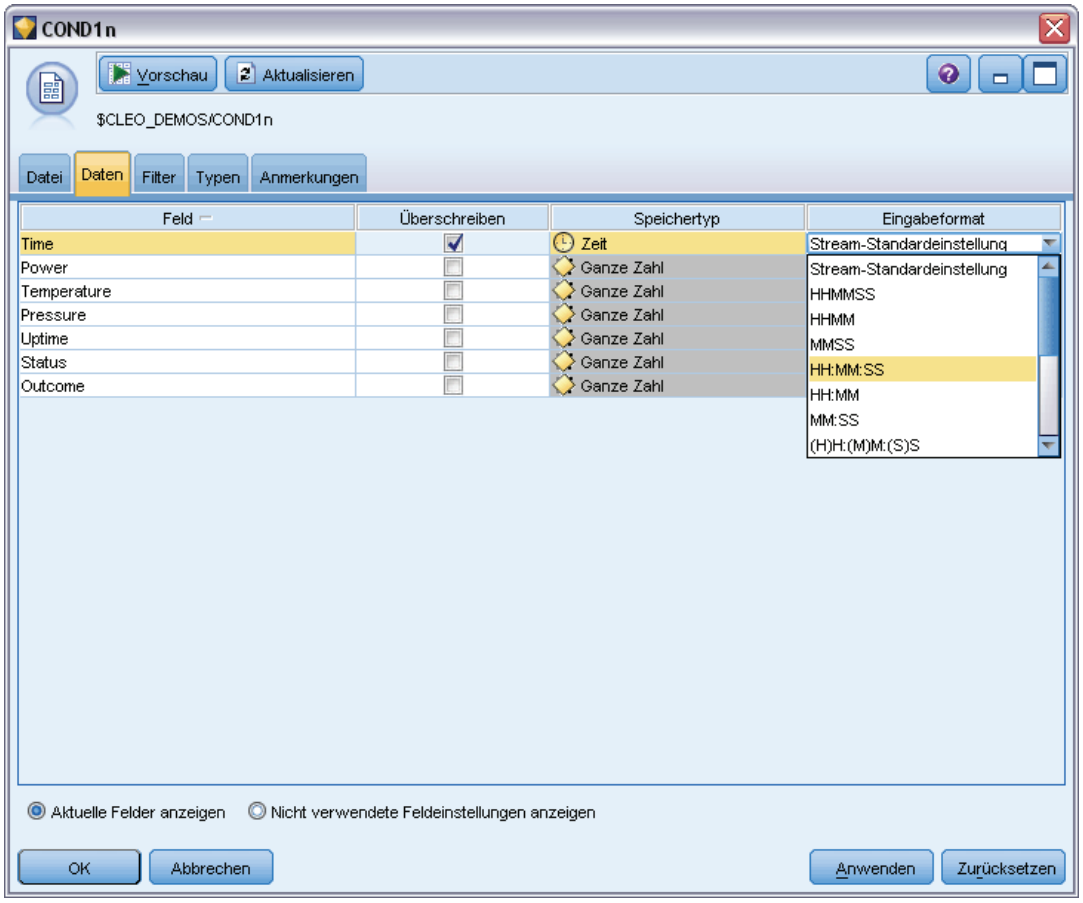
Klicken Sie bei der Bearbeitung dieses Dialogfelds zu einem beliebigen Zeitpunkt auf Aktualisieren, um Daten aus der Datenquelle neu zu laden. Diese Funktion ist nützlich, wenn Sie Datenverbindungen zum Quellenknoten ändern oder wenn Sie die verschiedenen Registerkarten des Dialogfelds bearbeiten.

Festlegen von Feldspeichertyp und Formatierung

Mit den Optionen auf der Registerkarte “Daten” für die Knoten “Datei (fest)” und “Datei (var.)”, “XML-Quelle” und “Eingabe” können Sie den Speichertyp für Felder festlegen, die in IBM® SPSS® Modeler importiert oder erstellt werden. Für die Knoten “Datei (fest)”, “Datei (var.)” und “Eingabe” können Sie außerdem die Feldformatierung und andere Metadaten festlegen.

Bei aus anderen Quellen eingelesenen Daten wird der Speichertyp automatisch ermittelt, kann jedoch mithilfe einer Konvertierungsfunktion, wie beispielsweise to_integer, in einem Füller- oder Ableitungsknoten geändert werden.

Abbildung 2-12
Überschreiben von Speichertyp und Feldformatierung beim Importieren



Feld. Mit der Spalte *Feld* zeigen Sie Felder im aktuellen Daten-Set an und wählen sie aus.

Überschreiben. Aktivieren Sie das Kontrollkästchen in der Spalte *Überschreiben*, um die Optionen in den Spalten *Speichertyp* und *Eingabeformat* zu aktivieren.

Datenspeichertyp

Der Speichertyp beschreibt die Art und Weise, wie Daten in einem Feld gespeichert werden. Beispiel: Ein Feld mit den Werten 1 und 0 speichert ganzzahlige Daten. Dies ist vom Messniveau zu unterscheiden, das die Verwendung der Daten beschreibt und sich nicht auf den Speichertyp auswirkt. Beispiel: Sie möchten das Messniveau für ein Feld ganzer Zahlen mit den Werten 1 und 0 auf *Flag* setzen. Das bedeutet normalerweise, dass 1=*True* und 0=*False* ist. Während der Speichertyp stets an der Quelle festgelegt werden muss, kann das Messniveau mithilfe eines Typknotens an jeder beliebigen Stelle im Stream geändert werden. Für weitere Informationen siehe Thema Messniveaus in Kapitel 4 auf S. 138.

Folgende Speichertypen sind verfügbar:

- **String.** Wird für Felder verwendet, die nicht numerische Daten enthalten (auch als alphanumerische Daten bezeichnet). Eine Zeichenkette kann jede beliebige Abfolge von Zeichen enthalten, beispielsweise *fred*, *Klasse 2* oder *1234*. Beachten Sie, dass die Zahlen in Zeichenketten nicht für Berechnungen verwendet werden können.
- **Ganze Zahl.** Ein Feld, bei dessen Werten es sich um ganze Zahlen handelt.
- **Reelle Zahl.** Bei den Werten handelt es sich um Zahlen, die Dezimalstellen enthalten können (nicht auf ganze Zahlen beschränkt). Das Anzeigeformat wird im Dialogfeld für die Stream-Eigenschaften angegeben und kann für einzelne Felder in einem Typknoten überschrieben werden (Registerkarte "Format").
- **Datum.** Datumswerte, angegeben in einem Standardformat, wie Jahr, Monat und Tag (z. B. 2007-09-26). Das jeweilige Format wird im Dialogfeld für die Stream-Eigenschaften angegeben.
- **Uhrzeit.** Als Dauer gemessene Zeit. Beispielsweise kann ein Service-Call, der 1 Stunde, 26 Minuten und 38 Sekunden dauerte, als 01:26:38 angegeben werden, je nachdem, welches Zeitformat aktuell im Dialogfeld für die Stream-Eigenschaften angegeben ist.
- **Zeitstempel.** Werte, die sowohl eine Datums- als auch eine Zeitkomponente enthalten, wie beispielsweise 2007–09–26 09:04:00; auch hier wieder abhängig von den aktuellen Formaten für Datum und Zeit im Dialogfeld "Stream-Eigenschaften". Beachten Sie, dass Zeiststempelwerte ggf. in Anführungszeichen gesetzt werden müssen, um sicherzustellen, dass sie als Einzelwert interpretiert werden und nicht als gesonderte Datums- und Zeitwerte. (Dies gilt beispielsweise bei der Eingabe von Werten in einem Benutzereingabeknoten.)

Speichertypkonvertierung. Der Speichertyp für ein Feld kann mit verschiedenen Konvertierungsfunktionen, z. B. to_string und to_integer, in einem Füllerknoten geändert werden. Für weitere Informationen siehe Thema Speichertypkonvertierung mithilfe des Füllerknotens in Kapitel 4 auf S. 181. Beachten Sie, dass die Konvertierungsfunktionen (und alle anderen Funktionen, für die ein spezieller Eingabetyp, wie beispielsweise ein Wert für Datum oder Uhrzeit, erforderlich ist) von den aktuell im Dialogfeld für die Stream-Eigenschaften angegebenen Formaten abhängen. Wenn Sie beispielsweise ein Zeichenkettenfeld mit den Werten *Jan 2003*, *Feb 2003* (usw.) in einen Datumsspeicher konvertieren müssen, wählen Sie MON JJJJ

als Standard-Datumsformat für den Stream aus. Konvertierungsfunktionen sind auch im Ableitungsknoten zur temporären Konvertierung während einer Ableitungsberechnung verfügbar. Mit dem Ableitungsknoten können Sie auch andere Bearbeitungen vornehmen wie beispielsweise die Umkodierung von Zeichenkettenfeldern mit kategorialen Werten. Für weitere Informationen siehe Thema Umkodieren von Werten mit dem Ableitungsknoten in Kapitel 4 auf S. 178.

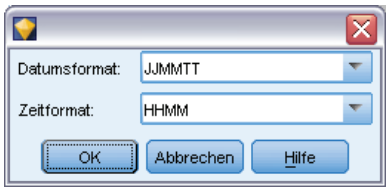
Einlesen gemischter Daten. Beachten Sie, dass beim Einlesen von Feldern mit numerischem Speichertyp (ganze Zahl, reelle Zahl, Zeit, Zeitstempel oder Datum) alle nicht numerischen Werte auf null oder auf systemdefiniert fehlend gesetzt werden. Dies liegt daran, dass SPSS Modeler im Gegensatz zu einigen anderen Anwendungen keine gemischten Speichertypen innerhalb eines Felds zulässt. Um dies zu vermeiden, sollten alle Felder mit gemischten Daten als Zeichenketten eingelesen werden, indem der Speichertyp im Quellenknoten oder in der externen Anwendung nach Bedarf geändert wird.

Feldeingabeformat (nur für die Knoten "Datei (fest)", "Datei (var.)" und "Eingabe")

Sie können für alle Speichertypen außer "Zeichenkette" und "Ganze Zahl" anhand der Dropdown-Liste Formatierungsoptionen für das ausgewählte Feld festlegen. Beispiel: Beim Verbinden von Daten verschiedener Ländereinstellungen müssen Sie einen Punkt (.) als Dezimaltrennzeichen für ein Feld festlegen, während ein anderes Feld ein Komma als Trennzeichen erfordert.

Im Quellenknoten festgelegte Eingabeoptionen überschreiben die Formatierungsoptionen, die im Dialogfeld "Stream-Eigenschaften" definiert sind. Sie sind jedoch später im Stream nicht persistent. Ihr Zweck besteht darin, Eingaben basierend auf Ihrem Wissen über die Daten korrekt zu analysieren. Die festgelegten Formate dienen als Richtlinie für die Analyse der Daten beim Einlesen in SPSS Modeler und bestimmen nicht das Format nach dem Einlesen in SPSS Modeler. Um die Formatierung für die einzelnen Felder an anderer Stelle im Stream festzulegen, verwenden Sie die Registerkarte "Format" eines Typknotens. Für weitere Informationen siehe Thema Feldformat – Registerkarte "Einstellungen" in Kapitel 4 auf S. 153.

Abbildung 2-13
Festlegen von Datums- und Zeitformaten für Zeitstempelfelder



Die Optionen sind je nach Speichertyp verschieden. Für den Speichertyp "Reelle Zahl" können Sie z. B. Punkt (.) oder Komma (,) als Dezimaltrennzeichen auswählen. Für Zeitstempelfelder wird ein separates Dialogfeld geöffnet, wenn Sie in der Dropdown-Liste Angeben wählen. Für weitere Informationen siehe Thema Festlegen der Feldformatierungsoptionen in Kapitel 4 auf S. 154.

Für alle Speichertypen können Sie auch Stream-Standardeinstellung wählen, um die Stream-Standardeinstellungen für den Import zu verwenden. Stream-Einstellungen werden im Dialogfeld "Stream-Eigenschaften" festgelegt.

Weitere Optionen

Auf der Registerkarte “Daten” können einige andere Optionen festgelegt werden:

- Zum Anzeigen von Speichertypeinstellungen für Daten, die nicht mehr über den aktuellen Knoten verbunden sind (z. B. Trainingsdaten), wählen Sie Nicht verwendete Feldeinstellungen anzeigen. Sie können die Legacyfelder löschen, indem Sie auf Löschen klicken.
- Klicken Sie bei der Bearbeitung dieses Dialogfelds zu einem beliebigen Zeitpunkt auf Aktualisieren, um Daten aus der Datenquelle neu zu laden. Diese Funktion ist nützlich, wenn Sie Datenverbindungen zum Quellenknoten ändern oder wenn Sie die verschiedenen Registerkarten des Dialogfelds bearbeiten.

Data Collection Knoten

Data Collection-Quellenknoten importieren Umfragedaten auf der Grundlage des IBM® SPSS® Data Collection Survey Reporter Developer Kit, das von der IBM Corp.-Marktforschungssoftware verwendet wird. Bei diesem Format wird zwischen **Falldaten**—, den tatsächlichen Antworten auf Fragen, die während einer Umfrage gesammelt wurden — und **Metadaten** unterschieden, die beschreiben, wie die Falldaten gesammelt und organisiert werden. Metadaten bestehen aus Informationen wie Fragetexten, Variablennamen und -beschreibungen, Variablendefinitionen für Mehrfachantworten, Übersetzungen der verschiedenen Textzeichenketten und der Definition der Struktur der Falldaten.

Hinweis: Für diesen Knoten ist das Data Collection Survey Reporter Developer Kit erforderlich, das zusammen mit Data Collection-Softwareprodukten von IBM Corp. ausgeliefert wird. Weitere Informationen finden Sie auf der Data Collection-Webseite unter *http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/*. Abgesehen von der Installation des Developer Kit ist keine weitere Konfiguration erforderlich.

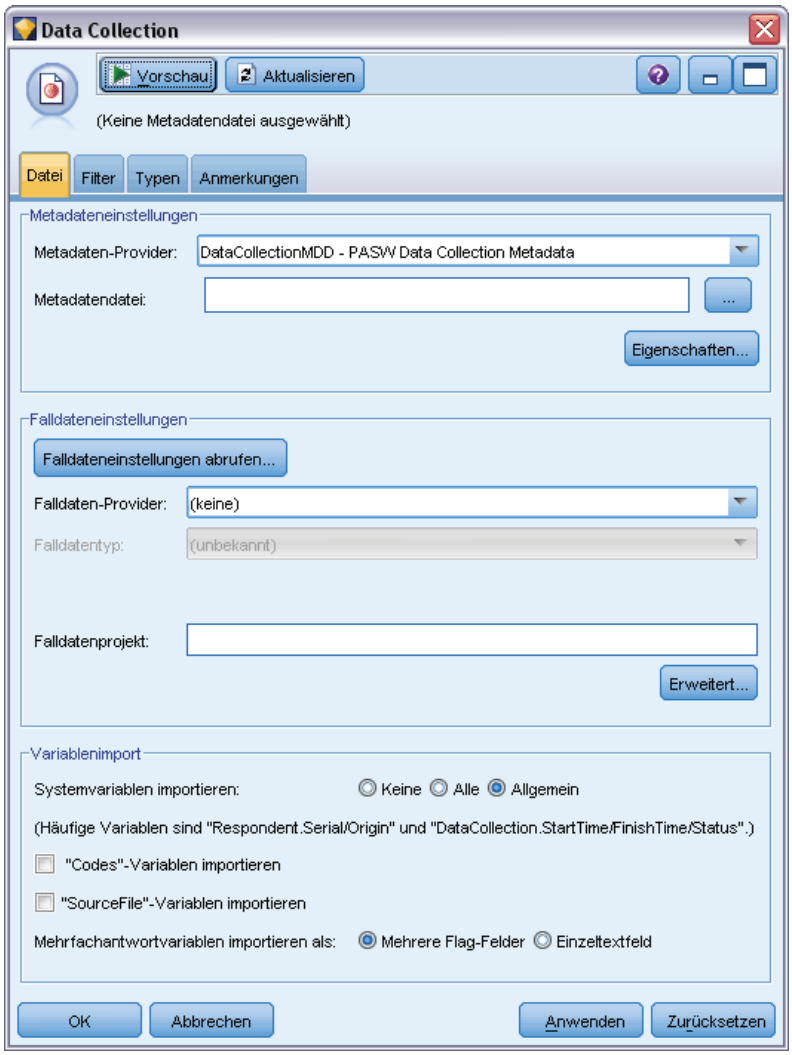
Kommentare

- Umfragedaten werden aus dem einfachen VDATA-Tabellenformat eingelesen oder aus Datenquellen im hierarchischen HDATA-Format, sofern diese eine Metadatenquelle beinhalten (erfordert Data Collection 4.5 oder höher).
- Die Typen werden automatisch mithilfe von Informationen aus den Metadaten instanziiert.
- Wenn Umfragedaten in IBM® SPSS® Modeler importiert werden, werden Fragen als Felder wiedergegeben, wobei für jeden Befragten ein Datensatz verwendet wird.

Dateioptionen für den Data Collection-Import

Auf der Registerkarte “Datei” im Data Collection-Knoten können Sie Optionen für die zu importierenden Metadaten und Falldaten angeben.

Abbildung 2-14
Data Collection-Quellenknoten – Dateioptionen



Metadateneinstellungen

Hinweis: Um die vollständige Liste der verfügbaren Provider-Dateitypen zu sehen, müssen Sie das IBM® SPSS® Data Collection Survey Reporter Developer Kit installieren, das mit der Data Collection-Software zur Verfügung steht. Weitere Informationen finden Sie auf der Data Collection-Webseite unter *http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/*.

Metadaten-Provider. Die Umfragedaten können aus einer Reihe von Formaten importiert werden, die von der Data Collection Survey Reporter Developer Kit-Software unterstützt werden. Folgende Provider-Typen werden unterstützt:

- DataCollectionMDD. Liest Metadaten aus einer Fragebogendefinitionsdatei (*.mdd*) ein. Dies ist das standardmäßige Data Collection Data Model-Format.

- ADO-Datenbank. Liest Falldaten und Metadaten aus ADO-Dateien ein. Geben Sie Namen und Speicherort der *.adoinfo*-Datei an, die die Metadaten enthält. Der interne Name dieser DSC lautet *mrADODsc*.
- In2data-Datenbank. Liest In2data-Falldaten und Metadaten. Der interne Name dieser DSC lautet *mrI2dDsc*.
- Datensammlungs-Protokolldatei. Liest Metadaten aus einer Standard-Data Collection-Protokolldatei. Typischerweise tragen Protokolldatei die Dateinamenerweiterung *.tmp*. Einige Protokolldateien können jedoch eine andere Dateinamenerweiterung aufweisen. Falls erforderlich, können Sie die Datei umbenennen, sodass sie die Dateinamenerweiterung *.tmp* erhält. Der interne Name dieser DSC lautet *mrLogDsc*.
- Quancept-Definitionsdatei. Konvertiert Metadaten in ein Quancept-Skript. Geben Sie den Namen der Quancept-Datei (*.qdi*) an. Der interne Name dieser DSC lautet *mrQdiDrsDsc*.
- Quanvert-Datenbank. Liest Quanvert-Falldaten und Metadaten. Geben Sie Namen und Speicherort der Datei *.qvinfo* bzw. *.pkd* an. Der interne Name dieser DSC lautet *mrQvDsc*.
- Datensammlungs-Teilnahmedatenbank. Liest die Stichproben- und Verlaufstabellen eines Projekts ein und erstellt abgeleitete kategoriale Variablen, die den Spalten in diesen Tabellen entsprechen. Der interne Name dieser DSC lautet *mrSampleReportingMDSC*.
- Statistikdatei. Liest Falldaten und Metadaten aus einer IBM® SPSS® Statistics-Datei (*.sav*) ein. Schreibt Falldaten zur Analyse in SPSS Statistics in eine SPSS Statistics-Datei (*.sav*). Schreibt Metadaten aus einer SPSS Statistics-Datei (*.sav*) in eine *.mdd*-Datei. Der interne Name dieser DSC lautet *mrSavDsc*.
- Surveycraft-Datei. Liest SurveyCraft-Falldaten und Metadaten. Geben Sie den Namen der SurveyCraft-Datei (*.vq*) an. Der interne Name dieser DSC lautet *mrSCDsc*.
- Datensammlungs-Skriptdatei. Liest aus Metadaten in einer mrScriptMetadata-Datei. Typischerweise tragen diese Dateien die Dateinamenerweiterung *.mdd* bzw. *.dms*. Der interne Name dieser DSC lautet *mrScriptMDSC*.
- Triple-S-XML-Datei. Liest Metadaten aus einer Triple-S-Datei im XML-Format. Der interne Name dieser DSC lautet *mrTripleSDsc*.

Metadateneigenschaften. Wählen Sie optional Eigenschaften aus, um die zu importierende Umfrageversion sowie die Sprache, den Kontext und den Beschriftungstyp anzugeben, die verwendet werden sollen. Für weitere Informationen siehe Thema IBM SPSS Data Collection-Import – Metadateneigenschaften auf S. 39.

Falldateneinstellungen

Hinweis: Um die vollständige Liste der verfügbaren Provider-Dateitypen zu sehen, müssen Sie das Data Collection Survey Reporter Developer Kit installieren, das mit der Data Collection-Software zur Verfügung steht. Weitere Informationen finden Sie auf der Data Collection-Webseite unter *http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/*.

Falldateneinstellungen abrufen. Wenn Sie Metadaten ausschließlich aus *.mdd*-Dateien einlesen, klicken Sie auf Falldateneinstellungen abrufen, um zu bestimmen, welche Falldatenquellen den ausgewählten Metadaten zugeordnet sind, sowie die konkreten Einstellungen, die für den Zugriff auf eine bestimmte Quelle erforderlich sind. Diese Option ist nur für *.mdd*-Dateien verfügbar.

Falldaten-Provider. Folgende Provider-Typen werden unterstützt:

- ADO-Datenbank. Liest Falldaten mithilfe der Microsoft ADO-Schnittstelle. Wählen Sie OLE-DB UDL für den Falldatentyp aus und geben Sie eine Verbindungszeichenkette im Feld "Falldaten-UDL" an. Für weitere Informationen siehe Thema Datenbankverbindungszeichenkette auf S. 40. Der interne Name dieser Komponente lautet *mrADODsc*.
- Textdatei mit Trennzeichen (Excel). Liest Falldaten aus einer kommagetrennten Datei (.CSV), wie sie von Excel ausgegeben werden kann. Der interne Name lautet *mrCsvDsc*.
- Datensammlungs-Datendatei. Liest Falldaten aus einer Datei im systemeigenen Data Collection-Datenformat (ab Data Collection 4.5). Der interne Name lautet *mrDataFileDsc*.
- In2data-Datenbank. Liest Falldaten und Metadaten aus einer In2data-Datenbank (*.i2d*) ein. Der interne Name lautet *mrI2dDsc*.
- Datensammlungs-Protokolldatei. Liest Falldaten aus einer Standard-Data Collection-Protokolldatei. Typischerweise tragen Protokolldatei die Dateinamenerweiterung *.tmp*. Einige Protokolldateien können jedoch eine andere Dateinamenerweiterung aufweisen. Falls erforderlich, können Sie die Datei umbenennen, sodass sie die Dateinamenerweiterung *.tmp* erhält. Der interne Name lautet *mrLogDsc*.
- Quantum-Datendatei. Liest Daten aus einer ASCII-Datei im Quantum-Format (*.dat*). Der interne Name lautet *mrPunchDsc*.
- Quancept-Datendatei. Liest Daten aus einer Quancept-Datei (*.drs*, *.drz* bzw..*dru*). Der interne Name lautet *mrQdiDrsDsc*.
- Quanvert-Datenbank. Liest Falldaten aus einer Quanvert-Datei (*qvinfo* bzw. *.pkd*). Der interne Name lautet *mrQvDsc*.
- Datensammlungs-Datenbank (MS SQL Server). Liest Falldaten in eine relationale Microsoft SQL Server-Datenbank ein. Für weitere Informationen siehe Thema Datenbankverbindungszeichenkette auf S. 40. Der interne Name lautet *mrRdbDsc2*.
- Statistikdatei. Liest Falldaten aus einer SPSS Statistics-Datei (*.sav*) ein. Der interne Name lautet *mrSavDsc*.
- Surveycraft-Datei. Liest Falldaten aus einer SurveyCraft-Datei (*.qdt*) ein. Die *.vq*- und *.qdt*-Dateien müssen sich im selben Verzeichnis befinden und es muss für beide Dateien Lese- und Schreibzugriff bestehen. Dies ist nicht die Standardvorgehensweise bei der Erstellung mit SurveyCraft. Daher muss eine der Dateien verschoben werden, um SurveyCraft-Daten importieren zu können. Der interne Name lautet *mrScDsc*.
- Triple-S-Datendatei. Liest Daten aus einer Triple-S-Datendatei, entweder im Format mit fester Länge oder im kommagetrennten Format. Der interne Name lautet *mr TripleDsc*.
- Datensammlungs-XML. Liest Falldaten aus einer Data Collection XML-Datendatei. Typischerweise kann dieses Format zur Übertragung von Falldaten von einem Speicherort an einen anderen verwendet werden. Der interne Name lautet *mrXmlDsc*.

Falldatentyp. Gibt an, ob Falldaten aus einer Datei, einem Ordner, aus OLE-DB UDL oder ODBC DSN gelesen werden sollen, und aktualisiert die Dialogfeldoptionen entsprechend. Welche Optionen gültig sind, hängt vom Provider-Typ ab. Bei Datenbank-Providern können Sie Optionen für die OLE-DB- bzw. ODBC-Verbindung angeben. Für weitere Informationen siehe Thema Datenbankverbindungszeichenkette auf S. 40.

Falldatenprojekt. Beim Lesen von Falldaten aus einer Data Collection-Datenbank, können Sie den Namen des Projekts eingeben. Bei allen anderen Falldatentypen sollte diese Einstellung leer bleiben.

Variablenimport

Systemvariablen importieren. Gibt an, ob Systemvariablen importiert werden sollen, einschließlich Variablen, die den Befragungsstatus angeben (läuft, abgeschlossen, Fertigstellungsdatum usw.). Sie haben die Auswahl zwischen Keine, Alle und Benutzerdefiniert.

"Codes"-Variablen importieren. Steuert den Import von Variablen, die Codes darstellen, die für offene Antworten vom Typ "Andere" bei kategorialenVariablen verwendet werden.

"SourceFile"-Variablen importieren. Steuert den Import von Variablen, die Dateinamen oder Bilder von gescannten Antworten enthalten.

Mehrfachantwortvariablen importieren als. Mehrfachantwortvariablen können als mehrere Flag-Felder (Set aus dichotomen Variablen) importiert werden (dies ist die Standardmethode für neue Streams). In Versionen von IBM® SPSS® Modeler vor 12.0 erstellte Streams importierten Mehrfachantworten in ein einzelnes Feld. Die Werte wurden dabei durch Kommas getrennt. Die ältere Methode wird weiterhin unterstützt, damit bestehende Streams weiterhin wie gehabt ausgeführt werden können, es wird jedoch empfohlen, ältere Streams für die Verwendung der neuen Methode zu aktualisieren. Für weitere Informationen siehe Thema Importieren von Mehrfachantworten-Sets auf S. 41.

IBM SPSS Data Collection-Import – Metadateneigenschaften

Beim Import der IBM® SPSS® Data Collection-Umfragedaten können Sie die zu importierende Umfrageversion sowie die Sprache, den Kontext und den Beschriftungstyp angeben, die verwendet werden sollen. Beachten Sie, dass jeweils nur eine Sprache, ein Kontext und ein Beschriftungstyp importiert werden kann.

Abbildung 2-15
IBM SPSS Data Collection-Import – Metadateneigenschaften

Version. Jede Umfrageversion lässt sich als Snapshot der für die Sammlung eines bestimmten Falldaten-Sets verwendeten Metadaten betrachten. Wenn sich ein Fragebogen ändert, können mehrere Versionen erstellt werden. Sie können die aktuellste Version, alle Versionen oder eine bestimmte Version erstellen.

- **Alle Versionen.** Wählen Sie diese Option, wenn eine Kombination (Obermenge) aller verfügbaren Versionen verwendet werden soll. (Dies wird manchmal als "Superversion" bezeichnet.) Bei einem Konflikt zwischen den Versionen haben die aktuelleren Versionen normalerweise Vorrang gegenüber den älteren Versionen. Wenn sich beispielsweise eine Kategoriebeschriftung in einer Version abweicht, wird der Text in der aktuellsten Version verwendet.
- **Neueste Version.** Wählen Sie diese Option, wenn Sie die aktuellste Version verwenden möchten.
- **Version angeben.** Wählen Sie diese Option, wenn Sie eine bestimmte Umfrageversion verwenden möchten.

Die Auswahl aller Versionen ist sinnvoll, wenn Sie beispielsweise Falldaten für mehrere Versionen exportieren möchten und Änderungen an den Variablen- und Kategoriedefinitionen durchgeführt wurden, die dazu führen, dass Falldaten, die mit einer bestimmten Version gesammelt wurden, in einer anderen Version nicht gültig sind. Die Auswahl aller Versionen, für die die Falldaten exportiert werden sollen, bedeutet, dass Sie im Allgemeinen die mit den verschiedenen Versionen gesammelten Falldaten gleichzeitig exportieren können, ohne dass Gültigkeitsfehler aufgrund der Unterschiede zwischen den Versionen gemeldet werden. Dennoch können, je nach den Versionsänderungen, dennoch einige Gültigkeitsfehler auftreten.

Sprache. Die Fragen und der zugehörige Text können in den Metadaten in mehreren Sprachen gespeichert werden. Sie können die Standardsprache für die Umfrage verwenden oder eine bestimmte Sprache angeben. Wenn ein Element in der angegebenen Sprache nicht verwendet wird, wird der Standard verwendet.

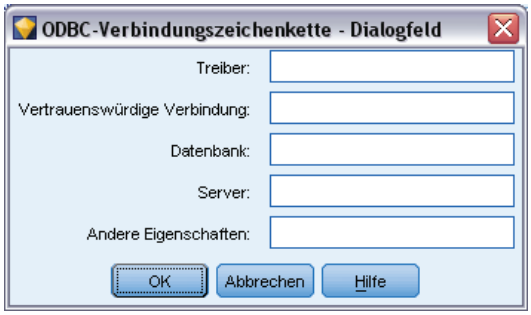
Kontext. Wählen Sie den zu verwendenden Benutzerkontext aus. Der Benutzerkontext regelt, welche Texte angezeigt werden. Wählen Sie beispielsweise Frage, um Fragetexte anzuzeigen, oder Analyse, um kürzere Texte anzuzeigen, die bei der Analyse der Daten für die Anzeige geeignet sind.

Beschriftungstyp. Listet die definierten Beschriftungstypen auf. Der Standard ist Beschriftung. Er wird für Fragetexte im Benutzerkontext "Frage" und für Variablenbeschreibungen im Benutzerkontext "Analyse" verwendet. Für Anweisungen, Beschreibungen usw. können weitere Beschriftungstypen definiert werden.

Datenbankverbindungszeichenkette

Bei Verwendung des IBM® SPSS® Data Collection-Knotens zum Import von Falldaten aus einer Datenbank via OLE-DB oder ODBC wählen Sie Bearbeiten aus der Registerkarte "Datei" aus, um auf das Dialogfeld "Verbindungszeichenkette" zuzugreifen, in dem Sie die Verbindungszeichenkette anpassen können, die zur Feinabstimmung der Verbindung an den Provider weitergeleitet wird.

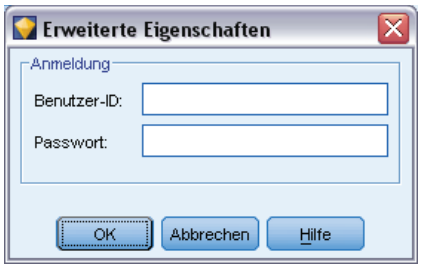
Abbildung 2-16
IBM SPSS Data Collection-Import – Verbindungszeichenkette



Erweiterte Eigenschaften

Bei Verwendung des IBM® SPSS® Data Collection-Knotens zum Importieren von Falldaten aus einer Datenbank, für die eine explizite Anmeldung erforderlich ist, wählen Sie Erweitert aus, um eine Benutzer-ID und ein Passwort für den Zugriff auf die Datenquelle anzugeben.

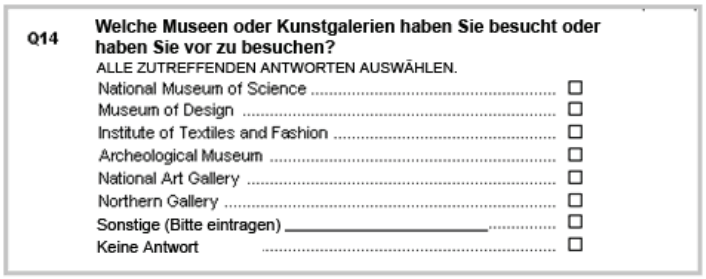
Abbildung 2-17
IBM SPSS Data Collection-Import – Erweiterte Eigenschaften



Importieren von Mehrfachantworten-Sets

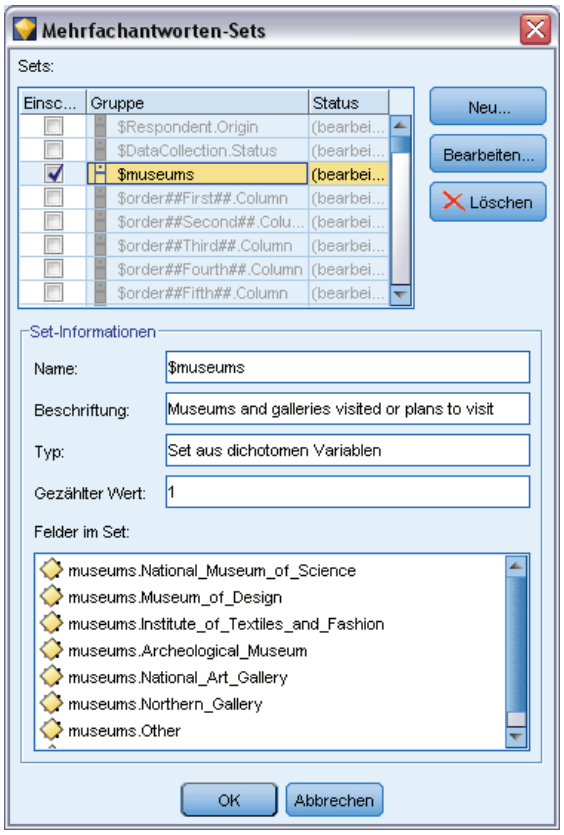
Mehrfachantwortvariablen können aus IBM® SPSS® Data Collection als Sets aus dichotomen Variablen mit einem gesonderten Flag-Feld für jeden möglichen Wert der Variablen importiert werden. Wenn die Befragten beispielsweise in einer Liste auswählen sollen, welche Museen sie besucht haben, enthält das Set ein gesondertes Flag-Feld für jedes aufgeführte Museum.

Abbildung 2-18
Frage mit Mehrfachantworten



Nach dem Import der Daten können Sie Mehrfachantworten-Sets über jeden Knoten, der die Registerkarte “Filter” enthält, hinzufügen und bearbeiten. Für weitere Informationen siehe Thema Bearbeiten von Mehrfachantworten-Sets in Kapitel 4 auf S. 160.

Abbildung 2-19
Dialogfeld “Mehrfachantworten-Sets”



Importieren von Mehrfachantworten in ein einzelnes Feld (für in früheren Versionen erstellte Streams)

In älteren Versionen von IBM® SPSS® Modeler wurden Mehrfachantworten nicht wie oben beschrieben importiert, sondern in ein einzelnes Feld. Die Werte wurden dabei durch Kommas getrennt. Diese Methode wird weiterhin unterstützt, um bestehende Streams zu unterstützen, es wird jedoch empfohlen, alle derartigen Streams für die Verwendung der neuen Methode zu aktualisieren.

Anmerkungen zum Import von IBM SPSS Data Collection-Spalten

Spalten aus den IBM® SPSS® Data Collection-Daten werden wie in der folgenden Tabelle zusammengefasst in IBM® SPSS® Modeler eingelesen.

Data Collection Spaltentyp	SPSS Modeler-Speicher	Messniveau
Boole’sche Flag (ja/nein)	Zeichenfolge	Flag (Werte 0 und 1)
Kategorial	Zeichenfolge	Nominal

Data Collection Spaltentyp	SPSS Modeler-Speicher	Messniveau
Datums- oder Zeitstempel	Zeitstempel	Stetig
Doppelt (Gleitkommawert innerhalb eines angegebenen Bereichs)	Reelle Zahl	Stetig
Lang (ganzzahliger Wert innerhalb eines angegebenen Bereichs)	Ganzzahl	Stetig
Text (Freitextbeschreibung)	Zeichenfolge	Typlos
Ebene (gibt Raster oder Schleifen innerhalb einer Frage an)	Kommt in VDATA nicht vor und wird nicht in SPSS Modeler importiert	
Objekt (Binärdaten wie beispielsweise ein Fax mit handgeschriebenem Text oder eine Tonaufnahme)	Wird nicht in SPSS Modeler importiert	
Keine (unbekannter Typ)	Wird nicht in SPSS Modeler importiert	
Respondent.Serial-Spalte (weist jedem Befragten eine eindeutige ID zu)	Ganzzahl	Typlos

Um mögliche Inkonsistenzen zwischen Wertbeschriftungen aus Metadaten und tatsächlichen Werten zu vermeiden, werden alle Metadatenwerte in Kleinbuchstaben umgewandelt. So wird beispielsweise der Wert der Beschriftung *E1720_Jahre* in *e1720_jahre* konvertiert.

IBM Cognos BI-Quellenknoten

Mit dem IBM Cognos IB-Quellenknoten können Sie Cognos BI-Datenbankdaten oder einzelne Listenberichte in Ihre Data Mining-Sitzung importieren. Auf diese Weise können Sie die Business Intelligence-Funktionen von Cognos mit den Vorhersageanalytikfunktionen von IBM® SPSS® Modeler kombinieren. Sie können relationale, dimensional modellierte relationale (DMR) und OLAP-Daten importieren.

Wählen Sie über eine Cognos-Serververbindung zunächst einen Speicherort aus, aus dem Daten bzw. Berichte importiert werden sollen. Ein Speicherort enthält ein Cognos-Modell und alle Ordner, Abfragen, Ansichten, Verknüpfungen, URLs und Aufgabendefinitionen, die diesem Modell zuordnet sind. Ein Cognos-Modell definiert Unternehmensregeln, Datenbeschreibungen, Datenbeziehungen, Geschäftsdimensionen und -Hierarchien sowie andere administrative Aufgaben.

Wenn Sie Daten importieren, wählen Sie die zu importierenden Objekte aus dem ausgewählten Paket aus. Zu den importierbaren Objekten gehören Abfragesubjekte (die für Datenbanktabellen stehen) oder einzelne Abfrageelemente (die für Datenbankspalten stehen). Für weitere Informationen siehe Thema Cognos-Objektsymbole auf S. 44.

Wenn für das Paket Filter definiert wurden, können Sie einen oder mehrere davon importieren. Wenn ein von Ihnen importierter Filter importierten Daten zugeordnet ist, wird der betreffende Filter angewendet, bevor die Daten importiert werden. *Hinweis*: Die zu importierenden Daten müssen im UTF-8-Format vorliegen.

Wenn Sie einen Bericht importieren, wählen Sie ein Paket bzw. einen Ordner in einem Paket mit einem oder mehreren Berichten aus. Anschließend wählen Sie den Bericht aus, der importiert werden soll. *Hinweis*: Es können nur einzelne Listenberichte importiert werden; mehrere Listen werden nicht unterstützt.

Wenn Parameter definiert wurden, entweder für ein Datenobjekt oder für einen Bericht, können Sie Werte für diese Parameter angeben, bevor Sie das Objekt bzw. den Bericht importieren.

Cognos-Objektsymbole

Die verschiedenen Objekttypen, die Sie aus einer Cognos BI-Datenbank importieren können, werden durch unterschiedliche Symbole dargestellt, wie in der folgenden Tabelle zu sehen.

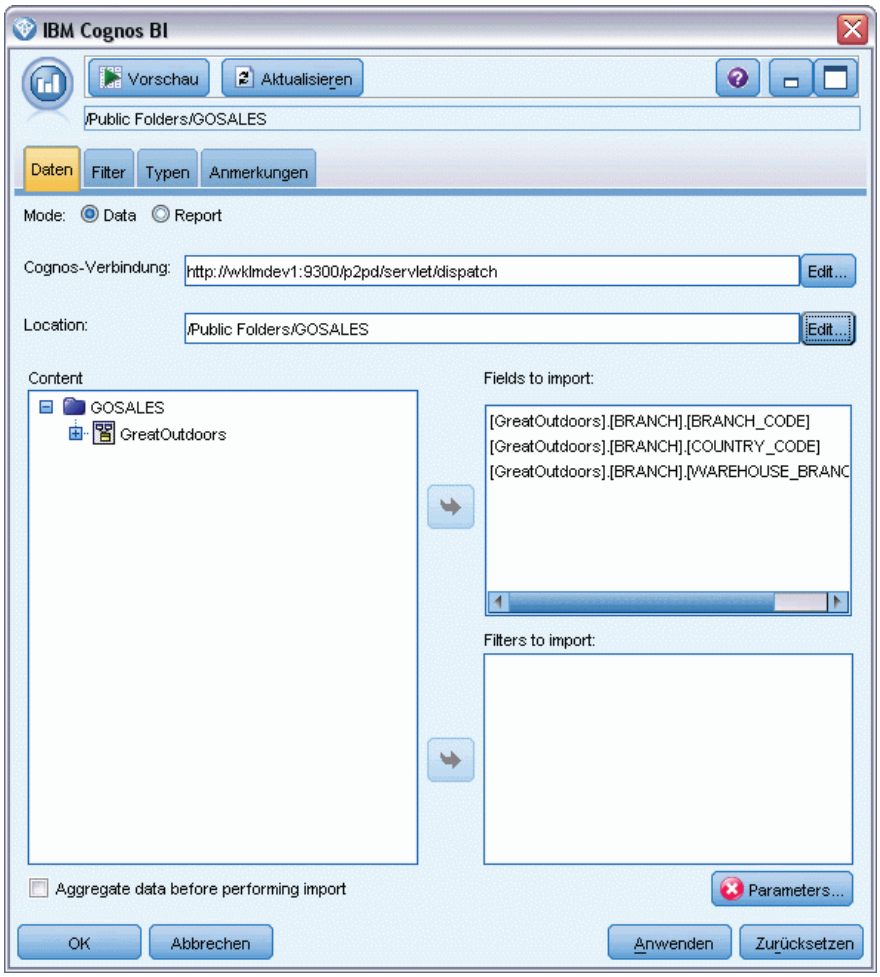
Tabelle 2-3
Cognos-Objektsymbole

Symbol	Objekt
	Paket
	Namespace
	Abfragesubjekt
	Abfrageelement
	Maßdimension
	Maß
	Dimension
	Ebenenhierarchie
	Ebene
	Filter
	Bericht
	Eigenständige Berechnung

Importieren von Cognos-Daten

Beim Importieren von Daten aus einer IBM Cognos BI-Datenbank müssen Sie sicherstellen, dass Modus auf Daten gesetzt ist, und das Dialogfeld wie folgt ausfüllen.

Abbildung 2-20
Importieren von Cognos-Daten



Verbindung. Klicken Sie auf die Schaltfläche Bearbeiten, um ein Dialogfeld anzuzeigen, in dem Sie die Details einer neuen Cognos-Verbindung, über die Daten bzw. Berichte importiert werden sollen, definieren können. Wenn Sie bereits bei einem Cognos-Server über IBM® SPSS® Modeler angemeldet sind, können Sie auch die Dateils der aktuellen Verbindung bearbeiten. Für weitere Informationen siehe Thema Cognos-Verbindungen auf S. 48.

Speicherort. Wenn Sie die Cognos-Serververbindung eingerichtet haben, klicken Sie auf die Schaltfläche Bearbeiten neben diesem Feld, um eine Liste der verfügbaren Pakete anzuzeigen, aus denen Sie Inhalte importieren können. Für weitere Informationen siehe Thema Auswahl des Cognos-Standorts auf S. 49.

Inhalt. Zeigt den Namen des ausgewählten Pakets und die dem Paket zugewiesenen Namespaces an. Doppelklicken Sie auf einen Namespace, um die Objekte anzuzeigen, die Sie importieren können. Die verschiedenen Objekttypen sind durch unterschiedliche Symbole gekennzeichnet. Für weitere Informationen siehe Thema Cognos-Objektsymbole auf S. 44.

Um ein Objekt für den Import auszuwählen, markieren Sie das Objekt und klicken Sie auf den oberen der beiden Rechtspfeile, um das Objekt in den Bereich Zu importierende Felder zu verschieben. Wenn Sie auf ein Abfragesubjekt klicken, werden alle entsprechenden Abfrageelemente importiert. Wenn Sie auf ein Abfragesubjekt doppelklicken, wird es erweitert, sodass Sie ein oder mehrere seiner individuellen Abfrageelemente auswählen können. Mit Strg-Klicken (individuelle Elemente auswählen), Umschalt-Klicken (mehrere Elemente auswählen) und Strg-A (alle Elemente auswählen) können Sie eine Mehrfachauswahl vornehmen.

Um einen anzuwendenden Filter auszuwählen (sofern für das Paket Filter definiert sind), navigieren Sie im Bereich "Inhalt" zu dem Filter, markieren Sie ihn und klicken Sie auf den unteren der beiden Rechtspfeile, um den Filter in den Bereich Anzuwendende Filter zu verschieben. Mit Strg-Klicken (einzelne Filter auswählen) und Umschalt-Klicken (zusammenhängenden Block von Filtern auswählen) können Sie eine Mehrfachauswahl vornehmen.

Zu importierende Felder. Listet die Datenbankobjekte auf, die laut Ihrer Auswahl zur Verarbeitung in SPSS Modeler importiert werden. Wenn Sie ein bestimmtes Objekt nicht mehr benötigen, wählen Sie es aus und klicken Sie auf den Linkspfeil, um es wieder in den Bereich Inhalt zu verschieben. Sie können Mehrfachauswahlen auf dieselbe Weise wie für Inhalt vornehmen.

Anzuwendende Filter. Listet die Filter auf, die laut Ihrer Auswahl vor dem Import auf die Daten angewendet werden sollen. Wenn Sie einen bestimmten Filter nicht mehr benötigen, wählen Sie ihn aus und klicken Sie auf den Linkspfeil, um ihn wieder in den Bereich Inhalt zu verschieben. Sie können Mehrfachauswahlen auf dieselbe Weise wie für Inhalt vornehmen.

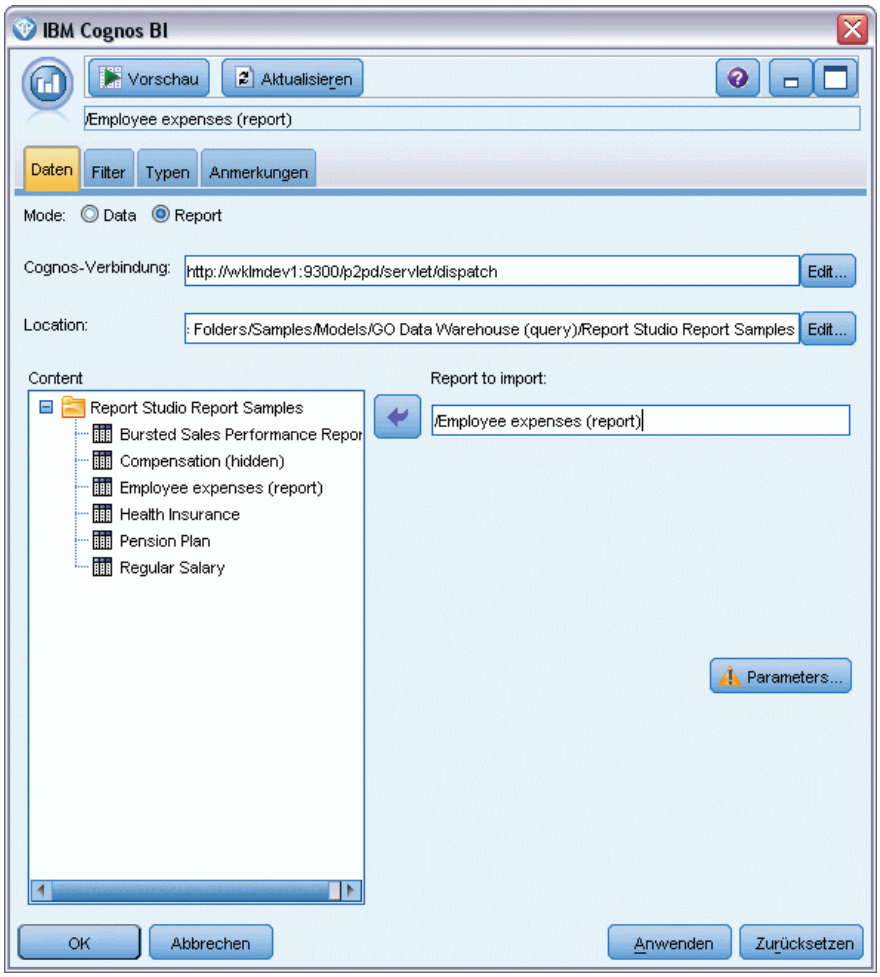
Parameter. Wenn diese Schaltfläche aktiviert ist, sind für das ausgewählte Objekt Parameter definiert. Mit Parametern können Sie Anpassungen vornehmen (beispielsweise eine parametrisierte Berechnung durchführen), bevor Sie die Daten importieren. Wenn Parameter definiert sind, aber kein Standard angegeben wurde, wird auf der Schaltfläche ein Warndreieck angezeigt. Klicken Sie auf die Schaltfläche, um die Parameter anzuzeigen und ggf. zu bearbeiten. Wenn die Schaltfläche deaktiviert ist, sind für den Bericht keine Parameter definiert.

Daten vor dem Importieren aggregieren. Aktivieren Sie dieses Kontrollkästchen, wenn Sie statt Rohdaten aggregierte Daten importieren möchten.

Importieren von Cognos-Berichten

Beim Importieren eines vordefinierten Berichts aus einer IBM Cognos BI-Datenbank müssen Sie sicherstellen, dass Modus auf Bericht gesetzt ist, und das Dialogfeld wie folgt ausfüllen. *Hinweis*: Es können nur einzelne Listenberichte importiert werden; mehrere Listen werden nicht unterstützt.

Abbildung 2-21
Importieren von Cognos-Berichten



Verbindung. Klicken Sie auf die Schaltfläche Bearbeiten, um ein Dialogfeld anzuzeigen, in dem Sie die Details einer neuen Cognos-Verbindung, über die Daten bzw. Berichte importiert werden sollen, definieren können. Wenn Sie bereits bei einem Cognos-Server über IBM® SPSS® Modeler angemeldet sind, können Sie auch die Dateils der aktuellen Verbindung bearbeiten. Für weitere Informationen siehe Thema Cognos-Verbindungen auf S. 48.

Speicherort. Wenn Sie die Cognos-Serververbindung eingerichtet haben, klicken Sie auf die Schaltfläche Bearbeiten neben diesem Feld, um eine Liste der verfügbaren Pakete anzuzeigen, aus denen Sie Inhalte importieren können. Für weitere Informationen siehe Thema Auswahl des Cognos-Standorts auf S. 49.

Inhalt. Zeigt den Namen des ausgewählten Pakets bzw. des ausgewählten Ordners an, der Berichte enthält. Navigieren Sie zu einem Bericht, wählen Sie ihn aus und klicken Sie auf den Rechtspfeil, um den Bericht in das Feld Zu importierender Bericht zu verschieben.

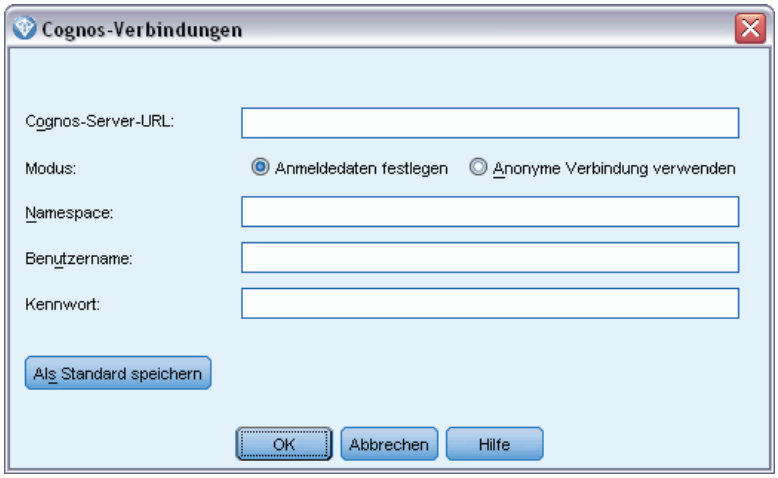
Zu importierender Bericht. Gibt den Bericht an, der laut Ihrer Auswahl in SPSS Modeler importiert wird. Wenn Sie den Bericht nicht mehr benötigen, wählen Sie ihn aus und klicken Sie auf den Linkspfeil, um ihn wieder in den Bereich Inhalt zu verschieben, oder verschieben Sie einen anderen Bericht in dieses Feld.

Parameter. Wenn diese Schaltfläche aktiviert ist, sind für den ausgewählten Bericht Parameter definiert. Sie können Parameter verwenden, um vor dem Import des Berichts Anpassungen vorzunehmen (z. B. Angabe eines Start- und Enddatums für Berichtdaten). Wenn Parameter definiert sind, aber kein Standard angegeben wurde, wird auf der Schaltfläche ein Warndreieck angezeigt. Klicken Sie auf die Schaltfläche, um die Parameter anzuzeigen und ggf. zu bearbeiten. Wenn die Schaltfläche deaktiviert ist, sind für den Bericht keine Parameter definiert.

Cognos-Verbindungen

Im Dialogfeld "Cognos-Verbindungen" können Sie den Cognos BI-Server auswählen, von dem Sie Datenbankobjekte importieren bzw. an den Sie Datenbankobjekte exportieren möchten.

Abbildung 2-22
Auswahl des Cognos-Servers



URL des Cognos-Servers. Geben Sie die URL des Cognos BI-Servers ein, den Sie für die Import- bzw. Exportvorgänge verwenden möchten. Dies ist der Wert der Umgebungseigenschaft "External dispatcher URI" (Externe Dispatcher-URI) der IBM Cognos-Konfiguration auf dem Cognos BI-Server. Wenden Sie sich an Ihren Cognos-Systemadministrator, wenn Sie sich nicht sicher sind, welche URL Sie verwenden müssen.

Modus. Wählen Sie Anmeldedaten festlegen, wenn Sie sich mit einem spezifischen Cognos-Namespace, Benutzernamen und Passwort anmelden möchten (z. B. als Administrator). Wählen Sie Anonyme Verbindung verwenden, um sich ohne Benutzer-Anmeldedaten anzumelden. Sie füllen in diesem Fall keine weiteren Felder aus.

Namespace. Geben Sie den Sicherheitsanbieter für die Authentifizierung bei Cognos an, mit dem Sie sich beim Server anmelden möchten. Der Authentifizierungsanbieter dient dazu, Benutzer, Gruppen und Rollen zu definieren und zu verwalten und den Authentifizierungsprozess zu steuern.

Benutzername. Geben Sie den Cognos-Benutzernamen ein, mit dem die Anmeldung beim Server erfolgen soll.

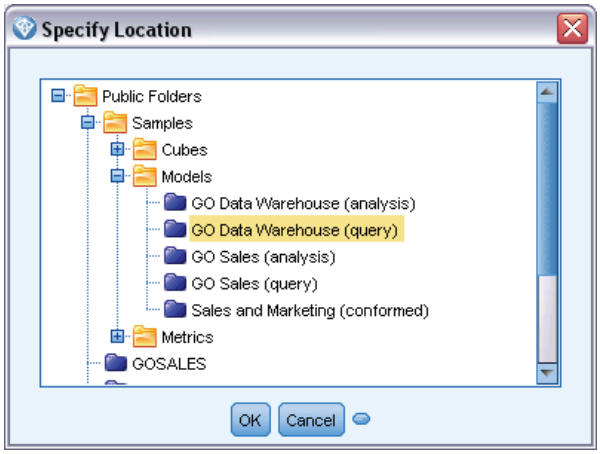
Passwort. Geben Sie das Passwort ein, das zum angegebenen Benutzernamen gehört.

Als Standard speichern. Klicken Sie auf diese Schaltfläche, um diese Einstellung als Standardeinstellungen wiederherzustellen, damit Sie sie nicht jedesmal, wenn Sie den Knoten öffnen, neu eingeben müssen.

Auswahl des Cognos-Standorts

Im Dialogfeld “Speicherort angeben” können Sie ein Cognos-Paket angeben, aus dem Daten importiert werden sollen, bzw. ein Paket bzw. einen Ordner, aus dem Berichte importiert werden sollen.

Abbildung 2-23
Auswahl des Cognos-Standorts



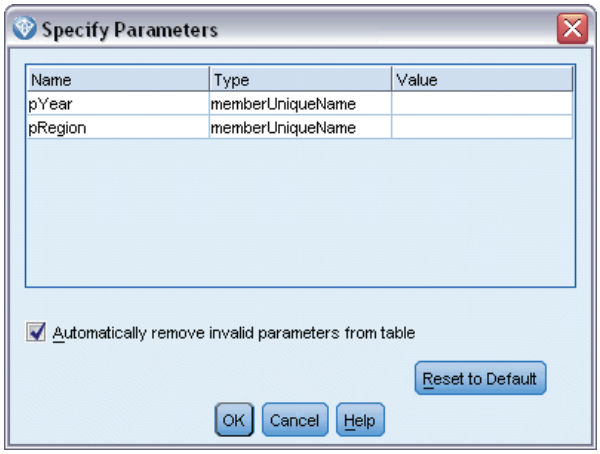
Öffentliche Ordner. Wenn Sie Daten importieren, werden hier die Pakete und Ordner aufgelistet, die auf dem ausgewählten Server zur Verfügung stehen. Wählen Sie das Paket aus, das Sie verwenden möchten, und klicken Sie auf OK. Sie können nur ein Paket pro Cognos BI-Quellenknoten auswählen.

Wenn Sie Berichte importieren, werden hier die Ordner und Pakete mit Berichten aufgelistet, auf denen ausgewählte Server zur Verfügung stehen. Wählen Sie ein Paket oder einen Berichtordner aus und klicken Sie auf OK. Sie können nur ein einziges Paket bzw. nur einen einzigen Berichtordner pro Cognos BI-Quellenknoten auswählen, die Berichtordner können jedoch andere Berichtordner sowie einzelne Berichte enthalten.

Angeben von Parametern für Daten bzw. Berichte

Wenn Parameter in Cognos BI definiert wurden, entweder für ein Datenobjekt oder für einen Bericht, können Sie Werte für diese Parameter angeben, bevor Sie das Objekt bzw. den Bericht importieren. Ein Beispiel für Parameter für einen Bericht wären die Anfangs- und Enddaten für die Berichtsinhalte.

Abbildung 2-24
Cognos-Parameter



Name. Der Name des Parameters laut Angabe in der Cognos BI-Datenbank.

Typ. Eine Beschreibung des Parameters.

Wert. Der dem Parameter zuzuweisende Wert. Doppelklicken Sie zur Eingabe bzw. Bearbeitung eines Werts auf die entsprechende Zelle in der Tabelle. Hier werden keine Werte validiert, etwaige ungültige Werte werden somit zur Laufzeit entdeckt.

Ungültige Parameter automatisch aus Tabelle entfernen. Diese Option ist standardmäßig ausgewählt und entfernt alle ungültigen Parameter, die im Datenobjekt bzw. Bericht gefunden werden.

SAS-Quellenknoten

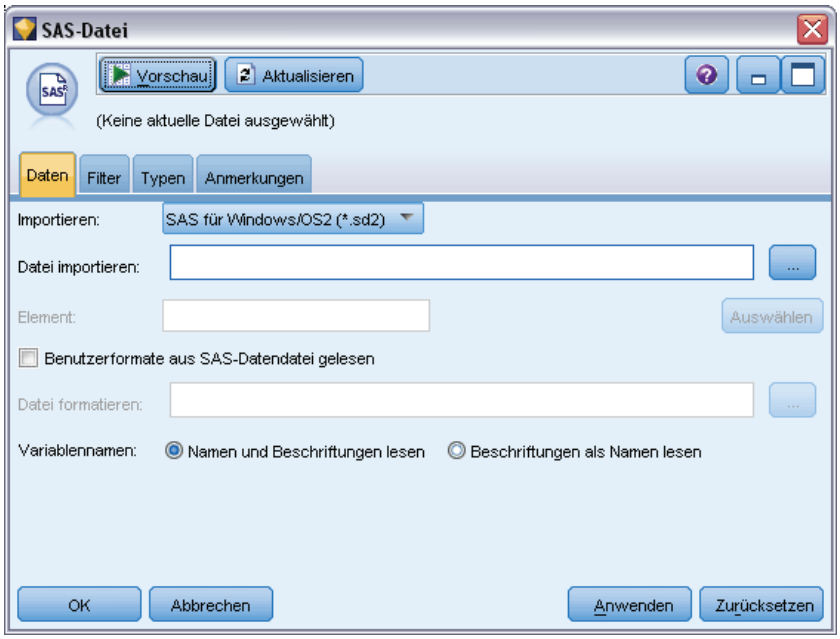
Hinweis: Diese Funktion steht in SPSS Modeler Professional und SPSS Modeler Premium zur Verfügung.

Mit dem SAS-Quellenknoten können Sie SAS-Daten in Ihre Data Mining-Sitzung importieren. Sie können vier Dateitypen importieren:

- SAS für Windows/OS2 (*.sd2*)
- SAS für UNIX (*.ssd*)
- SAS-Transportdatei (*.tpt*)
- SAS Version 7/8/9 (*.sas7bdat*)

Beim Importieren der Daten werden alle Variablen beibehalten und kein Variablentyp wird geändert. Alle Fälle werden ausgewählt.

Abbildung 2-25
Importieren einer SAS-Datei



Festlegen von Optionen für den SAS-Quellenknoten

Importieren. Wählen Sie den zu transportierenden SAS-Dateityp aus. Zur Auswahl stehen die Optionen SAS für Windows/OS2 (.sd2), SAS für UNIX (.SSD), SAS-Transportdatei (.tpt) oder SAS Version 7/8/9 (.sas7bdat).

Datei importieren. Geben Sie den Namen der Datei an. Sie können einen Dateinamen eingeben oder auf die Schaltfläche mit den Auslassungspunkten (...) klicken, um zum Speicherort der Datei zu blättern.

Element. Wählen Sie ein Element für den Import aus der oben ausgewählten SAS-Transportdatei aus. Sie können einen Elementnamen eingeben oder auf Auswählen klicken, um durch alle Elemente in der Datei zu blättern.

Benutzerformate aus SAS-Datendatei lesen. Wählen Sie diese Option, um Benutzerformate zu lesen. SAS-Dateien speichern Daten und Datenformate (wie Variablenlabels) in verschiedenen Dateien. In den meisten Fällen sollen die Formate ebenfalls importiert werden. Bei einem großen Daten-Set ist es jedoch empfehlenswert, diese Option zu deaktivieren, um Speicher zu sparen.

Formatdatei. Wenn eine Formatdatei erforderlich ist, ist dieses Textfeld aktiviert. Sie können einen Dateinamen eingeben oder auf die Schaltfläche mit den Auslassungspunkten (...) klicken, um zum Speicherort der Datei zu blättern.

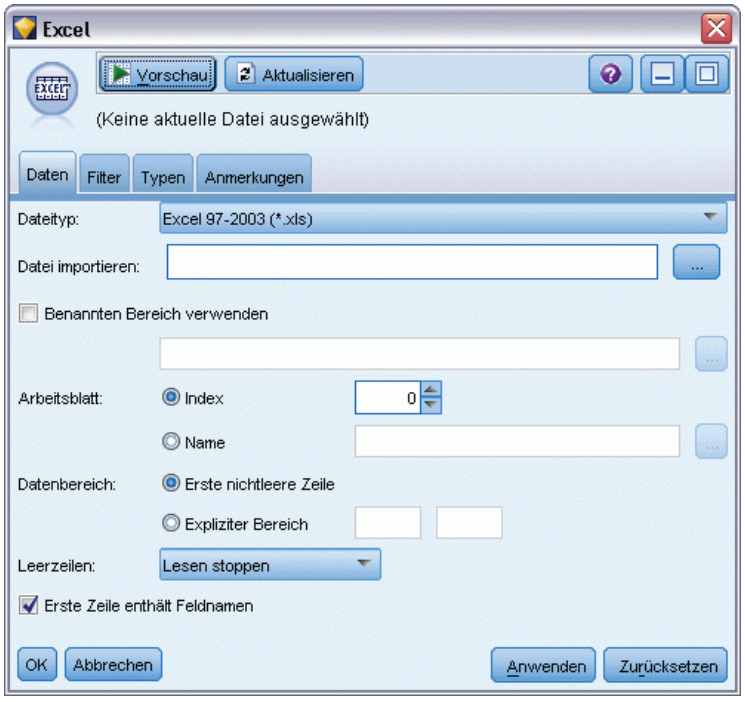
Variablennamen. Wählen Sie eine Methode zur Behandlung von Variablennamen und -beschriftungen beim Importieren aus einer SAS-Datei. Metadaten, die Sie hier einschließen, bleiben während Ihrer Arbeit in IBM® SPSS® Modeler erhalten und können zur Verwendung in SAS wieder exportiert werden.

- **Namen und Beschriftungen lesen.** Wählen Sie diese Option, wenn sowohl Variablennamen als auch -beschriftungen in SPSS Modeler eingelesen werden sollen. Standardmäßig ist diese Option ausgewählt und Variablennamen werden im Typknoten angezeigt. Beschriftungen können je nach den im Dialogfeld "Stream-Eigenschaften" angegebenen Optionen in Expression Builder, Diagrammen, Modellbrowsern und anderen Ausgabearten angezeigt werden.
- **Beschriftungen als Namen lesen.** Wählen Sie diese Option, um statt der kurzen Feldnamen die beschreibenden Variablenlabels aus der SAS-Datei zu lesen und diese Beschriftungen als Variablennamen in SPSS Modeler zu verwenden.

Excel-Quellenknoten

Mit dem Excel-Quellenknoten können Sie Daten aus einer beliebigen Version von Microsoft Excel importieren.

Abbildung 2-26
Excel-Quellenknoten



Dateityp. Wählen Sie den Excel-Dateityp, den Sie importieren möchten.

Datei importieren. Gibt Namen und Speicherort der zu importierenden Tabellenkalkulationsdatei an.

Benannten Bereich verwenden. Ermöglicht die Angabe eines benannten Zellenbereichs, wie im Excel-Arbeitsblatt definiert. Klicken Sie auf die Schaltfläche mit den Auslassungspunkten (...), um eine Auswahl aus der Liste der verfügbaren Bereiche zu treffen. Wenn ein benannter Bereich verwendet wird, sind andere Einstellungen für Arbeitsblatt und Datenbereich nicht mehr anwendbar und werden daher deaktiviert.

Arbeitsblatt wählen. Gibt das zu importierende Arbeitsblatt an, entweder nach Index oder nach Namen.

- **Nach Index.** Geben Sie den Indexwert für das zu importierende Arbeitsblatt an. Beginnen Sie mit 0 für das erste Arbeitsblatt, 1 für das zweite Arbeitsblatt usw.
- **Nach Namen.** Geben Sie den Namen des importierenden Arbeitsblattes an. Klicken Sie auf die Schaltfläche mit den Auslassungspunkten (...), um eine Auswahl aus der Liste der verfügbaren Arbeitsblätter zu treffen.

Bereich auf Arbeitsblatt. Sie können Daten beginnend mit der ersten nichtleeren Zeile oder mit einem expliziten Zellenbereich importieren.

- **Bereich beginnt mit der ersten nichtleeren Zeile.** Sucht die erste nichtleere Zelle und verwendet diese als linke obere Ecke des Datenbereichs.
- **Expliziter Zellbereich.** Ermöglicht die Angabe eines expliziten Bereichs nach Zeile und Spalte. Beispielsweise können Sie für den Excel-Bereich A1:D5 in das erste Feld A1 und in das zweite Feld D5 eingeben (oder alternativ R1C1 und R5C4). Alle Zeilen im angegebenen Bereich werden ausgegeben, einschließlich der Leerzeilen.

Bei Leerzeilen. Wenn mehrere leere Zeilen gefunden werden, können Sie mit Lesen stoppen angeben, dass der Lesevorgang angehalten werden soll, oder mit Leere Zeilen zurückgeben festlegen, dass alle Daten bis zum Ende des Arbeitsblatts gelesen werden sollen, einschließlich Leerzeilen.

Erste Zeile enthält Spaltennamen. Gibt an, dass die erste Zeile im angegebenen Bereich als Feldnamen (Spaltennamen) verwendet werden soll. Wenn diese Option nicht ausgewählt ist, werden Feldnamen automatisch generiert.

Feld-Speichertyp und -Messniveau

Beim Lesen von Werten aus Excel werden Felder mit numerischem Speicher standardmäßig mit einem Messniveau von *Stetig* und Zeichenkettenfelder als *Nominal* eingelesen. Sie können auf der Registerkarte “Typ” manuell das Messniveau (“Stetig” bzw. “Nominal”) ändern, der Speichertyp wird jedoch automatisch bestimmt (allerdings kann er, falls erforderlich, mithilfe einer Konvertierungsfunktion, wie beispielsweise to_integer, in einem Füller- oder Ableitungsknoten geändert werden. Für weitere Informationen siehe Thema Festlegen von Feldspeichertyp und Formatierung auf S. 32.

Standardmäßig werden Felder mit einer Mischung aus numerischen Werten und Zeichenkettenwerten als Zahlen eingelesen. Alle Zeichenkettenwerte werden also in IBM® SPSS® Modeler auf null (systemdefiniert fehlend) gesetzt. Dies liegt daran, dass SPSS Modeler im Gegensatz zu Excel keine gemischten Speichertypen innerhalb eines Felds zulässt. Um dies zu vermeiden, können Sie das Zellenformat in der Excel-Tabelle manuell auf Text setzen. Dadurch werden alle Werte (einschließlich Zahlen) als Zeichenketten eingelesen.

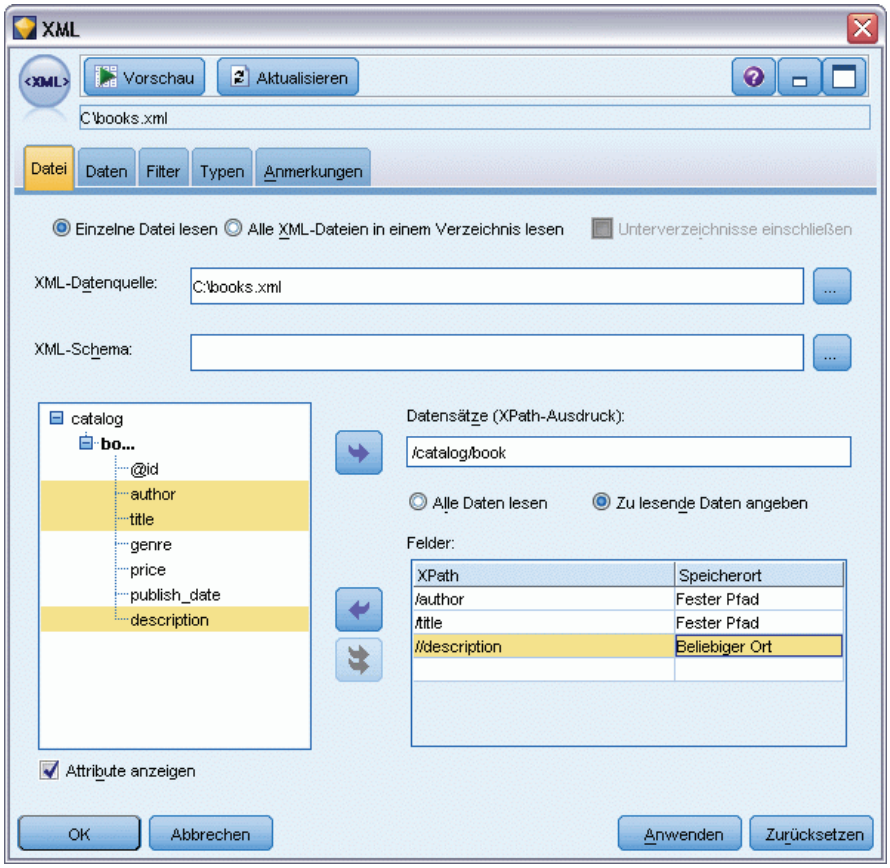
XML-Quellenknoten

Hinweis: Diese Funktion steht in SPSS Modeler Professional und SPSS Modeler Premium zur Verfügung.

Mit dem XML-Quellenknoten können Sie die Daten aus einer Datei im XML-Format in einen IBM® SPSS® Modeler-Stream importieren. XML ist eine Standardsprache für den Datenaustausch und gilt für viele Unternehmen als das bevorzugte Format für diesen Zweck. So möchte beispielsweise eine Steuerbehörde Daten aus Steuereinnahmen analysieren, die online übermittelt wurden und im XML-Format stehen.

Durch Importieren von XML-Daten in einen SPSS Modeler-Stream können Sie zahlreiche Vorhersageanalysefunktionen an der Quelle ausführen. Die XML-Daten werden in ein Tabellenformat gegliedert, bei dem die Spalten den verschiedenen Verschachtelungsniveaus der XML-Elemente und Attribute entsprechen. Die XML-Objekte werden im XPath-Format angezeigt (siehe *http://www.w3.org/TR/xpath20/*).

Abbildung 2-27
XML-Daten importieren



Eine einzelne Datei lesen. Standardmäßig liest SPSS Modeler eine einzelne Datei, die Sie im Feld XML-Datenquelle angeben.

Alle XML-Dateien in einem Verzeichnis lesen. Wenn Sie diese Option wählen, werden alle XML-Dateien in einem bestimmten Verzeichnis gelesen. Geben Sie die Position in dem Feld Verzeichnis an, das angezeigt wird. Aktivieren Sie das Kontrollkästchen Unterverzeichnisse einschließen, um zusätzlich XML-Dateien aus allen Unterverzeichnissen des angegebenen Verzeichnisses zu lesen.

XML-Datenquelle. Geben Sie den vollständigen Pfad und Dateinamen der XML-Quellendatei an, die Sie importieren möchten, oder nutzen Sie die Schaltfläche “Durchsuchen”, um die Datei zu finden.

XML-Schema. (Optional) Geben Sie den vollständigen Pfad und Dateinamen einer XSD- oder DTD-Datei an, aus der die XML-Struktur gelesen werden soll, oder verwenden Sie die Schaltfläche “Durchsuchen”, um diese Datei zu finden. Wenn Sie dieses Feld frei lassen, wird die Struktur aus der XML-Quellendatei gelesen. Eine XSD- oder DTD-Datei kann mehr als ein Wurzelelement besitzen. In diesem Fall wird ein Dialogfeld angezeigt, in dem Sie das gewünschte Wurzelelement auswählen, wenn Sie den Fokus auf ein anderes Feld wechseln. Für weitere Informationen siehe Thema Auswahl aus mehreren Wurzelelementen auf S. 56.

XML-Struktur. Ein hierarchischer Baum, der die Struktur der XML-Quellendatei anzeigt (oder das Schema, sofern Sie eines im Feld XML-Schema angegeben haben). Zum Definieren einer Datensatzgrenze wählen Sie ein Element aus und klicken auf die Rechtspfeil-Schaltfläche, um das Objekt in das Feld Datensätze zu kopieren.

Attribute anzeigen. Zeigt die Attribute der XML-Elemente in dem Feld XML-Struktur an oder blendet sie aus.

Datensätze (XPath-Ausdruck). Zeigt die XPath-Syntax für ein Element, das aus dem Feld “XML-Struktur” kopiert wurde. Dieses Element wird dann in der XML-Struktur hervorgehoben und definiert die Datensatzgrenze. Jedes Mal, wenn dieses Element in der Quellendatei gefunden wird, wird ein neuer Datensatz erstellt. Wenn das Feld leer ist, wird das erste untergeordnete Element unter dem Stamm als Datensatzgrenze verwendet.

Alle Daten lesen. Standardmäßig werden alle Daten in der Quellendatei in den Stream eingelesen.

Zu lesende Daten angeben. Wählen Sie diese Option, wenn Sie einzelne Elemente, Attribute oder beides importieren möchten. Durch Auswählen dieser Option wird die Feldertabelle aktiviert, in der Sie die zu importierenden Daten angeben können.

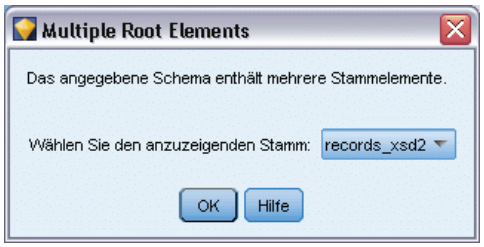
Felder. In dieser Tabelle werden die für den Import ausgewählten Elemente und Attribute angezeigt, wenn Sie die Option Zu lesende Daten angeben ausgewählt haben. Sie können die XPath-Syntax eines Elements oder Attributs entweder direkt in die XPath-Spalte eingeben oder ein Element oder Attribut in der XML-Struktur auswählen und auf die Rechtspfeil-Schaltfläche klicken, um das Objekt in die Tabelle zu kopieren. Zum Kopieren aller untergeordneten Elemente und Attribute eines Elements wählen Sie das Element in der XML-Struktur aus und klicken Sie auf die Doppelpfeilschaltfläche.

- **XPath.** Die XPath-Syntax der zu importierenden Objekte.
- **Ort.** Die Position in der XML-Struktur der zu importierenden Objekte. Fester Weg zeigt den Weg des Objekts im Verhältnis zu dem in der XML-Struktur hervorgehobenen Element (oder dem ersten untergeordneten Element unter dem Stamm, wenn kein Element hervorgehoben ist). Beliebiger Ort kennzeichnet ein Objekt mit dem angegebenen Namen an einem beliebigen Ort in der XML-Struktur. Benutzerdefiniert wird angezeigt, wenn Sie den Ort direkt in die XPath-Spalte eingeben.

Auswahl aus mehreren Wurzelelementen

Während eine ordnungsgemäß erstellte XML-Datei nur ein einzelnes Wurzelelement besitzen kann, kann eine XSD- oder DTD-Datei mehrere Wurzelelemente enthalten. Wenn eine der Wurzeln mit der Wurzel in der XML-Quellendatei übereinstimmt, wird dieses Wurzelelement verwendet. Andernfalls müssen Sie das zu verwendende Wurzelelement auswählen.

Abbildung 2-28
Auswahl aus mehreren Wurzelelementen



Wählen Sie die Wurzel, die angezeigt werden soll. Wählen Sie das zu verwendende Wurzelelement aus. Standardmäßig wird das erste Wurzelelement in der XSD- oder DTD-Struktur verwendet.

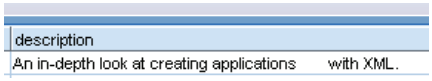
Entfernen unerwünschter Leerzeichen aus XML-Quelldaten

Zeilenumbrüche können in XML-Quelldaten mit der Zeichenkombination [CR][LF] erzeugt werden. In manchen Fällen können diese Zeilenumbrüche mitten im Text auftreten, zum Beispiel:

```
<Beschreibung>Ein tiefer Einblick in das Erstellen von Anwendungen[CR][LF]
mit XML.</Beschreibung>
```

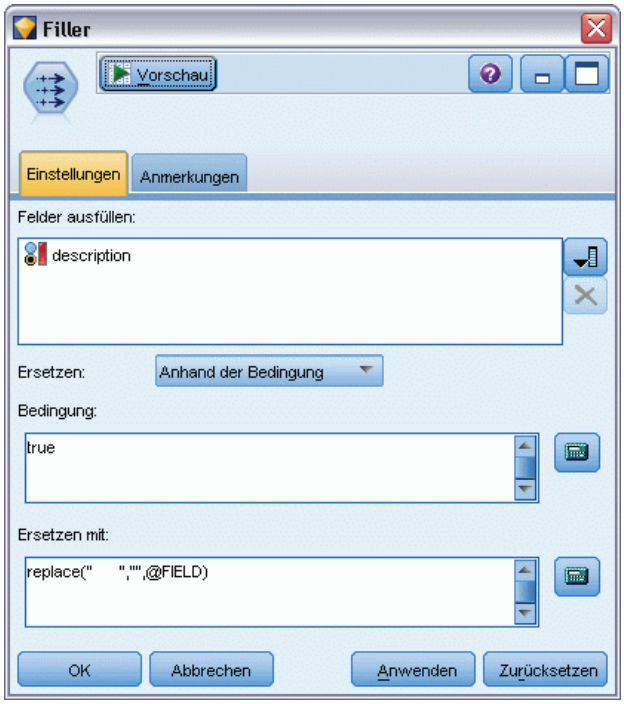
Diese Zeilenumbrüche sind unter Umständen nicht sichtbar, wenn die Datei in bestimmten Anwendungen, etwa einem Webbrowser, geöffnet wird. Wenn die Daten jedoch durch den XML-Quellenknoten in den Stream eingelesen werden, werden die Zeilenumbrüche in eine Reihe von Leerzeichen umgewandelt, zum Beispiel:

Abbildung 2-29
XML-Datensatz mit als Leerzeichen angezeigtem Zeilenumbruch



Sie können diese unerwünschten Leerzeichen mithilfe eines Füllerknotens beseitigen:

Abbildung 2-30
Füllerknoten mit Einstellungen zum Entfernen von Leerzeichen

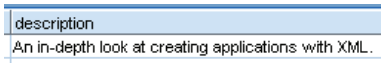


Wie Sie dazu vorgehen, erfahren Sie an einem Beispiel:

- Hängen Sie einen Füllerknoten an den XML-Quellenknoten an.
- Öffnen Sie den Füllerknoten und wählen Sie mithilfe der Feldauswahl das Feld mit den unerwünschten Leerzeichen aus.
- Setzen Sie die Option Ersetzen auf Anhand der Bedingung und die Option Bedingung auf wahr.
- Geben Sie im Feld Ersetzen durch`replace(" ","",@FIELD)` ein und klicken Sie auf "OK".
- Hängen Sie einen Tabellenknoten an den Füllerknoten an und führen Sie den Stream aus.

In der Ausgabe des Tabellenknotens wird der Text nun wie folgt angezeigt:

Abbildung 2-31
XML-Datensatz mit entfernten unerwünschten Leerzeichen



Benutzereingabeknoten

Der Benutzereingabeknoten bietet eine einfache Möglichkeit, künstliche Daten zu erstellen. Dazu können entweder neue Daten ohne Vorlage erstellt oder vorhandene Daten geändert werden. Diese Funktion ist nützlich, wenn Sie z. B. ein Test-Daten-Set für die Modellierung erstellen möchten.

Erstellen von Daten ohne Vorlage

Der Benutzereingabeknoten ist von der Palette der Datenquellen aus verfügbar und kann direkt zum Stream-Zeichenbereich hinzugefügt werden.

- Klicken Sie auf die Registerkarte Datenquellen der Knotenpalette.
- Fügen Sie den Benutzereingabeknoten durch Ziehen und Ablegen oder durch Doppelklicken zum Stream-Zeichenbereich hinzu.
- Öffnen Sie das Dialogfeld durch Doppelklicken und geben Sie Felder und Werte an.

Hinweis: In der Palette der Datenquellen ausgewählte Benutzereingabeknoten sind komplett leer und enthalten keine Felder oder Dateninformationen. So können Sie künstliche Daten vollkommen neu ohne Vorlage erstellen.

Erzeugen von Daten von einer vorhandenen Datenquelle aus

Abbildung 2-32
Benutzereingabeknoten aus einem Stream-Knoten erzeugt

Einen Benutzereingabeknoten können Sie auch von jedem Nichtendknoten im Stream aus erzeugen:

- Überlegen Sie sich, an welchem Punkt des Streams Sie einen Knoten ersetzen möchten.
- Klicken Sie mit der rechten Maustaste auf den Knoten, der seine Daten in den Benutzereingabeknoten speist, und wählen Sie im Menü die Option Benutzereingabeknoten generieren.
- Der Benutzereingabeknoten wird mit allen Prozessen weiter unten im Stream angezeigt und ersetzt den vorhandenen Knoten an dem ausgewählten Punkt Ihres Daten-Streams. Nachdem

der Knoten erzeugt wurde, übernimmt er die gesamte Datenstruktur und Feldtypinformationen (sofern verfügbar) von den Metadaten.

Hinweis: Wenn Daten nicht alle Knoten im Stream durchlaufen haben, sind die Knoten nicht vollständig als Instanz generiert. Dies bedeutet, dass der Speichertyp und Datenwerte unter Umständen nicht verfügbar sind, wenn der Knoten durch einen Benutzereingabeknoten ersetzt wird.

Festlegen von Optionen für den Benutzereingabeknoten

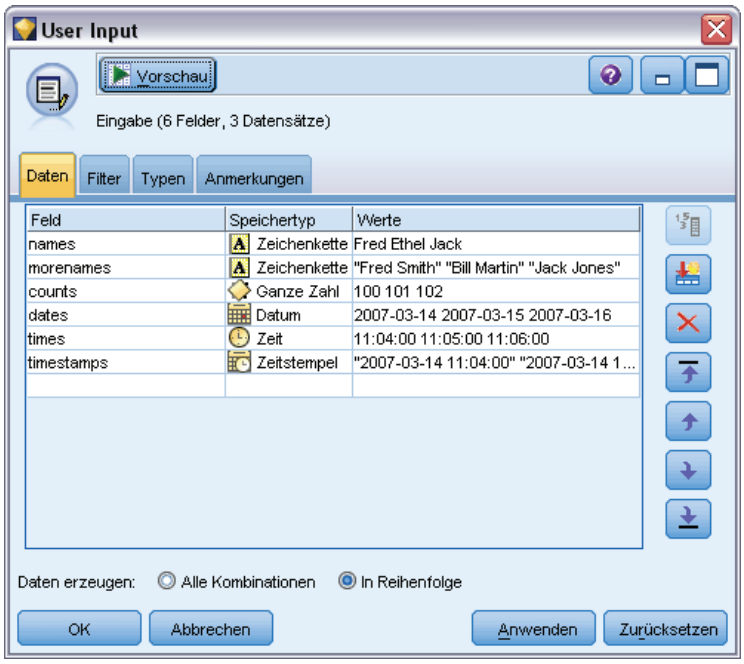
Das Dialogfeld für einen Benutzereingabeknoten enthält mehrere Tools, mit denen Sie Werte eingeben und die Datenstruktur für künstliche Daten definieren können. Bei einem generierten Knoten enthält die Tabelle auf der Registerkarte "Daten" Feldnamen aus der ursprünglichen Datenquelle. Bei einem Knoten, der von der Palette der Datenquellen hinzugefügt wurde, ist die Tabelle leer. Anhand der Tabellenoptionen können Sie folgende Aufgaben durchführen:

- Neue Felder mit der Schaltfläche "Neues Feld hinzufügen" rechts neben der Tabelle hinzufügen
- Vorhandene Felder umbenennen
- Den Datenspeichertyp für jedes Feld festlegen
- Werte angeben
- Die Reihenfolge der Felder in der Anzeige ändern.

Eingeben von Daten

Für jedes Feld können Sie Werte festlegen oder vom ursprünglichen Daten-Set aus über die Schaltfläche zur Wertauswahl rechts neben der Tabelle einfügen. Weitere Informationen zur Angabe von Werten finden Sie in den unten beschriebenen Regeln. Sie können das Feld auch leer lassen. Leere Felder werden mit dem systemdefinierten Nullwert (`$null$`) aufgefüllt.

Abbildung 2-33
Festlegen des Speichertyps für Felder in einem generierten Benutzereingabeknoten



Zeichenkettenwerte können Sie einfach durch Leerzeichen getrennt in die Werte-Spalte eingeben:

Fritz Tanja Martin

Zeichenketten, die Leerzeichen enthalten, können in doppelte Leerzeichen gesetzt werden:

"Willi Schmidt" "Fritz Martin" "Jochen Berger"

Bei numerischen Feldern können Sie mehrere Werte auf gleiche Weise eingeben (d. h. mit Leerzeichen):

10 12 14 16 18 20

Sie können jedoch diese Reihe von Werten auch angeben, indem Sie die Grenzwerte (10, 20) festlegen und die Schritte dazwischen (2). Bei dieser Methode lautet die Eingabe wie folgt:

10,20,2

Beide Methoden können auch miteinander kombiniert werden, indem die eine in die andere eingebettet wird. Beispiel:

1 5 7 10,20,2 21 23

Das Ergebnis dieser Eingabe sind folgende Werte:

1 5 7 10 12 14 16 18 20 21 23

Datums- und Zeitwerte können unter Verwendung des aktuellen Standardformats eingegeben werden, das im Dialogfeld “Stream-Eigenschaften” ausgewählt wird. Beispiele:

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

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

Bei Zeitstempelwerten, die aus einer Datums- und einer Zeitkomponente bestehen, müssen doppelte Anführungszeichen verwendet werden:

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

Weitere Details finden Sie weiter unten in den Kommentaren zum Datenspeichertyp.

Daten erzeugen. Ermöglicht die Angabe, wie die Datensätze generiert werden sollen, wenn Sie den Stream ausführen.

- **Alle Kombinationen**. Generiert Datensätze, die jede mögliche Kombination der Feldwerte enthalten, sodass jeder Feldwert in mehreren Datensätzen enthalten ist. Dadurch können zuweilen mehr Daten generiert werden als gewünscht. Daher wird nach diesem Knoten häufig ein Stichprobenknoten eingefügt.
- **In Reihenfolge**. Generiert Datensätze in der Reihenfolge, in der die Datenfeldwerte angegeben werden. Jeder Feldwert kommt in einem Datensatz jeweils nur einmal vor. Die Gesamtzahl der Datensätze entspricht der höchsten Anzahl von Werten für ein einzelnes Feld. Wenn die Felder weniger Werte enthalten als die größte Anzahl, werden nicht definierte Werte ($null$) eingefügt.

Durch die folgenden Einträge werden beispielsweise die in den Tabellen unten aufgeführten Datensätze generiert.

- **Alter.** 30,60,10
- **BP.** NIEDRIG
- **Cholesterol.** NORMAL HOCH
- **Medikament.** (leer)

Daten erzeugen gesetzt auf Alle Kombinationen:

Alter	BP	Cholesterol	Medikament
30	NIEDRIG	NORMAL	$null$
30	NIEDRIG	HOCH	$null$
40	NIEDRIG	NORMAL	$null$
40	NIEDRIG	HOCH	$null$
50	NIEDRIG	NORMAL	$null$
50	NIEDRIG	HOCH	$null$
60	NIEDRIG	NORMAL	$null$
60	NIEDRIG	HOCH	$null$

Daten erzeugen gesetzt auf In Reihenfolge:

Alter	BP	Cholesterol	Medikament
30	NIEDRIG	NORMAL	$null$
40	$null$	HOCH	$null$
50	$null$	$null$	$null$
60	$null$	$null$	$null$

Datenspeichertyp

Der Speichertyp beschreibt die Art und Weise, wie Daten in einem Feld gespeichert werden. Beispiel: Ein Feld mit den Werten 1 und 0 speichert ganzzahlige Daten. Dies ist vom Messniveau zu unterscheiden, das die Verwendung der Daten beschreibt und sich nicht auf den Speichertyp auswirkt. Beispiel: Sie möchten das Messniveau für ein Feld ganzer Zahlen mit den Werten 1 und 0 auf *Flag* setzen. Das bedeutet normalerweise, dass 1=*True* und 0=*False* ist. Während der Speichertyp stets an der Quelle festgelegt werden muss, kann das Messniveau mithilfe eines Typknotens an jeder beliebigen Stelle im Stream geändert werden. Für weitere Informationen siehe Thema Messniveaus in Kapitel 4 auf S. 138.

Folgende Speichertypen sind verfügbar:

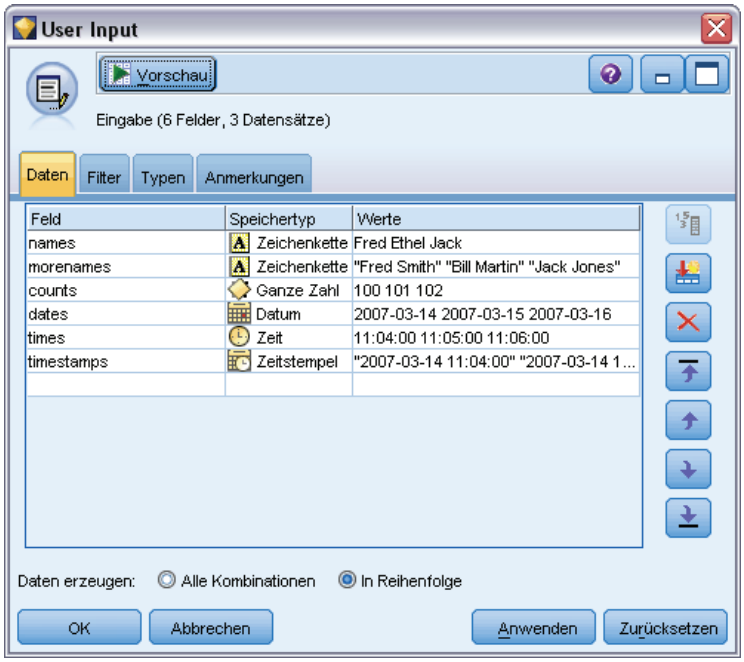
- **String.** Wird für Felder verwendet, die nicht numerische Daten enthalten (auch als alphanumerische Daten bezeichnet). Eine Zeichenkette kann jede beliebige Abfolge von Zeichen enthalten, beispielsweise *fred*, *Klasse 2* oder *1234*. Beachten Sie, dass die Zahlen in Zeichenketten nicht für Berechnungen verwendet werden können.
- **Ganze Zahl.** Ein Feld, bei dessen Werten es sich um ganze Zahlen handelt.
- **Reelle Zahl.** Bei den Werten handelt es sich um Zahlen, die Dezimalstellen enthalten können (nicht auf ganze Zahlen beschränkt). Das Anzeigeformat wird im Dialogfeld für die Stream-Eigenschaften angegeben und kann für einzelne Felder in einem Typknoten überschrieben werden (Registerkarte “Format”).
- **Datum.** Datumswerte, angegeben in einem Standardformat, wie Jahr, Monat und Tag (z. B. 2007-09-26). Das jeweilige Format wird im Dialogfeld für die Stream-Eigenschaften angegeben.
- **Uhrzeit.** Als Dauer gemessene Zeit. Beispielsweise kann ein Service-Call, der 1 Stunde, 26 Minuten und 38 Sekunden dauerte, als 01:26:38 angegeben werden, je nachdem, welches Zeitformat aktuell im Dialogfeld für die Stream-Eigenschaften angegeben ist.
- **Zeitstempel.** Werte, die sowohl eine Datums- als auch eine Zeitkomponente enthalten, wie beispielsweise 2007–09–26 09:04:00; auch hier wieder abhängig von den aktuellen Formaten für Datum und Zeit im Dialogfeld “Stream-Eigenschaften”. Beachten Sie, dass Zeiststempelwerte ggf. in Anführungszeichen gesetzt werden müssen, um sicherzustellen, dass sie als Einzelwert interpretiert werden und nicht als gesonderte Datums- und Zeitwerte. (Dies gilt beispielsweise bei der Eingabe von Werten in einem Benutzereingabeknoten.)

Speichertypkonvertierung. Der Speichertyp für ein Feld kann mit verschiedenen Konvertierungsfunktionen, z. B. to_string und to_integer, in einem Füllerknoten geändert werden. Für weitere Informationen siehe Thema Speichertypkonvertierung mithilfe des Füllerknotens in Kapitel 4 auf S. 181. Beachten Sie, dass die Konvertierungsfunktionen (und alle anderen Funktionen, für die ein spezieller Eingabetyp, wie beispielsweise ein Wert für Datum oder Uhrzeit, erforderlich ist) von den aktuell im Dialogfeld für die Stream-Eigenschaften angegebenen Formaten abhängen. Wenn Sie beispielsweise ein Zeichenkettenfeld mit den Werten *Jan 2003*, *Feb 2003* (usw.) in einen Datumsspeicher konvertieren müssen, wählen Sie MON JJJJ als Standard-Datumsformat für den Stream aus. Konvertierungsfunktionen sind auch im Ableitungsknoten zur temporären Konvertierung während einer Ableitungsberechnung verfügbar. Mit dem Ableitungsknoten können Sie auch andere Bearbeitungen vornehmen wie beispielsweise die Umkodierung von Zeichenkettenfeldern mit kategorialen Werten. Für weitere Informationen siehe Thema Umkodieren von Werten mit dem Ableitungsknoten in Kapitel 4 auf S. 178.

Einlesen gemischter Daten. Beachten Sie, dass beim Einlesen von Feldern mit numerischem Speichertyp (ganze Zahl, reelle Zahl, Zeit, Zeitstempel oder Datum) alle nicht numerischen Werte auf null oder auf systemdefiniert fehlend gesetzt werden. Dies liegt daran, dass IBM® SPSS® Modeler im Gegensatz zu einigen anderen Anwendungen keine gemischten Speichertypen innerhalb eines Felds zulässt. Um dies zu vermeiden, sollten alle Felder mit gemischten Daten als Zeichenketten eingelesen werden, indem der Speichertyp im Quellenknoten oder in der externen Anwendung nach Bedarf geändert wird.

Hinweis: Generierte Benutzereingabeknoten enthalten unter Umständen bereits Speicherinformationen, die aus dem Quellenknoten gesammelt wurden, sofern der Knoten als Instanz generiert wurde. Ein nicht als Instanz generierter Knoten enthält keine Speichertyp- oder Verwendungstypinformationen.

Abbildung 2-34
Festlegen des Speichertyps für Felder in einem generierten Benutzereingabeknoten



Regeln für das Festlegen von Werten

Bei symbolischen Feldern sollten zwischen den Werten Leerzeichen stehen. Beispiel:

HOCH MITTEL NIEDRIG

Bei numerischen Feldern können Sie mehrere Werte auf gleiche Weise eingeben (d. h. mit Leerzeichen):

10 12 14 16 18 20

Sie können jedoch diese Reihe von Werten auch angeben, indem Sie die Grenzwerte (10, 20) festlegen und die Schritte dazwischen (2). Bei dieser Methode lautet die Eingabe wie folgt:

10,20,2

Beide Methoden können auch miteinander kombiniert werden, indem die eine in die andere eingebettet wird. Beispiel:

1 5 7 10,20,2 21 23

Das Ergebnis dieser Eingabe sind folgende Werte:

1 5 7 10 12 14 16 18 20 21 23

Allgemeine Registerkarten für Quellenknoten

Folgende Optionen können für alle Quellenknoten festgelegt werden, indem Sie auf die entsprechende Registerkarte klicken:

- **Registerkarte "Daten".** Dient zum Ändern des Standardspeichertyps.
- **Registerkarte "Filter".** Dient zum Entfernen oder Umbenennen von Datenfeldern. Die Registerkarte bietet dieselben Funktionen wie der Filterknoten. Für weitere Informationen siehe Thema Festlegen der Filteroptionen in Kapitel 4 auf S. 157.
- **Registerkarte "Typen".** Wird verwendet, um Messniveaus festzulegen. Die Registerkarte bietet dieselben Funktionen wie der Typknoten.
- **Registerkarte "Anmerkungen".** Wird für alle Knoten verwendet. Die Registerkarte bietet Optionen zum Umbenennen von Knoten, zum Anzeigen einer benutzerdefinierten QuickInfo und zum Speichern einer längeren Anmerkung.

Festlegen von Messniveaus im Quellenknoten

Die Feldeigenschaften können in einem Quellenknoten oder in einem separaten Typknoten angegeben werden. Die Funktionsweise ist bei beiden Knoten ähnlich. Folgende Eigenschaften stehen zur Verfügung:

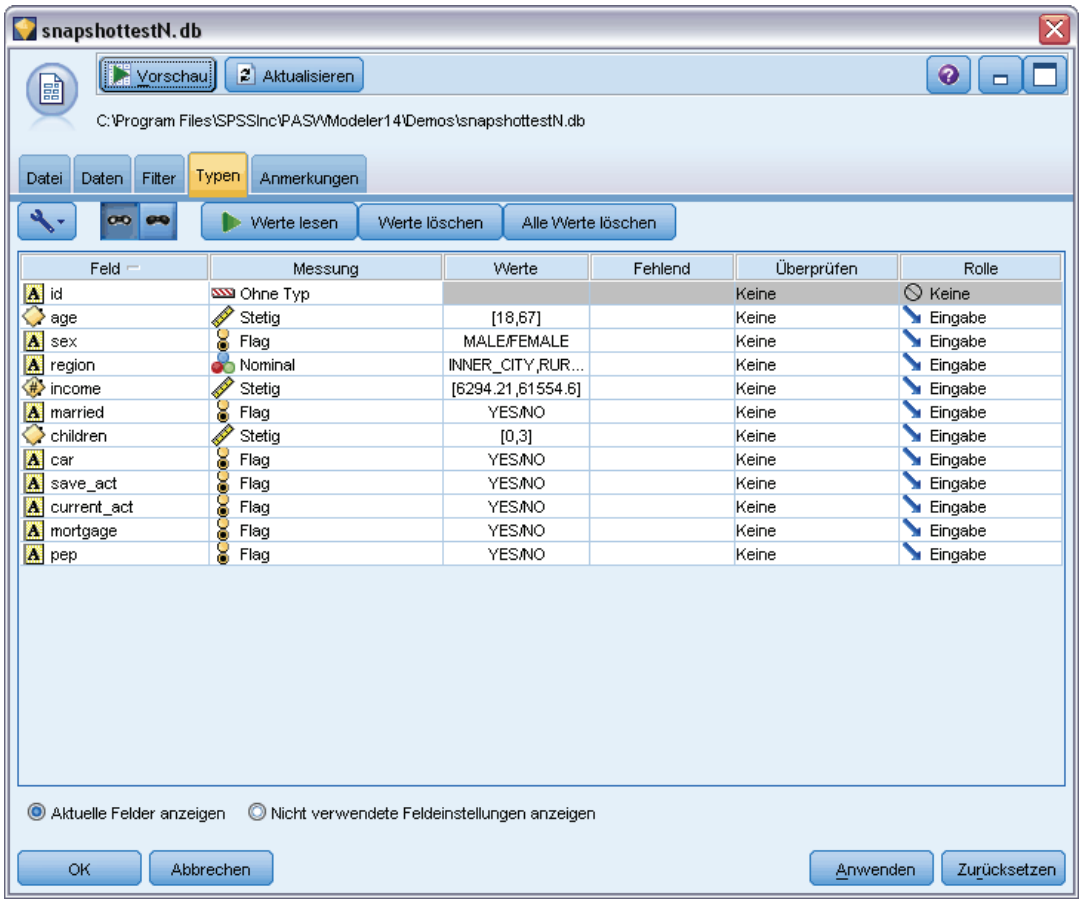
- **Feld.** Doppelklicken Sie auf einen beliebigen Feldnamen, um Werte- und Feldbeschriftungen für Daten in IBM® SPSS® Modeler anzugeben. So können aus IBM® SPSS® Statistics beispielsweise importierte Feldmetadaten hier angezeigt oder geändert werden. Auf ähnliche Weise können Sie auch neue Beschriftungen für Felder und ihre Werte erstellen. Die Beschriftungen, die Sie hier angeben, werden überall in SPSS Modeler angezeigt, je nach der von Ihnen im Dialogfeld "Stream-Eigenschaften" getroffenen Auswahl.
- **Messung.** Dies ist das Messniveau, das zur Beschreibung der Eigenschaften von Daten in einem bestimmten Feld verwendet wird. Wenn alle Details eines Felds bekannt sind, wird es als **vollständig instanziiert** bezeichnet. Für weitere Informationen siehe Thema Messniveaus in Kapitel 4 auf S. 138.

  Anmerkung: Das Messniveau eines Felds ist etwas anderes als sein Speichertyp, der angibt, ob die Daten als Zeichenkette, ganze Zahl, reelle Zahl, Datum, Zeit oder Zeitstempel gespeichert werden sollen.
- **Werte.** In dieser Spalte können Sie Optionen zum Lesen von Datenwerten aus dem Daten-Set auswählen oder die Option Angeben verwenden, um Messniveaus und Werte in einem separaten Dialogfeld anzugeben. Sie können auch Felder übergeben, ohne ihre Werte zu lesen. Für weitere Informationen siehe Thema Datenwerte in Kapitel 4 auf S. 143.

- **Fehlend.** Wird verwendet, um anzugeben, wie fehlende Werte für das Feld behandelt werden. Für weitere Informationen siehe Thema Fehlende Werte definieren in Kapitel 4 auf S. 149.
- **Überprüfen.** In dieser Spalte können Sie Optionen festlegen, um sicherzustellen, dass die Feldwerte den angegebenen Werten oder Bereichen entsprechen. Für weitere Informationen siehe Thema Überprüfen von Typenwerten in Kapitel 4 auf S. 149.
- **Rolle.** Wird verwendet, um Modellierungsknoten mitzuteilen, ob es sich bei Feldern um Eingabefelder (Prädiktorfelder) oder Zielfelder (vorhergesagte Felder) für einen Maschinenlernprozess handelt. Beides und Keine sind auch verfügbare Rollen, zusammen mit Partition, das ein Feld bezeichnet, das für die Aufteilung von Datensätzen in separate Stichproben zu Training-, Test- und Validierungszwecken verwendet wird. Der Wert Aufteilung gibt an, dass für jeden möglichen Wert des Felds separate Modelle erstellt werden. Für weitere Informationen siehe Thema Festlegen der Feldrolle in Kapitel 4 auf S. 150.

Für weitere Informationen siehe Thema Typknoten in Kapitel 4 auf S. 136.

Abbildung 2-35
Optionen der Registerkarte "Typen"



Zeitpunkt der Instanziierung am Quellenknoten

Es gibt zwei Möglichkeiten, Informationen über den Datenspeichertyp und die Werte Ihrer Felder abzurufen. Diese **Instanziierung** kann entweder am Quellenknoten erfolgen, wenn Sie Daten erstmals in IBM® SPSS® Modeler importieren, oder durch Einfügen eines Typknotens in den Daten-Stream.

Die Instanziierung am Quellenknoten ist in folgenden Fällen nützlich:

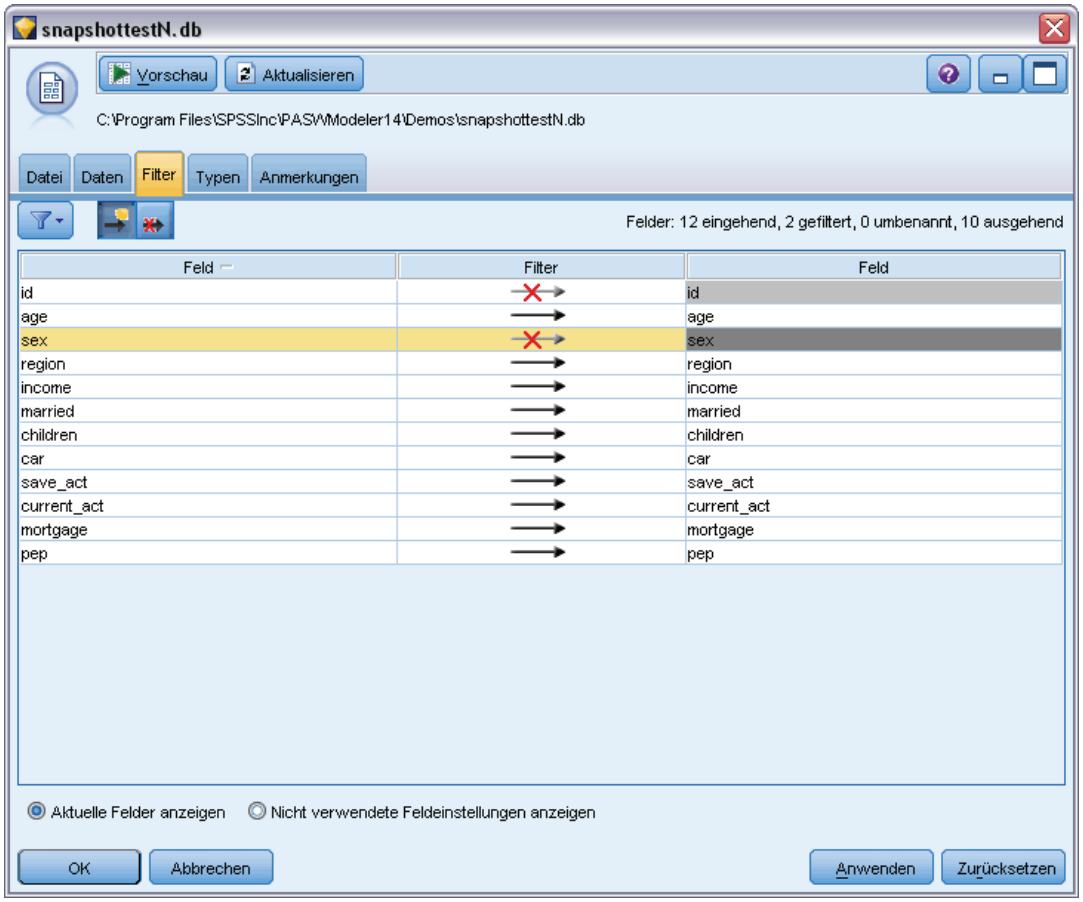
- Wenn das Daten-Set recht klein ist.
- Wenn Sie beabsichtigen, mit Expression Builder neue Felder abzuleiten (durch Instanziierung werden die Feldwerte von Expression Builder verfügbar gemacht).

Im Allgemeinen ist bei nicht allzu großen Daten-Sets und wenn keine Felder später im Stream hinzugefügt werden sollen, eine Instanziierung am Quellenknoten die praktischste Methode.

Filtern von Feldern am Quellenknoten

Mit der Registerkarte “Filter” des Dialogfelds eines Quellenknotens können Sie Felder basierend auf Ihrer anfänglichen Untersuchung der Daten aus Vorgängen weiter unten im Stream ausschließen Diese Funktion ist nützlich, wenn z. B. doppelte Felder in den Daten vorhanden sind oder wenn Sie bereits ausreichend mit den Daten vertraut sind und irrelevante Felder ausschließen können. Alternativ können Sie weiter unten im Stream einen gesonderten Filterknoten einfügen. Die Funktionsweise ist in beiden Fällen ähnlich. Für weitere Informationen siehe Thema Festlegen der Filteroptionen in Kapitel 4 auf S. 157.

Abbildung 2-36
Filtern von Feldern aus dem Quellenknoten



Kapitel
3

Datensatzoperationsknoten

Überblick über die Datensatzoperationen

Mit Datensatzoperationsknoten werden Änderungen an Daten auf der Datensatzebene vorgenommen. Diese Operationen sind wichtig während der **Datenverständnis-** und **Datenvorbereitungs**-Phase des Data Mining, da Sie damit die Daten für Ihre jeweiligen geschäftlichen Anforderungen zuschneiden können.

Sie könnten beispielsweise auf der Grundlage der Ergebnisse des Data Audit, das mit dem Data Audit-Knoten (Ausgabepalette) durchgeführt wurde, zu dem Schluss kommen, dass die Datensätze über die Einkäufe der Kunden für die letzten drei Monate zusammengeführt werden sollten. Mit einem Zusammenführungsknoten ("Mergen") können Sie Datensätze auf der Grundlage der Werte eines Schlüsselfelds, beispielsweise *Kunden-ID*, zusammenführen. Oder Sie könnten feststellen, dass eine Datenbank mit Informationen über Website-Aufrufe unüberschaubar ist, da sie mehr als eine Million Datensätze enthält. Mithilfe von Beispielknoten können Sie eine Untergruppe von Daten für die Modellierung auswählen.

Die Palette "Datensatzoperationen" enthält folgende Knoten:

Der Auswahlknoten wählt auf der Grundlage einer bestimmten Bedingung eine Untergruppe von Datensätzen aus einem Daten-Stream aus oder verwirft sie. Sie können beispielsweise die Datensätze auswählen, die zu einer bestimmten Verkaufsregion gehören. Für weitere Informationen siehe Thema Auswahlknoten auf S. 69.

Der Stichprobenknoten wählte eine Teilmenge der Datensätze aus. Es wird eine Vielzahl von Stichprobentypen unterstützt, darunter geschichtete, gruppierte (Klumpenstichproben) und nichtzufällige (strukturierte) Stichproben. Eine Stichprobenziehung kann nützlich zur Verbesserung der Leistungsfähigkeit und zur Auswahl von verwandten Datensätzen bzw. Transaktionen für die Analyse sein. Für weitere Informationen siehe Thema Stichprobenknoten auf S. 71.

Der Balancierungsknoten korrigiert Ungleichgewichte in einem Daten-Set, sodass dieses eine bestimmte Bedingung erfüllt. Die Balancierungsanweisung passt den Anteil der Datensätze, bei denen eine Bedingung wahr ist, um den angegebenen Faktor an. Für weitere Informationen siehe Thema Balancierungsknoten auf S. 80.

Der Aggregatknoten ersetzt eine Sequenz von Eingabedatensätzen durch zusammengefasste, aggregierte Ausgabedatensätze. Für weitere Informationen siehe Thema Aggregatknoten. auf S. 82.

Mit dem Knoten "RFM-Aggregat" (Recency-, Frequency-, Monetary-Aggregat) können Sie Daten über die früheren Transaktionen von Kunden verwenden, alle nicht benötigten Daten entfernen und alle verbliebenen Transaktionsdaten zu einer einzigen Zeile zusammenfassen, die angibt, wann der betreffende Kunde zuletzt mit Ihnen in Geschäftskontakt stand, wie viele Transaktionen er vorgenommen hat und wie hoch der Gesamtwert dieser Transaktionen ist. Für weitere Informationen siehe Thema RFM-Aggregatknoten auf S. 85.

Der Sortierknoten sortiert Datensätze anhand der Werte eines oder mehrerer Felder in aufsteigender oder absteigender Reihenfolge. Für weitere Informationen siehe Thema Sortierknoten auf S. 88.

Der Zusammenführungsknoten erstellt aus mehreren Eingabedatensätzen einen einzelnen Ausgabedatensatz mit einigen oder allen der Eingabefelder. Er wird zum Zusammenführen von Daten aus verschiedenen Quellen verwendet, beispielsweise Daten über Auslandskunden und erworbene demografische Daten. Für weitere Informationen siehe Thema Zusammenführungsknoten (“Mergen”) auf S. 89.

Der Anhangknoten verkettet Gruppen von Datensätzen miteinander. Er ist insbesondere nützlich für die Kombination von Daten-Sets mit ähnlicher Struktur, aber unterschiedlichen Daten. Für weitere Informationen siehe Thema Anhangknoten auf S. 100.

Der Duplikatknoten entfernt doppelte Datensätze, entweder indem jeweils der erste Datensatz an den Daten-Stream übergeben wird oder aber indem der erste Datensatz verworfen wird und stattdessen etwaige Duplikate an den Stream übergeben werden. Für weitere Informationen siehe Thema Duplikatknoten auf S. 102.

Für viele der Knoten in der Palette “Datensatzoperationen” ist die Verwendung eines CLEM-Ausdrucks erforderlich. Wenn Sie mit CLEM vertraut sind, können Sie einen Ausdruck in das Feld eingeben. Alle Ausdrucksfelder enthalten jedoch eine Schaltfläche zum Öffnen des CLEM Expression Builder, mit dem solche Ausdrücke automatisch erstellt werden.

Abbildung 3-1
Schaltfläche für Expression Builder

Auswahlknoten

Mit Auswahlknoten können Sie eine Untergruppe von Datensätzen aus dem Stream auswählen bzw. verwerfen. Dafür werden spezielle Bedingungen verwendet, beispielsweise BD (Blutdruck) = "HOCH".

Abbildung 3-2
Dialogfeld "Auswahlknoten"

Modalwert. Gibt an, ob Datensätze, die die Bedingung erfüllen, in den Daten-Stream eingeschlossen oder daraus ausgeschlossen werden.

- **Einschließen.** Wählen Sie diese Option, um Datensätze einzuschließen, die die Auswahlbedingung erfüllen.
- **Verwerfen.** Wählen Sie diese Option, um Datensätze auszuschließen, die die Auswahlbedingung erfüllen.

Bedingung. Zeigt die Auswahlbedingung an, die zum Testen der einzelnen Datensätze verwendet wird, die Sie mithilfe eines CLEM-Ausdrucks angeben. Geben Sie entweder einen Ausdruck in das Fenster ein oder verwenden Sie den Expression Builder, den Sie mit der Taschenrechnerschaltfläche rechts neben dem Fenster aufrufen können.

Sie können Datensätze auf der Basis einer Bedingung wie der folgenden verwerfen:

(*var1*='*Wert1*' and *var2*='*Wert2*')

Der Auswahlknoten verwirft in diesem Fall standardmäßig auch Datensätze, die Nullwerte für alle Auswahlfelder enthalten. Um dies zu vermeiden, hängen Sie die folgende Bedingung an die Originalbedingung an:

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

Auswahlknoten werden auch zur Auswahl eines Anteils der Datensätze verwendet. Normalerweise wird für diesen Vorgang ein anderer Knoten, der Stichprobenknoten, verwendet. Wenn die Bedingung, die Sie angeben möchten, jedoch komplexer ist als die zur Verfügung stehenden Parameter, können Sie mithilfe des Auswahlknotens Ihre eigene Bedingung erstellen. Sie können beispielsweise Bedingungen der folgenden Art erstellen:

BD = "HOCH" and random(10) <= 4

Dadurch werden ungefähr 40 % der Datensätze mit hohem Blutdruck ausgewählt und zur weiteren Analyse im Stream weitergegeben.

Stichprobenknoten

Mithilfe von Stichprobenknoten können Sie eine Teilmenge der Datensätze für die Analyse auswählen oder einen Anteil von Datensätzen auswählen, der verworfen werden soll. Es wird eine Vielzahl von Stichprobentypen unterstützt, darunter geschichtete, gruppierte (Klumpenstichproben) und nichtzufällige (strukturierte) Stichproben. Stichprobenziehungen können aus verschiedenen Gründen durchgeführt werden:

- Zur Verbesserung der Leistung durch Schätzung von Modellen anhand einer Teilmenge der Daten. Modelle, die aus einer Stichprobe geschätzt wurden, sind häufig ebenso genau wie Modelle, die aus dem vollständigen Datensatz abgeleitet werden. Das gilt insbesondere, wenn sie durch die verbesserte Leistungsfähigkeit in der Lage sind, mit unterschiedlichen Methoden zu experimentieren, die Sie andernfalls nicht ausprobiert hätten.
- Zur Auswahl von Gruppen verwandter Datensätze oder Transaktionen für die Analyse, beispielsweise alle Artikel in einem Online-Warenkorb oder alle Eigenschaften in einem bestimmten Umfeld.
- Zur Ermittlung von Einheiten oder Fällen zur zufälligen Untersuchung im Rahmen von Qualitätssicherung, Betrugsprävention oder Sicherheitsmaßnahmen.

Anmerkung: Wenn Sie die Daten einfach nur zum Zwecke der Validierung in eine Trainings- und eine Teststichprobe unterteilen möchten, kann stattdessen ein Partitionsknoten verwendet werden. Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 auf S. 208.

Typen von Stichproben

Klumpenstichproben. Hierbei werden Gruppen bzw. Klumpen als Stichprobe gezogen, nicht einzelne Einheiten. Nehmen Sie beispielsweise an, Sie haben eine Datendatei mit einem Datensatz pro Schüler. Wenn Sie nach Schule gruppieren und der Stichprobenumfang 50 % beträgt, werden 50 % der Schulen ausgewählt und aus jeder ausgewählten Schule werden alle Schüler ausgewählt. Die Schüler in den nicht ausgewählten Schulen werden verworfen. Durchschnittlich wäre zu erwarten, dass ungefähr 50 % der Schüler ausgewählt werden, da jedoch die Schulen unterschiedlich groß sind, wird dieser Prozentsatz vermutlich nicht genau erreicht. Auf ähnliche Weise können Sie Artikel in einem Warenkorb nach Transaktions-ID zu Klumpen zusammenfassen, um sicherzustellen, dass alle Artikel aus ausgewählten Transaktionen verwendet werden Ein Beispiel, in dem Immobilien nach Gemeinde zu Klumpen gruppiert werden, finden Sie im Beispiel-Stream *complexsample_property.str*.

Geschichtete Stichproben. Hierbei werden die Stichproben unabhängig innerhalb von sich nicht überschneidenden Untergruppen der Grundgesamtheit, den so genannten Schichten, ausgewählt. So können Sie beispielsweise sicherstellen, dass Männer und Frauen zu gleichen Anteilen ausgewählt werden oder dass jede Region oder sozioökonomische Gruppe innerhalb der Einwohner einer Stadt repräsentiert wird. Außerdem können Sie für jede Schicht einen anderen Stichprobenumfang angeben (z. B. wenn Sie annehmen, dass eine Gruppe in den ursprünglichen

Daten unterrepräsentiert ist). Ein Beispiel, in dem Immobilien nach County geschichtet werden, finden Sie im Beispiel-Stream *complexsample_property.str*.

Systematische Stichprobenziehung (Stichprobenziehung vom Typ "1 in n"). Wenn eine zufällige Auswahl schwer zu erzielen ist, können die Stichprobeneinheiten systematisch (in festgelegten Intervallen) oder sequenziell gezogen werden.

Stichprobengewichte. Stichprobengewichte werden beim Ziehen einer komplexen Stichprobe automatisch berechnet und entsprechen ungefähr der "Häufigkeit" der einzelnen gezogenen Einheiten in den ursprünglichen Daten. Daher sollte die Summe der Gewichte in der gesamten Stichprobe eine Schätzung des Umfangs der ursprünglichen Daten darstellen.

Stichprobenrahmen

Ein Stichprobenrahmen definiert die potenzielle Quelle der in eine Stichprobe oder Studie aufzunehmenden Fälle. In einigen Fällen kann es möglich sein, jedes einzelne Mitglied einer Grundgesamtheit zu ermitteln und jedes beliebige davon in eine Stichprobe aufzunehmen. Dies ist beispielsweise bei der Stichprobenziehung aus Artikeln von einem Fließband der Fall. In den meisten Fällen besteht jedoch nicht auf jeden möglichen Fall Zugriff. So können Sie beispielsweise nicht sicher sein, welche Personen bei einer Wahl abstimmen wird, bis die Wahl stattgefunden hat. In diesem Fall können Sie in den USA beispielsweise das Wählerregister als Stichprobenrahmen verwenden, auch wenn einige registrierte Personen nicht abstimmen werden und wenn einige Personen möglicherweise abstimmen, obwohl sie zu dem Zeitpunkt, als Sie Einsicht in das Register nahmen, noch nicht aufgeführt waren. Personen, die sich nicht im Stichprobenrahmen befinden, können auch nicht in die Stichprobe aufgenommen werden. Ob Ihr Stichprobenrahmen hinsichtlich seiner Natur hinreichend große Ähnlichkeit mit der Grundgesamtheit aufweist, die Sie evaluieren möchten, ist eine Frage, die für jeden realen Fall gesondert zu untersuchen ist.

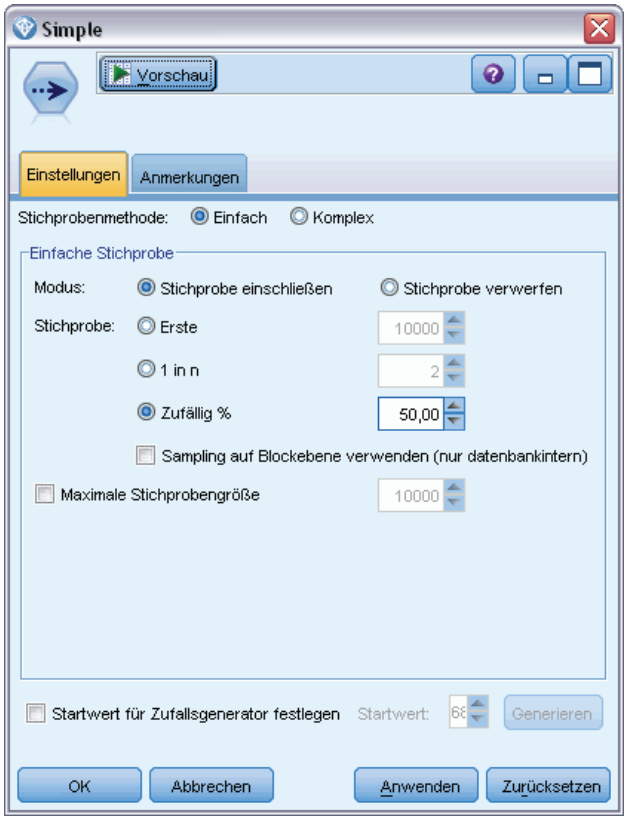
Optionen für Stichprobenknoten

Sie können, je nach Anforderung, die Methode Einfach oder Komplex auswählen.

Einfache Stichproben – Optionen

Mit der Methode "Einfach" können Sie einen Zufallsprozentsatz von Datensätzen, zusammenhängende Datensätze oder einfach jeden *n-ten* Datensatz auswählen.

Abbildung 3-3
Optionen für einfache Stichproben



Modalwert. Wählen Sie aus, ob Datensätze für die folgenden Modi übergeben (eingeschlossen) oder verworfen (ausgeschlossen) werden sollen:

- **Stichprobe einschließen.** Nimmt die ausgewählten Datensätze in den Daten-Stream auf und verwirft alle anderen. Beispiel: Wenn Sie den Modus auf Stichprobe einschließen und die Option 1 in n auf "5" setzen, wird jeder 5. Datensatz in den Daten-Stream aufgenommen und es ergibt sich ein Daten-Set mit ungefähr einem Fünftel der ursprünglichen Größe. Dies ist der Standardmodus bei der Stichprobenziehung von Daten und der einzige Modus, der bei der Methode "Komplex" zur Verfügung steht.
- **Stichprobe verwerfen.** Schließt die ausgewählten Datensätze aus und nimmt alle anderen auf. Beispiel: Wenn Sie den Modus auf Stichprobe verwerfen und die Option 1 in n auf "5" setzen, wird jeder 5. Datensatz verworfen. Dieser Modus ist nur bei der Methode "Einfach" verfügbar.

Beispiel. Wählen Sie die Methode der Stichprobenziehung aus den folgenden Optionen aus:

- **Erste.** Verwenden Sie diese Option, um eine Stichprobenziehung mit zusammenhängenden Daten durchzuführen. Beispiel: Wenn die maximale Stichprobengröße auf "1000" gesetzt ist, werden die ersten 10.000 Datensätze ausgewählt.

- **1 in n.** Wählen Sie diese Option aus, um Stichproben zu ziehen, indem jeder *n*-te Datensatz übergeben bzw. verworfen wird. Wenn z. B. "*n*" auf "5" gesetzt ist, wird jeder 5. Datensatz ausgewählt.
- **Zufällig %.** Wählen Sie diese Option aus, um per Zufallsgenerator einen festgelegten Prozentsatz der Daten als Stichprobe zu ziehen. Beispiel: Wenn der Prozentsatz auf "20" gesetzt wird, werden 20 % der Daten entweder an den Daten-Stream übergeben oder verworfen, je nach dem ausgewählten Modus. Geben Sie mithilfe des Felds einen Prozentsatz für die Stichprobenziehung an. Mit dem Steuerelement Startwert für Zufallsgenerator festlegen können Sie außerdem einen Startwert bestimmen.

  Sampling auf Blockebene verwenden (nur datenbankintern). Diese Option ist nur aktiviert, wenn Sie beim Durchführen von In-Database Mining auf einer Oracle- oder IBM DB2-Datenbank einen Zufallsprozentsatz für die Stichprobenbeziehung auswählen. In diesem Fall kann es effizienter sein, Sampling auf Blockebene zu verwenden.

Maximale Stichprobengröße. Gibt an, wie viele Datensätze maximal in die Stichprobe aufgenommen werden sollen. Diese Option ist redundant und wird daher deaktiviert, wenn Erste und Einschließen ausgewählt wurden. Beachten Sie auch, dass diese Einstellung bei Verwendung in Kombination mit der Option Zufällig % dazu führen kann, dass bestimmte Datensätze nicht ausgewählt werden. Wenn Ihr Daten-Set beispielsweise 10 Millionen Datensätze enthält und Sie bei einem maximalen Stichprobenumfang von 3 Millionen 50 % der Datensätze auswählen, führt das dazu, dass nur die ersten 6 Millionen Datensätze jeweils mit 50%iger Wahrscheinlichkeit ausgewählt werden und die restlichen vier Millionen Datensätze keine Chance haben, in die Stichprobe aufgenommen zu werden. Um diese Einschränkung zu vermeiden, müssen Sie die Methode Komplex für die Stichprobenziehung auswählen und eine Standardstichprobe von 3 Millionen Datensätzen anfordern, ohne eine Klumpen- oder Schichtungsvariable anzugeben.

Komplexe Stichproben – Optionen

Die Optionen für komplexe Stichproben gestatten eine feinere Steuerung der Stichprobe. So können beispielsweise neben anderen Optionen gruppierte (Klumpenstichproben), geschichtete und gewichtete Stichproben festgelegt werden.

Abbildung 3-4
Optionen für komplexe Stichproben

Klumpen und Schichtung. Mit dieser Option können Sie bei Bedarf Felder für Klumpen, Schichtung und Eingabegewichtung angeben. Für weitere Informationen siehe Thema Einstellungen unter "Klumpen und Schichtung" auf S. 76.

Stichprobentyp.

- **Zufällig.** Wählt Klumpen oder Datensätze innerhalb der einzelnen Schichten nach dem Zufallsprinzip aus.
- **Systematisch.** Wählt Datensätze in festen Intervallen aus. Diese Option hat dieselbe Wirkung wie die Methode *1 in n*, mit der Ausnahme, dass sich die Position des ersten Datensatzes in Abhängigkeit von einem Zufallsstartwert ändert. Der Wert von *n* wird automatisch auf der Grundlage der Stichprobengröße bzw. des Anteils ermittelt.

Stichprobeneinheiten. Sie können Anteile oder Anzahl (beobachtete Werte) als Grundeinheiten für die Stichprobe auswählen.

Stichprobenumfang. Es gibt mehrere Möglichkeiten zur Festlegung des Stichprobenumfangs:

- **Fest.** Hiermit können Sie den Gesamtumfang der Stichprobe als Anzahl (beobachtete Werte) oder Anteil angeben.

- **Benutzerdefiniert.** Ermöglicht die Angabe des Stichprobenumfangs für die einzelnen Untergruppen oder Schichten. Diese Option ist nur verfügbar, wenn im Unterdialogfeld "Klumpen und Schichtung" ein Schichtungsfeld angegeben wurde.
- **Variable.** Ermöglicht dem Benutzer die Auswahl eines Felds, mit dem der Stichprobenumfang für die einzelnen Untergruppen oder Schichten definiert werden kann. Dieses Feld sollte für jeden Datensatz innerhalb einer bestimmten Schicht denselben Wert aufweisen; wenn die Stichprobe beispielsweise nach County geschichtet wird, müssen alle Datensätze mit *county = Surrey* denselben Wert aufweisen. Das Feld muss numerisch sein und seine Werte müssen mit den ausgewählten Stichprobeneinheiten übereinstimmen. Bei Anteilen sollten die Werte größer als 0 und kleiner als 1 sein; bei den beobachteten Werten ist der Mindestwert 1.

Minimale Stichprobe pro Schicht. Gibt die Mindestanzahl an Datensätzen (bzw. die Mindestanzahl an Clustern, wenn ein Cluster-Feld angegeben wurde) an.

Maximale Stichprobe pro Schicht. Legt die maximale Anzahl an Datensätzen oder Clustern fest. Wenn Sie diese Option auswählen, ohne einen Cluster oder ein Schichtungsfeld anzugeben, wird eine Zufallsstichprobe oder systematische Stichprobe mit der angegebenen Größe ausgewählt.

Startwert für Zufallsgenerator festlegen. Bei der Stichprobenziehung oder Partitionierung von Datensätzen auf der Grundlage eines Zufallsprozentsatzes können Sie mit dieser Option dieselben Ergebnisse in einer anderen Sitzung replizieren. Wenn Sie den vom Zufallszahlengenerator verwendeten Startwert angeben, stellen Sie sicher, dass bei jeder Ausführung des Knotens dieselben Datensätze zugewiesen werden. Geben Sie den gewünschten Startwert ein oder klicken Sie auf die Schaltfläche Generieren, um automatisch einen Startwert zu generieren. Wenn diese Option nicht ausgewählt ist, wird bei jeder Ausführung des Knotens eine andere Stichprobe generiert.

Anmerkung: Bei Verwendung der Option Startwert für Zufallsgenerator festlegen mit Datensätzen, die aus einer Datenbank eingelesen wurden, ist möglicherweise vor der Stichprobenziehung ein Sortierknoten erforderlich, um zu gewährleisten, dass bei jeder Ausführung des Knotens dasselbe Ergebnis erzielt wird. Dies liegt daran, dass der Startwert für den Zufallsgenerator von der Reihenfolge der Datensätze abhängt, die in relationalen Datenbanken nicht unbedingt gleich bleibt. Für weitere Informationen siehe Thema Sortierknoten auf S. 88.

Einstellungen unter "Klumpen und Schichtung"

Im Dialogfeld "Klumpen und Schichtung" können Sie beim Ziehen einer komplexen Stichprobe Felder für Klumpen, Schichtung und Gewichtung auswählen.

Abbildung 3-5
Einstellungen für Klumpen- und Schichtungsfelder

Cluster. Gibt ein kategoriales Feld an, das für die Gruppierung von Datensätzen verwendet wird. Datensätze werden anhand Ihrer Zugehörigkeit zu bestimmten Klumpen bei der Stichprobenziehung berücksichtigt. Dabei werden bestimmte Klumpen aufgenommen und andere nicht. Wenn jedoch ein Datensatz aus einem bestimmten Klumpen aufgenommen wird, werden auch alle anderen aufgenommen. Bei der Analyse von Verbindungen zwischen Produkten in Warenkörben könnten Sie beispielsweise die Artikel nach Transaktions-ID gruppieren, um sicherzustellen, dass alle Artikel aus den ausgewählten Transaktionen verwendet werden. Anstatt bei der Stichprobenziehung einzelne Datensätze auszuwählen, wodurch Informationen darüber, welche Artikel gemeinsam verkauft wurden, verloren gehen würden, können Sie bei der Stichprobenziehung Transaktionen auswählen, um sicherzugehen, dass alle Datensätze der ausgewählten Transaktionen erhalten bleiben.

Schichten nach. Dient zur Angabe eines kategorialen Felds, mit dem Datensätze geschichtet werden können, sodass die Stichproben unabhängig innerhalb von sich nicht überschneidenden Untergruppen der Grundgesamtheit, den so genannten Schichten, ausgewählt werden. Wenn Sie beispielsweise eine Stichprobe mit dem Umfang 50 % auswählen, die nach Geschlecht geschichtet ist, werden zwei 50-Prozent-Stichproben gezogen, eine für die Männer und eine für die Frauen. Weitere Beispiele für Schichten sind sozioökonomische Gruppen, Berufskategorien, Altersgruppen oder ethnische Gruppen. Mithilfe von Schichten können Sie angemessene Stichprobengrößen für relevante Untergruppen gewährleisten. Wenn im ursprünglichen Daten-Set dreimal mehr Frauen als Männer enthalten sind, wird dieses Verhältnis durch die separate Stichprobenziehung aus jeder der Gruppen beibehalten. Es können auch mehrere Schichtungsfelder angegeben werden (beispielsweise die Stichprobenziehung von Produktlinien innerhalb von Regionen oder umgekehrt).

Anmerkung: Wenn Sie die Schichtung anhand eines Felds mit fehlenden Werten (null oder systemdefiniert fehlende Werte, leere Zeichenketten, leere Bereiche und Leerstellen oder benutzerdefiniert fehlende Werte) vornehmen, können Sie keine benutzerdefinierten Stichprobengrößen für die Schichten angeben. Wenn Sie bei der Schichtung nach einem Feld mit fehlenden Werten oder Leerwerten benutzerdefinierte Stichprobengrößen verwenden möchten, müssen Sie die fehlenden Werte weiter oben im Stream ergänzen.

Eingabegewichtung verwenden. Dient zur Angabe eines Felds, das zur Gewichtung von Datensätzen vor der Stichprobenziehung verwendet werden soll. Wenn beispielsweise das Gewichtungsfeld Werte im Bereich von 1 bis 5 aufweist, werden die Datensätze mit dem Gewicht 5 mit der 5-fachen Wahrscheinlichkeit ausgewählt. Die Werte in diesem Feld werden mit den endgültigen Ausgabegewichtungen überschrieben, die vom Knoten generiert wurden (siehe folgender Absatz)

Neue Ausgabegewichtung. Gibt den Namen des Felds an, in das die endgültigen Gewichte geschrieben werden, wenn kein Feld für die Eingabegewichtung angegeben wurde. (Wenn ein Feld für die Eingabegewichtung angegeben wurde, werden seine Werte, wie oben angegeben, durch die endgültigen Gewichte ersetzt und es wird kein separates Feld für die Ausgabegewichtungen erstellt.) Die Werte für die Ausgabegewichtungen geben die Anzahl der Datensätze an, die durch die einzelnen Stichprobendatensätze in den ursprünglichen Daten repräsentiert werden. Die Summe der Gewichtswerte ergibt eine Schätzung des Stichprobenumfangs. Wenn beispielsweise eine 10 % umfassende Zufallsstichprobe gezogen wurde, ist das Ausgabegewicht für alle Datensätze 10, was anzeigt, dass jeder Datensatz in der Stichprobe ungefähr 10 Datensätze in den ursprünglichen Daten repräsentiert. Bei einer geschichteten oder gewichteten Stichprobe können die Werte der Ausgabegewichte je nach dem Stichprobenanteil der einzelnen Schichten variieren.

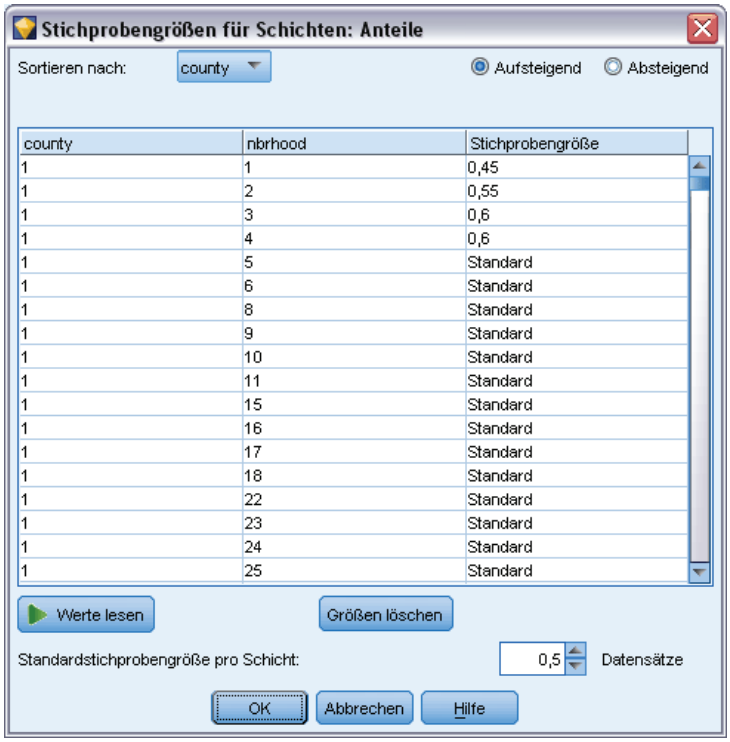
Kommentare

- Die Ziehung von Klumpenstichproben ist sinnvoll, wenn Sie keine vollständige Auflistung der Grundgesamtheit, aus der die Stichprobe gezogen werden soll, beschaffen können, aber vollständige Listen für bestimmte Gruppen bzw. Klumpen zugänglich sind. Außerdem wird sie verwendet, wenn eine Zufallsstichprobe zu einer Liste mit Testsubjekten führen würde, mit denen eine Kontaktaufnahme nicht praktikabel wäre. Es wäre beispielsweise einfacher, alle Bauern in einem bestimmten Landkreis bzw. County zu besuchen, als eine Auswahl von Bauern, die über alle Landkreise oder Counties des Staates verstreut sind.
- Sie können auch sowohl ein Klumpen- als auch ein Schichtungsfeld angeben, um Klumpen innerhalb der einzelnen Schichten unabhängig voneinander zu ziehen. Sie können beispielsweise eine Stichprobe der Eigentumswerte ziehen, die nach County geschichtet ist, und dann innerhalb der einzelnen Counties Klumpen bilden, die auf den Gemeinden beruhen. Dadurch wird sichergestellt, dass aus jedem County eine unabhängige Stichprobe der Gemeinden gezogen wird. Einige Gemeinden werden aufgenommen, andere dagegen nicht, aber bei jeder aufgenommenen Gemeinde werden alle Eigenschaften innerhalb der Gemeinde aufgenommen.
- Um eine Zufallsstichprobe der Einheiten aus den einzelnen Clustern zu ziehen, können Sie zwei Stichprobenknoten miteinander verknüpfen. So können Sie beispielsweise zuerst eine nach County geschichtete Stichprobe der Gemeinden ziehen. Anschließend können Sie einen zweiten Stichprobenknoten anfügen und *town* (Gemeinde) als Schichtungsfeld auswählen. Dadurch können Sie aus jeder Gemeinde einen Anteil an Datensätzen als Stichprobe ziehen.
- In Fällen, in denen für eine eindeutige Identifizierung der Klumpen eine Kombination von Feldern erforderlich ist, können Sie mithilfe eines Ableitungsknotens ein neues Feld generieren. Beispiel: Wenn mehrere Läden dasselbe Nummerierungssystem für Transaktionen verwenden, könnten Sie ein neues Feld ableiten, das die Geschäfts- und die Transaktions-ID miteinander verknüpft.

Stichprobengrößen für Schichten

Beim Ziehen einer geschichteten Stichprobe besteht die Standardoption darin, aus jeder Schicht denselben Anteil an Datensätzen oder Klumpen zu ziehen. Wenn eine Gruppe einer anderen beispielsweise zahlenmäßig um den Faktor 3 überlegen ist, soll dieses Verhältnis normalerweise in der Stichprobe erhalten bleiben. Andernfalls können Sie die Stichprobengröße für jede Schicht separat angeben.

Abbildung 3-6
Angabe der Stichprobengröße für Schichten



Im Dialogfeld “Stichprobengrößen für Schichten” werden die einzelnen Werte des Schichtungsfelds aufgeführt, sodass Sie den Standardwert für die betreffende Schicht außer Kraft setzen können. Wenn mehrere Schichtungsfelder ausgewählt sind, wird jede mögliche Wertekombination aufgelistet, sodass Sie beispielsweise die Größe für die verschiedenen ethnischen Gruppen in den verschiedenen Städten oder die Größe der verschiedenen Gemeinden innerhalb der einzelnen Counties angeben können. Die Größen werden als Anteile oder als Anzahl der beobachteten Werte angegeben, je nachdem was in der aktuellen Einstellung im Stichprobenknoten festgelegt ist.

So geben Sie Stichprobengrößen für Schichten an:

- Wählen Sie im Stichprobenknoten die Option Komplex und wählen Sie mindestens ein Schichtungsfeld aus. Für weitere Informationen siehe Thema Einstellungen unter “Klumpen und Schichtung” auf S. 76.
- Wählen Sie die Option Angepasst und dann Größen angeben.

- Klicken Sie im Dialogfeld “Stichprobengrößen für Schichten” auf die Schaltfläche Werte lesen links unten, um die Anzeige auszufüllen. Ggf. müssen Sie Werte in einem weiter oben liegenden Quellen- oder Typknoten instanziieren. Für weitere Informationen siehe Thema Was ist Instanziierung? in Kapitel 4 auf S. 142.

- Klicken Sie in eine Zeile, um die Standardgröße für die betreffende Schicht zu überschreiben.

Hinweise zur Stichprobengröße

Benutzerdefinierte Stichprobengrößen können nützlich sein, wenn verschiedene Schichten eine unterschiedliche Varianz aufweisen, beispielsweise, um Stichprobengrößen proportional zur Standardabweichung zu machen. (Wenn die Fälle innerhalb der Schicht eine größere Variation aufweisen, müssen Sie mehr davon ziehen, um eine repräsentative Stichprobe zu erhalten.) Bei kleinen Schichten kann ein höherer Stichprobenanteil sinnvoll sein, um sicherzustellen, dass eine Mindestanzahl von Beobachtungen aufgenommen wird.

Hinweis: Wenn Sie die Schichtung anhand eines Felds mit fehlenden Werten (null oder systemdefiniert fehlende Werte, leere Zeichenketten, leere Bereiche und Leerstellen oder benutzerdefiniert fehlende Werte) vornehmen, können Sie keine benutzerdefinierten Stichprobengrößen für die Schichten angeben. Wenn Sie bei der Schichtung nach einem Feld mit fehlenden Werten oder Leerwerten benutzerdefinierte Stichprobengrößen verwenden möchten, müssen Sie die fehlenden Werte weiter oben im Stream ergänzen.

Balancierungsknoten

Mithilfe von Balancierungsknoten können Sie Unausgewogenheiten in den Daten-Sets korrigieren, sodass Sie den angegebenen Testkriterien entsprechen. Beispiel: Angenommen, ein Daten-Set weist nur zwei Werte auf – *niedrig* und *hoch* – und 90 % der Fälle sind *niedrig* und nur 10 % der Fälle *hoch*. Bei vielen Modellierungsverfahren gibt es Schwierigkeiten mit solchen verzerrten Daten, weil sie in der Regel nur die *niedrigen* Ergebnisse berücksichtigen und die *hohen* ignorieren, da diese seltener sind. Bei ausgewogenen Daten mit ungefähr gleich vielen Ergebnissen vom Typ *niedrig* und *hoch* haben die Modelle eine bessere Chance, Muster zu finden, die zur Unterscheidung zwischen den beiden Gruppen dienen können. In diesem Fall kann mit einem Balancierungsknoten eine Balancierungsanweisung erstellt werden, die die Anzahl der Fälle mit dem Ergebnis vom Typ *niedrig* reduziert.

Die Balancierung erfolgt durch das Duplizieren und anschließende Verwerfen von Datensätzen auf der Grundlage der von Ihnen angegebenen Bedingungen. Datensätze, für die keine Bedingung gilt, werden immer übergeben. Da dieser Vorgang auf der Duplizierung und/oder dem Verwerfen von Datensätzen beruht, kann die ursprüngliche Sequenz Ihrer Daten in den weiter unten im Stream liegenden Operationen nicht erhalten bleiben. Daher müssen Sie alle sequenzbezogenen Werte ableiten, bevor Sie einen Balancierungsknoten zum Daten-Stream hinzufügen.

Hinweis: Balancierungsknoten können automatisch aus Verteilungsdiagrammen und Histogrammen generiert werden. Beispielsweise können Sie die Daten balancieren, sodass sie in allen Kategorien eines kategorialen Felds gleiche Anteile anzeigen, wie in einem Verteilungsdiagramm gezeigt.

Beispiel. Bei der Erstellung eines RFM-Streams zur Ermittlung aktueller Kunden, die positiv auf frühere Marketingkampagnen reagiert haben, verwendet die Marketingabteilung einer Vertriebsgesellschaft einen Balancierungsknoten, um die Unterschiede zwischen den Wahr- und den Falsch-Antworten in den Daten auszugleichen.

Festlegen der Optionen für den Balancierungsknoten

Abbildung 3-7
Einstellungen im Balancierungsknoten

Anweisungen für Datensatzgewichtung. Listet die aktuellen Gewichtungsanweisungen auf. Zu jeder Anweisung gehören ein Faktor und eine Bedingung, die die Software anweist, den Anteil der Datensätze um einen angegebenen Faktor zu erhöhen, wenn die Bedingung wahr ist. Bei einem Faktor von unter 1,0 wird der Anteil der angegebenen Datensätze verringert. Beispiel: Angenommen, Sie möchten die Anzahl der Datensätze, bei denen die Behandlung mit Medikament Y erfolgt, verringern. Dann können Sie beispielsweise eine Gewichtungsanweisung mit dem Faktor 0,7 und der Bedingung Medikament = "MedikamentY" erstellen. Diese Anweisung führt dazu, dass die Anzahl der Datensätze, bei denen die Behandlung mit Medikament Y erfolgt, für alle Operationen weiter unten im Stream auf 70 % reduziert wird.

Hinweis: Balancierungsfaktoren für die Reduzierung können bis auf vier Dezimalstellen angegeben werden. Faktoren, die unter 0,0001 festgelegt werden, führen zu einem Fehler, da die Ergebnisse nicht ordnungsgemäß berechnet werden.

- **Zum Erstellen von Bedingungen** klicken Sie auf die Schaltfläche rechts neben dem Textfeld. Dadurch wird eine leere Zeile zur Eingabe neuer Bedingungen eingefügt. Um einen CLEM-Ausdruck für die Bedingung zu erstellen, klicken Sie auf die Schaltfläche “Expression Builder”.

- **Zum Löschen von Anweisungen** verwenden Sie die rote Löschschaltfläche.
- **Zum Sortieren von Anweisungen** verwenden Sie die oben und nach unten weisenden Pfeil-Schaltflächen.

Balancierung nur für Trainingsdaten durchführen. Wenn im Stream ein Partitionsfeld vorhanden ist, wird bei Auswahl dieser Option die Balancierung ausschließlich in der Trainingspartition durchgeführt. Dies kann insbesondere bei der Generierung von Scores für die korrigierte Neigung nützlich sein, da diese eine nicht balancierte Test- bzw. Validierungspartition erfordern. Wenn im Stream kein Partitionsfeld vorhanden ist (oder wenn mehrere Partitionsfelder angegeben sind), wird diese Option ignoriert und alle Daten werden balanciert.

Aggregatknoten.

Aggregation ist eine Vorbereitungsaufgabe, die häufig zur Reduzierung der Größe eines Daten-Sets verwendet wird. Bevor Sie mit der Aggregation fortfahren, sollten Sie sich die Zeit nehmen, die Daten zu bereinigen. Achten Sie dabei insbesondere auf fehlende Daten. Bei der Aggregierung können potenziell nützliche Informationen zu fehlenden Werten verloren gehen.

Mit einem Aggregatknoten können Sie eine Sequenz von Eingabedatensätzen mit aggregierten Übersichts-Ausgabedatensätzen ersetzen. Beispielsweise könnten Sie folgendes Set von Eingabe-Verkaufsdatensätzen besitzen:

Alter	Geschlecht	Region	Zweigstelle	Verkäufe
23	M	S	8	4
45	M	S	16	4
37	M	S	8	5
30	M	S	5	7
44	M	X	4	9
25	M	X	2	11
29	F	S	16	6
41	F	X	4	8
23	F	X	6	2
45	F	X	4	5
33	F	X	6	10

Sie können diese Datensätze mit *Geschlecht* und *Region* als Schlüsselfelder aggregieren. Legen Sie anschließend fest, dass *Alter* mit dem Modus Mittelwert und *Umsatz* mit dem Modus Summe aggregiert werden soll. Wenn Sie anschließend im Dialogfeld des Aggregatknotens die Option Datensatzanzahl einschließen in Feld auswählen, lautet die aggregierte Ausgabe wie folgt:

Alter (Mittelwert)	Geschlecht	Region	Verkäufe (Summe)	Datensatzanzahl
35.5	F	X	25	4
29	F	S	6	1
34.5	M	X	20	2
33.75	M	S	20	4

Daraus können Sie beispielsweise entnehmen, dass das Durchschnittsalter der vier weiblichen Angehörigen des Vertriebspersonals in der Region “Nord” 35,5 Jahre beträgt und dass sie insgesamt 25 Einheiten verkauft haben.

Hinweis: Felder wie *Zweigstelle* werden automatisch verworfen, wenn kein Aggregatmodus angegeben wurde.

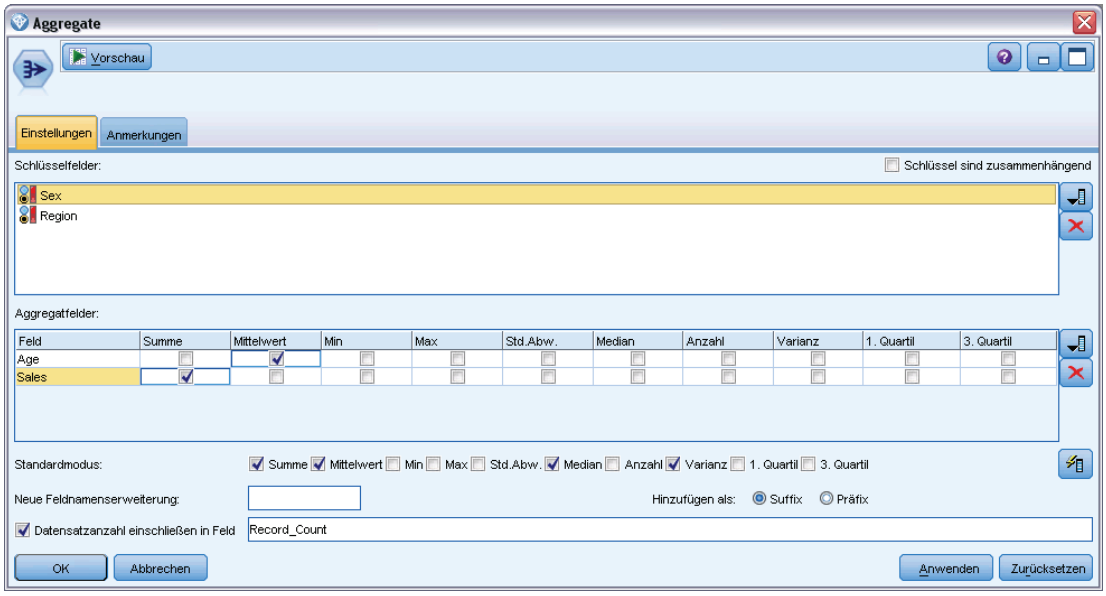
Festlegen der Optionen für den Aggregatknoten

Im Aggregatknoten geben Sie Folgendes an:

- Ein oder mehrere Schlüsselfelder zur Verwendung als Kategorien für die Aggregation
- Ein oder mehrere Aggregatfelder, für die die Aggregatwerte berechnet werden sollen
- Ein oder mehrere Aggregatmodi (Aggregattypen), die für die einzelnen Aggregatfelder ausgegeben werden sollen

Sie können auch die Standard-Aggregationsmodi angeben, die für neu hinzugefügte Felder verwendet werden sollen.

Abbildung 3-8
Aggregatknoten – Dialogfeld



Schlüsselfelder. Listet Felder auf, die als Kategorien für die Aggregation verwendet werden können. Sowohl stetige (numerische) als auch kategoriale Felder können als Schlüssel verwendet werden. Bei Auswahl von mehreren Schlüsselfeldern werden die Werte kombiniert und ergeben einen Schlüsselwert für die Aggregation von Datensätzen. Für jedes eindeutige Schlüsselfeld wird jeweils ein (1) aggregierter Datensatz generiert. Bei den Schlüsselfeldern *Geschlecht* und *Region* beispielsweise erhält jede eindeutige Kombination von *M* und *W* mit den Regionen *N* und *S* (vier eindeutige Kombinationen) einen aggregierten Datensatz. Verwenden Sie zum Hinzufügen eines Schlüsselfelds die Feldauswahl-Schaltfläche auf der rechten Seite des Fensters.

Schlüssel sind zusammenhängend. Wählen Sie diese Option aus, wenn Sie wissen, dass alle Datensätze mit denselben Schlüsselwerten in der Eingabe als zusammenhängende Gruppe vorliegen (z. B. wenn die Eingabe nach Schlüsselfeldern sortiert ist) Dadurch lässt sich eventuell die Leistungsfähigkeit verbessern.

Aggregatfelder. Listet die Felder auf, für die Werte aggregiert werden, sowie die ausgewählten Aggregationsmodi. Mit der Feldauswahl-Schaltfläche auf der rechten Seite können Sie Felder zu dieser Liste hinzufügen. Die folgenden Aggregationsmodi stehen zur Verfügung:

Hinweis: Einige Modi gelten nicht für nichtnumerische Felder (z. B. Summe für Datums-/Zeitfelder). Modi, die bei einem ausgewählten Aggregatfeld nicht verwendet werden können, sind deaktiviert.

- **Summe.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination summierte Werte auszugeben. Die Summe der Werte über alle Fälle mit nicht fehlenden Werten.
- **Mittelwert.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Mittelwerte auszugeben. Der Mittelwert ist ein Lagemaß, bei dem es sich um den arithmetischen Durchschnitt (die Summe dividiert durch die Anzahl der Fälle) handelt.
- **Min.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination Mindestwerte auszugeben.
- **Max.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination Höchstwerte auszugeben.
- **Std.Abw.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Standardabweichung auszugeben. Die Standardabweichung ist ein Maß für die Streuung um den Mittelwert, definiert als positive Wurzel der Varianzmessung.
- **Median.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Medianwerte auszugeben. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann). Auch als 50. Perzentil bzw. 2. Quartil bezeichnet.
- **Häufigkeiten.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Anzahl der Werte auszugeben, bei denen es sich nicht um Nullwerte handelt.
- **Varianz.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Varianzwerte auszugeben. Die Varianz ist ein Maß der Streuung um den Mittelwert. Sie ist gleich dem Quotienten aus der Summe der quadrierten Abweichung vom Mittelwert und der um 1 verringerten Fallanzahl.
- **1. Quartil.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Werte für das 1. Quartil (25. Perzentil) auszugeben.
- **3. Quartil.** Wählen Sie diese Option aus, um für jede Schlüsselfeldkombination die Werte für das 3. Quartil (75. Perzentil) auszugeben.

Standardmodus. Geben Sie den Standard-Aggregationsmodus an, der für neu hinzugefügte Felder verwendet werden soll. Wenn Sie häufig dieselbe Aggregation verwenden, wählen Sie hier einen oder mehrere Knoten aus und verwenden Sie die Schaltfläche “Auf alle anwenden” auf der rechten Seiten, um die ausgewählten Modi auf alle oben aufgeführten Felder zu übernehmen.

Neue Feldnamenerweiterung. Wählen Sie diese Option aus, um ein Suffix oder ein Präfix, beispielsweise “1” oder “neu” zu den duplizierten aggregierten Feldern hinzuzufügen. So führt eine Aggregation mit Mindestwerten beim Feld *Alter* zu einem Feld mit der Bezeichnung *Alter_Min_1*, wenn Sie Suffixoption ausgewählt und “1” als Erweiterung angegeben haben. *Hinweis*: Aggregationserweiterungen wie *_Min* oder *Max_* werden automatisch zum neuen Feld hinzugefügt und geben den Typ der durchgeführten Aggregation an. Wählen Sie Suffix bzw. Präfix aus, um die bevorzugte Erweiterungsart anzugeben.

Datensatzanzahl einschließen in Feld. Wählen Sie diese Option aus, um standardmäßig ein zusätzliches Feld mit der Bezeichnung *Datensatzanzahl* in jeden Ausgabedatensatz einzufügen. Dieses Feld gibt an, wie viele Eingabedatensätze aus den einzelnen Aggregatdatensätzen aggregiert wurden. Im Bearbeitungsfeld können Sie einen benutzerdefinierten Namen für dieses Feld angeben.

Hinweis: Systemdefinierte Nullwerte werden bei der Berechnung von Aggregaten ausgeschlossen, in der Datensatzanzahl sind sie jedoch enthalten. Leere Werte dagegen sind sowohl in der Aggregation als auch in der Datensatzanzahl enthalten. Um leere Werte auszuschließen, ersetzen Sie mithilfe eines Füllerknotens Leerstellen durch Nullwerte. Außerdem können Sie Leerstellen mithilfe eines Auswahlknotens entfernen.

Leistung

Aggregationsvorgänge können ggf. durch Parallelverarbeitung beschleunigt werden.

RFM-Aggregatknoten

Mit dem Knoten “RFM-Aggregat” (Recency-, Frequency-, Monetary-Aggregat) können Sie Daten über die früheren Transaktionen von Kunden verwenden, alle nicht benötigten Daten entfernen und alle verbliebenen Transaktionsdaten zu einer einzigen Zeile zusammenfassen (mit der eindeutigen Kunden-ID als Schlüssel), die angibt, wann der betreffende Kunde zuletzt mit Ihnen in Geschäftskontakt stand (Recency, Aktualität), wie viele Transaktionen er vorgenommen hat (Frequency, Häufigkeit) und wie hoch der Gesamtwert dieser Transaktionen ist (Monetary, Geldwert).

Bevor Sie mit der Aggregation fortfahren, sollten Sie sich die Zeit nehmen, die Daten zu bereinigen. Achten Sie dabei insbesondere auf etwaige fehlende Daten.

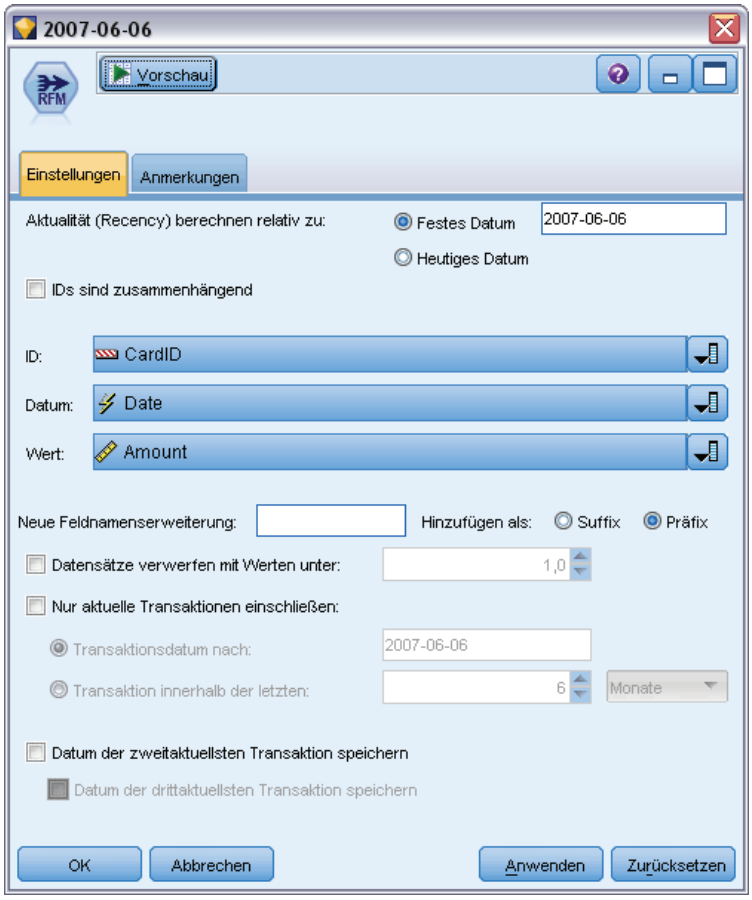
Sobald Sie die Daten mithilfe des RFM-Aggregatknotens identifiziert und transformiert haben, können Sie mithilfe eines RFM-Analyseknotens weitere Analysen durchführen. Für weitere Informationen siehe Thema Knoten “RFM-Analyse” in Kapitel 4 auf S. 204.

Beachten Sie: Nach dem Durchlaufen des RFM-Aggregatknotens enthält die Datendatei keinerlei Zielwerte; daher können Sie sie erst dann als Eingabe für weitere Prognoseanalysen mit anderen Modellierungsknoten, wie beispielsweise C5.0 oder CHAID verwenden, nachdem Sie sie mit anderen Kundendaten zusammengeführt haben (beispielsweise durch Abgleich der Kunden-IDs). Für weitere Informationen siehe Thema Zusammenführungsknoten (“Mergen”) auf S. 89.

Die Knoten “RFM-Aggregat” und “RFM-Analyse” in IBM® SPSS® Modeler sind für die Verwendung einer unabhängigen Klassierung eingerichtet. Damit werden also Daten für jedes der Maße Aktualität, Häufigkeit und Geldwert in Ränge eingeteilt und klassiert, ohne Berücksichtigung ihrer Werte oder der beiden anderen Maße.

Festlegen der Optionen für den RFM-Aggregatknoten

Abbildung 3-9
Einstellungen für RFM-Aggregat



Aktualität (Recency) berechnen relativ zu. Dient zur Angabe des Datums, ausgehend von dem die Aktualität der Transaktionen berechnet werden soll. Hierfür können Sie entweder unter Festes Datum ein Datum eingeben oder Heutiges Datum auswählen, wobei das aktuelle Datum laut Systemeinstellungen verwendet wird. Der Wert Heutiges Datum wird standardmäßig eingegeben und automatisch aktualisiert, wenn der Knoten ausgeführt wird.

IDs sind zusammenhängend. Wenn Ihre Daten vorsortiert sind, sodass alle Datensätze mit derselben ID zusammen im Daten-Stream erscheinen, wählen Sie diese Option, um die Verarbeitung zu beschleunigen. Wenn Ihre Daten nicht vorsortiert sind (oder Sie nicht sicher sind), lassen Sie diese Option deaktiviert. Die Daten werden dann vom Knoten automatisch sortiert.

ID. Dient zur Auswahl des für die Identifizierung des Kunden und seiner Transaktionen zu verwendenden Felds. Die zur Auswahl zur Verfügung stehenden Felder können Sie mit der Feldauswahl-Schaltfläche auf der rechten Seite anzeigen.

Datum. Dient zur Auswahl des Datumsfelds, das für die Berechnung der Aktualität (Recency) verwendet werden soll. Die zur Auswahl zur Verfügung stehenden Felder können Sie mit der Feldauswahl-Schaltfläche auf der rechten Seite anzeigen.

Beachten Sie, dass hierfür ein Feld mit dem Speichertyp "Datum" oder "Zeitstempel" im entsprechenden Format zur Verwendung als Eingabe erforderlich ist. Wenn Ihnen beispielsweise ein Zeichenkettenfeld mit Werten wie *Jan 2000*, *Feb 2000* usw. vorliegt, können Sie dieses mithilfe eines Füllerknotens und der Funktion to_date() in ein Datumsfeld konvertieren. Für weitere Informationen siehe Thema Speichertypkonvertierung mithilfe des Füllerknotens in Kapitel 4 auf S. 181.

Wert. Dient zur Auswahl des Felds, das für die Berechnung des Gesamtwerts der Transaktionen des Kunden verwendet werden soll. Die zur Auswahl zur Verfügung stehenden Felder können Sie mit der Feldauswahl-Schaltfläche auf der rechten Seite anzeigen. *Hinweis*: Dieser Wert muss ein numerischer Wert sein.

Neue Feldnamenerweiterung. Wählen Sie diese Option aus, wenn Sie die neu erstellten Felder für Aktualität, Häufigkeit und Geldwert durch ein Suffix oder Präfix, wie beispielsweise "12_Monate", ergänzen möchten. Wählen Sie Suffix bzw. Präfix aus, um die bevorzugte Erweiterungsart anzugeben. Dies kann beispielsweise bei der Untersuchung mehrerer Zeitperioden nützlich sein.

Datensätze verwerfen mit Werten unter. Falls erforderlich, können Sie hier einen Mindestwert für die bei der Berechnung der RFM-Gesamtwerte verwendeten Transaktionsdetails angeben. Die für den Wert geltenden Einheiten beziehen sich auf das ausgewählte Feld Wert.

Nur aktuelle Transaktionen einschließen. Bei der Analyse großer Datenbanken können Sie angeben, dass nur die aktuellsten Datensätze verwendet werden sollen. Sie können auswählen, ob die nach einem bestimmten Datum oder innerhalb eines bestimmten Zeitraums protokollierten Daten verwendet werden sollen:

- **Transaktionsdatum nach.** Dient zur Angabe des Transaktionsdatums, nach dem die Datensätze in die Analyse aufgenommen werden sollen.
- **Transaktion innerhalb der letzten.** Hier können Sie anhand von Anzahl und Typ der Zeiträume (Tage, Wochen, Monate oder Jahre) angeben, wie weit ausgehend von Aktualität (Recency) berechnen relativ zu die in die Analyse aufzunehmenden Datensätze zurückliegen dürfen.

Datum der zweitaktuellsten Transaktion speichern. Aktivieren Sie dieses Kontrollkästchen, wenn Sie das Datum der zweitaktuellsten Transaktion für die einzelnen Kunden ermitteln möchten. Zusätzlich können Sie auch das Kontrollkästchen Datum der drittaktuellsten Transaktion speichern aktivieren. Dies kann Ihnen beispielsweise dabei helfen, Kunden zu identifizieren, die möglicherweise vor längerer Zeit zahlreiche Transaktionen getätigt haben, aber nur eine aktuelle Transaktion.

Sortierknoten

Mit Sortierknoten können Sie Datensätze anhand der Werte eines oder mehrerer Felder in aufsteigender oder absteigender Reihenfolge kopieren. Sortierknoten werden beispielsweise häufig verwendet, um Datensätze mit den häufigsten Datenwerten anzuzeigen und auszuwählen. Üblicherweise werden die Daten zuerst mit dem Aggregatknoten aggregiert und die aggregierten Daten anschließend mit dem Sortierknoten in absteigender Reihenfolge nach Datensatzanzahl sortiert. Durch die Anzeige dieser Ergebnisse in einer Tabelle können Sie die Daten untersuchen und Entscheidungen treffen. Beispielsweise könnten Sie die Datensätze der 10 besten Kunden auswählen.

Abbildung 3-10
Sortierknoten – Dialogfeld

Sortieren nach. Alle Felder, die zur Verwendung als Sortierschlüssel ausgewählt wurden, werden in einer Tabelle angezeigt. Schlüsselfelder sind am besten für die Sortierung geeignet, wenn sie numerisch sind.

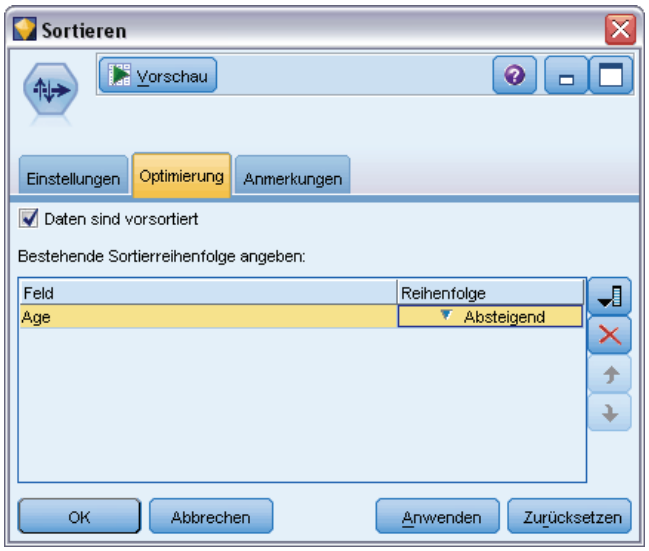
- **Zum Hinzufügen von Feldern** zu dieser Liste verwenden Sie die Feldauswahl-Schaltfläche auf der rechten Seite.
- **Zur Auswahl einer Reihenfolge** klicken Sie auf den Pfeil Aufsteigend oder Absteigend in der Spalte *Reihenfolge* der Tabelle.
- **Zum Löschen von Feldern** verwenden Sie die rote Löschschaltfläche.
- **Zum Sortieren von Anweisungen** verwenden Sie die oben und nach unten weisenden Pfeil-Schaltflächen.

Standard-Sortierreihenfolge. Als Standard-Sortierreihenfolge, die für das Hinzufügen neuer Felder verwendet wird, können Sie entweder Aufsteigend oder Absteigend auswählen.

Optimierungseinstellungen für das Sortieren

Wenn Sie mit Daten arbeiten, die bereits nach bestimmten Schlüsselfeldern sortiert sind, können Sie diese Sortierungsfelder angeben, sodass die restlichen Daten effizienter sortiert werden können. Beispiel: Die Daten sollen nach *Alter* (absteigend) und *Medikament* (aufsteigend) sortiert werden; Sie wissen jedoch, dass die Daten bereits nach *Alter* (absteigend) sortiert sind.

Abbildung 3-11
Optimierungseinstellungen



Daten sind vorsortiert. Gibt an, ob die Daten bereits nach einem oder mehreren Feldern sortiert sind.

Bestehende Sortierreihenfolge angeben. Gibt die Felder an, die bereits sortiert sind. Über das Dialogfeld "Felder auswählen" nehmen Sie Felder in die Liste auf. Geben Sie in der Spalte *Reihenfolge* jeweils an, ob die Felder in aufsteigender oder absteigender Reihenfolge sortiert sind. Wenn Sie mehrere Felder angeben, achten Sie darauf, die Felder in der richtigen Sortierreihenfolge aufzuführen. Mit den Pfeilen rechts neben der Liste ordnen Sie die Felder in der richtigen Reihenfolge an. Wenn Sie die vorhandene Sortierreihenfolge nicht richtig angeben, tritt beim Ausführen des Streams ein Fehler auf. In der Fehlermeldung wird dabei die Nummer des Datensatzes angezeigt, bei dem die Sortierung nicht mit Ihren Angaben übereinstimmt.

Hinweis: Die Sortierung kann ggf. durch Parallelverarbeitung beschleunigt werden.

Zusammenführungsknoten ("Mergen")

Die Funktion von Zusammenführungsknoten ("Mergen") besteht darin, aus mehreren Eingabedatensätzen einen einzelnen Ausgabedatensatz mit allen oder einigen der Eingabefelder zu erstellen. Dies ist ein nützlicher Vorgang, wenn Daten aus verschiedenen Quellen,

wie beispielsweise interne Kundendaten und käuflich erworbene demografische Daten, zusammengeführt werden sollen. Sie können Daten auf folgende Weisen zusammenführen:

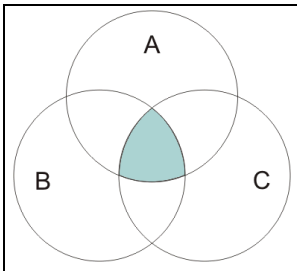
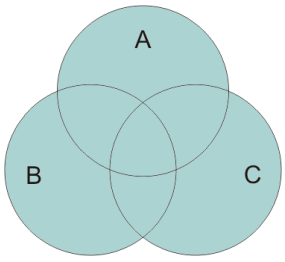
- **Beim Zusammenführen (Mergen) nach Reihenfolge** werden die entsprechenden Datensätze aus allen Quellen in der Reihenfolge der Eingabe miteinander verkettet, bis die kleinste Datenquelle erschöpft ist. Vor der Verwendung dieser Option müssen die Daten unbedingt mit einem Sortierknoten sortiert worden sein.
- **Das Zusammenführen (Mergen) mit einem Schlüsselfeld**, wie beispielsweise *Kunden-ID*, dient zur Angabe, wie die Datensätze aus einer Datenquelle mit Datensätzen aus den anderen Quellen abgeglichen werden können. Mehrere Arten von Joins sind möglich, beispielsweise Inner Join, Full Outer Join, Partieller Outer Join und Anti-Join. Für weitere Informationen siehe Thema Join-Typen auf S. 90.
- **Beim Zusammenführen (Mergen) nach Bedingung** können Sie eine Bedingung angegeben, die erfüllt sein muss, damit das Zusammenführen stattfindet. Sie können die Bedingung direkt im Knoten angeben, oder die Bedingung mithilfe des Expression Builder erstellen.

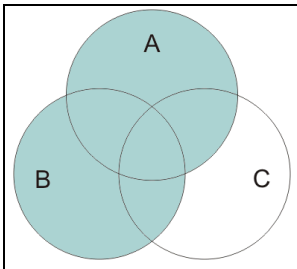
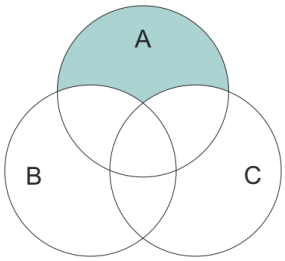
Join-Typen

Bei der Verwendung eines Schlüsselfelds zum Zusammenführen von Daten sollten Sie vorher überlegen, welche Datensätze ausgeschlossen und welche eingeschlossen werden sollen. Es gibt eine Vielzahl von Joins, die unten im Detail erörtert werden.

Die beiden Join-Grundtypen heißen “Inner Join” und “Outer Join”. Diese Methoden werden häufig zur Zusammenführung von Tabellen aus verwandten Daten-Sets auf der Grundlage gemeinsamer Werte eines Schlüsselfelds, beispielsweise *Kunden-ID*, verwendet. Inner Joins ergeben eine saubere Zusammenführung und ein Ausgabe-Daten-Set, das nur vollständige Datensätze enthält. Outer Joins beinhalten ebenfalls vollständige Datensätze aus den zusammengeführten Daten, doch sie ermöglichen auch die Aufnahme eindeutiger Daten aus einer oder mehreren Eingabetabellen.

Die zulässigen Join-Typen sind weiter unten detaillierter beschrieben.

	Ein **Inner Join** enthält nur Datensätze, bei denen ein Wert für das Schlüsselfeld bei allen Eingabetabellen gleich ist. Nicht übereinstimmende Datensätze werden nicht in das Ausgabe-Daten-Set aufgenommen.
	Bei einem **Full Outer Join** werden alle Datensätze (übereinstimmend und nicht übereinstimmend) aus den Eingabetabellen eingeschlossen. Linke und rechte Outer Joins werden als partielle Outer Joins bezeichnet und werden im Folgenden beschrieben.

	Ein **partieller Outer Join** enthält alle Datensätze, deren Übereinstimmung anhand des Schlüsselfelds abgeglichen wurde, sowie nicht übereinstimmende Datensätze aus den angegebenen Tabellen. (Oder anders gesagt: Alle Datensätze aus bestimmten Tabellen und nur passende Datensätze aus anderen Tabellen.) Tabellen (wie beispielsweise A und B in der Abbildung) können mithilfe der Schaltfläche “Auswahl” auf der Registerkarte “Verbinden” ausgewählt werden. Partielle Joins werden auch linke bzw. rechte Outer Joins genannt, wenn nur zwei Tabellen zusammengeführt werden. Da IBM® SPSS® Modeler die Zusammenführung von mehr als zwei Tabellen erlaubt, wird dieser Vorgang hier als partieller Outer Join bezeichnet.
	Bei **Anti-Join** werden nur nicht übereinstimmende Datensätze für die erste Eingabetabelle (Tabelle A in der Abbildung) aufgenommen. Bei diesem Join-Typ handelt es sich um das Gegenteil eines Inner Join. Es werden keine vollständigen Datensätze in das Ausgabe-Daten-Set aufgenommen.

Wenn beispielsweise Informationen über Bauernhöfe in einem Daten-Set vorliegen und Versicherungsansprüche zu Bauernhöfen in einem zweiten Daten-Set, dann können Sie die Datensätze aus der ersten Quelle mithilfe der Zusammenführungsoptionen mit den Datensätzen aus der zweiten Quelle abgleichen.

Um festzustellen, ob ein Kunde in diesem Bauernhof-Beispiel einen Versicherungsanspruch angemeldet hat, rufen Sie mit der Option “Inner Join” eine Liste mit allen IDs ab, die in beiden Daten-Sets vorkommen.

Abbildung 3-12
Beispielausgabe für eine Zusammenführung mit Inner Join

	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...

Bei einem Full Outer Join werden alle Datensätze (übereinstimmend und nicht übereinstimmend) aus den Eingabetabellen eingeschlossen. Bei unvollständigen Werten wird der systemdefiniert fehlende Wert ($null$) verwendet.

Abbildung 3-13
Beispielausgabe für eine Zusammenführung mit Full Outer Join

	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

Ein partieller Outer Join enthält alle Datensätze, deren Übereinstimmung anhand des Schlüsselfelds abgeglichen wurde, sowie nicht übereinstimmende Datensätze aus den angegebenen Tabellen. Die Tabelle zeigt alle Datensätze, die mit dem ID-Feld übereinstimmen, sowie alle Datensätze, die mit dem ersten Daten-Set übereinstimmen.

Abbildung 3-14
Beispielausgabe für eine Zusammenführung mit partiellem Outer Join

	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

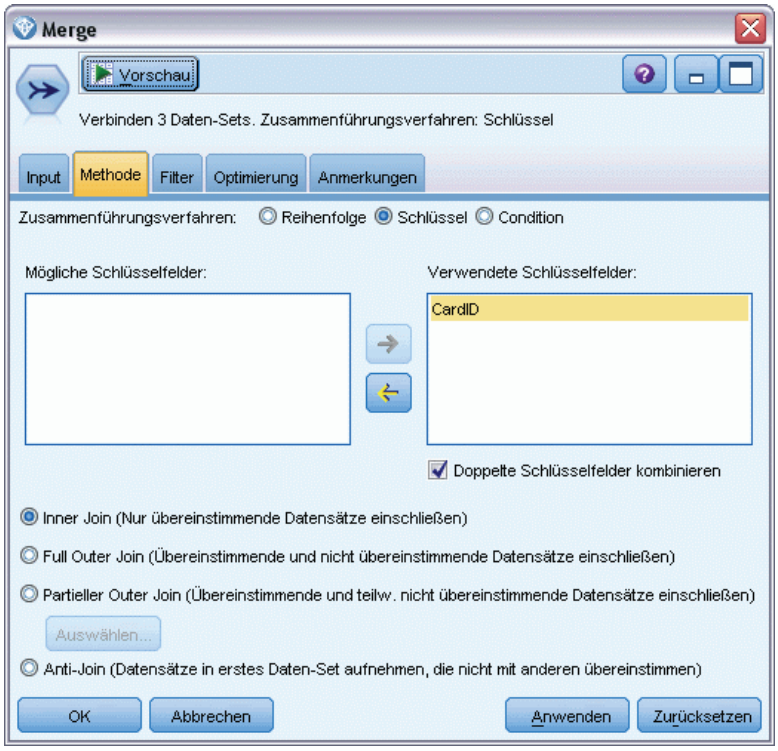
Bei Anti-Join gibt die Tabelle nur nicht übereinstimmende Datensätze für die erste Eingabetabelle aus.

Abbildung 3-15
Beispielausgabe für eine Zusammenführung mit Anti-Join

	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

Angeben eines Zusammenführungsverfahrens und von Schlüsseln

Abbildung 3-16
Verwenden der Registerkarte "Merge" zur Festlegung der Optionen für das Zusammenführungsverfahren



Zusammenführungsverfahren. Wählen Sie entweder Reihenfolge oder Schlüssel aus, um die Methode für die Zusammenführung von Datensätzen anzugeben. Durch die Auswahl von Schlüssel wird der untere Teil des Dialogfelds aktiviert.

- **Reihenfolge.** Führt Datensätze nach der Reihenfolge zusammen. Beispielsweise wird der *n*-te Datensatz aus jeder Eingabe zusammengeführt, um den *n*-ten Ausgabedatensatz zu erstellen. Wenn für einen Datensatz kein übereinstimmender Eingabedatensatz mehr vorhanden ist, werden keine weiteren Ausgabedatensätze erstellt. Die Anzahl der erstellten Datensätze ist also gleich der Anzahl der Datensätze im kleinsten Daten-Set.
- **Erläuterungen** Verwendet ein Schlüsselfeld wie *Transaktions-ID*, um Datensätze mit demselben Wert im Schlüsselfeld zusammenzuführen. Dies entspricht einem Datenbank-"Equi-Join". Wenn ein Schlüsselwert mehrmals vorkommt, werden alle möglichen Kombinationen ausgegeben. Beispiel: Wenn Datensätze mit demselben Schlüsselfeldwert *A* verschiedene Werte für *B*, *C* und *D* in anderen Feldern enthalten, erstellen die zusammengeführten Felder einen separaten Datensatz für die einzelnen Kombinationen von *A* mit Wert *B*, *A* mit Wert *C* und *A* mit Wert *D*.

  Hinweis: Nullwerte werden bei der Zusammenführung nach Schlüssel nicht als identisch betrachtet und werden nicht zusammengeführt.
- **Bedingung.** Verwenden Sie diese Option, um eine Bedingung für die Zusammenführung anzugeben. Für weitere Informationen siehe Thema Angeben von Bedingungen für das Zusammenführen auf S. 94.

Mögliche Felder. Listet nur die Felder mit identischen Namen in allen Eingabedatenquellen auf. Wählen Sie ein Feld aus dieser Liste aus und fügen Sie es mithilfe der Pfeilschaltflächen als Schlüsselfeld für die Zusammenführung von Datensätzen hinzu. Es können mehrere Schlüsselfelder verwendet werden. Sie können nicht übereinstimmende Eingabefelder mittels eines Filterknotens oder über die Registerkarte "Filter" eines Quellenknotens umbenennen.

Verwendete Schlüsselfelder. Listet alle Felder auf, die für die Zusammenführung der Datensätze aus allen Eingabedatenquellen auf der Grundlage der Schlüsselfeldwerte verwendet werden. Um einen Schlüssel aus der Liste zu entfernen, wählen Sie ihn aus und verschieben Sie ihn mithilfe der Pfeilschaltfläche zurück in die Liste "Mögliche Schlüsselfelder". Bei Auswahl mehrerer Schlüssel wird die unten stehende Option aktiviert.

Doppelte Schlüsselfelder kombinieren. Wenn oben mehrere Felder ausgewählt wurden, gewährleistet diese Option, dass nur ein einziges Ausgabefeld dieses Namens vorhanden ist. Die aktivierte Option ist standardmäßig aktiviert, außer wenn Streams aus früheren Versionen von IBM® SPSS® Modeler importiert wurden. Wenn diese Option deaktiviert ist, müssen doppelte Schlüsselfelder mithilfe der Registerkarte "Filter" im Dialogfeld des Zusammenführungsknotens ("Mergen") umbenannt oder ausgeschlossen werden.

Nur übereinstimmende Datensätze einschließen (Inner Join). Wählen Sie diese Option aus, um nur vollständige Datensätze zusammenzuführen.

Übereinstimmende und nicht übereinstimmende Datensätze einschließen (Full Outer Join) Wählen Sie diese Option aus, um einen Full Outer Join durchzuführen. Wenn also Werte für das Schlüsselfeld nicht in allen Eingabetabellen vorhanden sind, werden die unvollständigen Datensätze dennoch beibehalten. Der nicht definierte Wert ($null$) wird zum Schlüsselfeld hinzugefügt und in den Ausgabedatensatz aufgenommen.

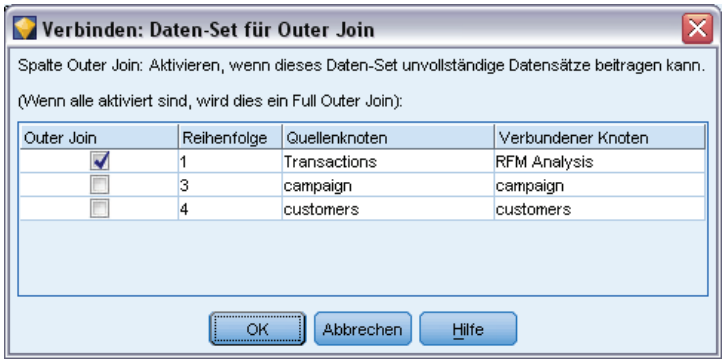
Partieller Outer Join (Übereinstimmende und teilweise nicht übereinstimmende Datensätze einschließen) Mit dieser Option können Sie einen partiellen Outer Join der Tabellen durchführen, die Sie in einem Unterdialogfeld ausgewählt haben. Klicken Sie auf Auswählen, um die Tabellen anzugeben, für die die unvollständigen Datensätze in der Zusammenführung beibehalten werden.

Anti-Join (Datensätze in erstes Daten-Set aufnehmen, die nicht mit anderen übereinstimmen) Wählen Sie diese Option aus, um eine Art "Anti-Join" durchzuführen, bei dem nur nicht übereinstimmende Datensätze aus dem ersten Daten-Set an die weiter unten im Stream liegenden Knoten übergeben werden. Mithilfe der Pfeile auf der Registerkarte "Eingaben" können Sie die Reihenfolge der Eingabe-Daten-Sets angeben. Bei diesem Join-Typ werden keine vollständigen Datensätze in das Ausgabe-Daten-Set eingeschlossen. Für weitere Informationen siehe Thema Join-Typen auf S. 90.

Auswählen von Daten für partielle Joins

Für einen partiellen Outer Join müssen Sie die Tabelle(n) auswählen, für die unvollständige Datensätze beibehalten werden sollen. Beispielsweise könnten Sie alle Datensätze aus einer Kundentabelle beibehalten, jedoch nur übereinstimmende Datensätze aus der Tabelle "Hypothekendarlehen".

Abbildung 3-17
Auswählen von Daten für den partiellen Outer Join

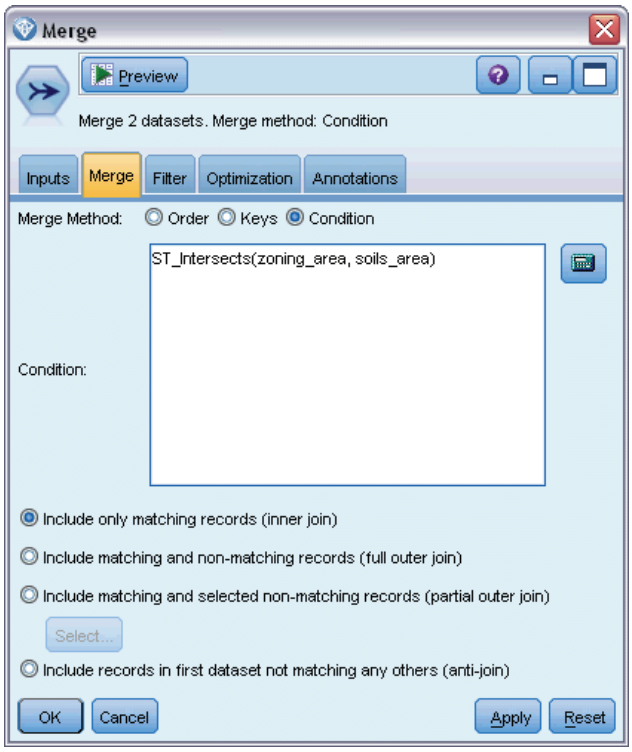


Spalte "Outer Join". Wählen Sie in der Spalte *Outer Join* die Daten-Sets aus, die vollständig aufgenommen werden sollen. Bei einem partiellen Join werden überlappende Datensätze sowie unvollständige Datensätze für die hier ausgewählten Daten-Sets beibehalten. Für weitere Informationen siehe Thema Join-Typen auf S. 90.

Angeben von Bedingungen für das Zusammenführen

Wenn Sie das Zusammenführungsverfahren auf Bedingung setzen, können Sie eine oder mehrere Bedingungen angegeben, die erfüllt sein müssen, damit das Zusammenführen stattfindet.

Abbildung 3-18
Festlegen von Bedingungen für das Zusammenführen

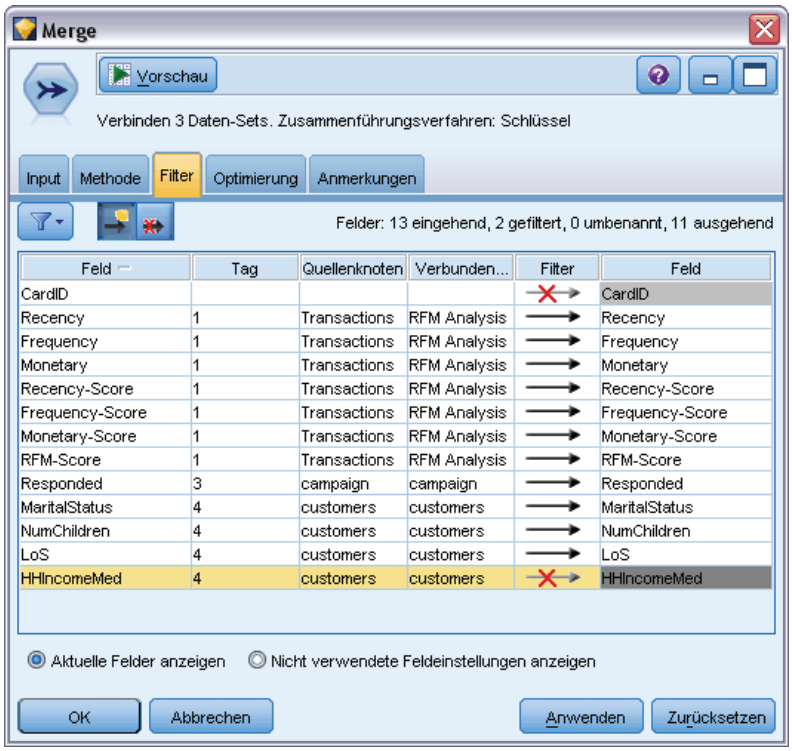


Sie können die Bedingungen entweder direkt in das Feld “Bedingung” eingeben oder sie mithilfe des Expression Builder erstellen, indem Sie auf das Rechnersymbol rechts neben dem Feld klicken.

Filtern von Feldern aus dem Zusammenführungsknoten (“Mergen”)

Zusammenführungsknoten (“Mergen”) bieten eine bequeme Möglichkeit zum Filtern oder Umbenennen doppelter Felder, die beim Zusammenführen mehrerer Datenquellen entstehen. Klicken Sie auf die Registerkarte Filter im Dialogfeld, um Filteroptionen auszuwählen.

Abbildung 3-19
Filtern von Feldern aus dem Zusammenführungsknoten (“Mergen”)



Die hier verfügbaren Optionen sind annähernd mit denen für den Filterknoten identisch. Im Filtermenü stehen jedoch zusätzliche Optionen zur Verfügung, die hier nicht erörtert werden. Für weitere Informationen siehe Thema Filtern bzw. Umbenennen von Feldern in Kapitel 4 auf S. 156.

Feld. Zeigt die Eingabefelder aus den aktuell verbundenen Datenquellen an.

Tag. Listet den Tag-Namen (bzw. die Nummer) auf, der der Datenquellenverknüpfung zugeordnet ist. Klicken Sie auf die Registerkarte Eingaben, um aktive Verknüpfungen mit diesem Zusammenführungsknoten (“Mergen”) zu ändern.

Quellenknoten. Zeigt den Quellenknoten an, dessen Daten zusammengeführt werden.

Verbundener Knoten. Zeigt den Knotennamen für den Knoten an, der mit dem Zusammenführungsknoten (“Mergen”) verbunden ist. Häufig sind beim komplexen Data Mining mehrere Zusammenführungs- bzw. Anhangsvorgänge erforderlich, die denselben Quellenknoten beinhalten können. Der Name des verbundenen Knotens bietet eine Möglichkeit zur Unterscheidung

Filter. Zeigt die aktiven Verbindungen zwischen Eingabe- und Ausgabefeld an. Aktive Verbindungen weisen einen nicht unterbrochenen Pfeil auf. Verbindungen mit einem roten X weisen auf gefilterte Felder hin.

Feld. Führt die Ausgabefelder nach dem Zusammenführen oder Anhängen auf. Doppelte Felder werden in roter Farbe angezeigt. Klicken Sie auf das Filterfeld oben, um doppelte Felder zu deaktivieren.

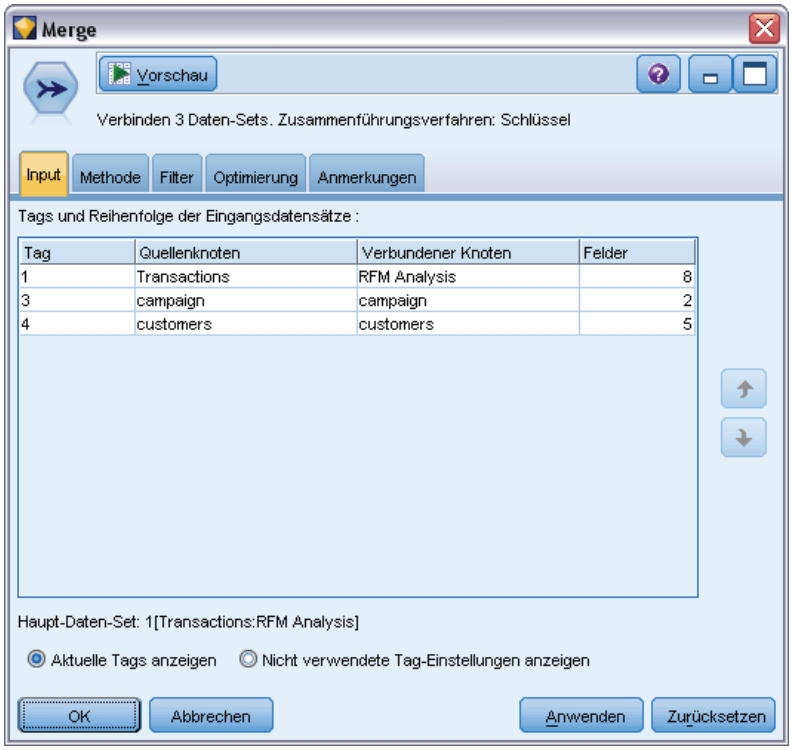
Aktuelle Felder anzeigen. Verwenden Sie diese Option, um Informationen zu den Feldern auszuwählen, die als Schlüsselfelder verwendet werden sollen.

Nicht verwendete Feldeinstellungen anzeigen. Wählen Sie diese Option aus, um Informationen zu Feldern auszuwählen, die derzeit nicht verwendet werden.

Festlegen der Eingabereihenfolge und Tag-Kennzeichnung

Auf der Registerkarte "Eingaben" der Dialogfelder der Zusammenführungs- und Anhangknoten können Sie die Reihenfolge der Eingabedatenquellen angeben und etwaige Änderungen an den Tag-Namen für die einzelnen Quellen vornehmen.

Abbildung 3-20
Verwenden der Registerkarte "Eingaben" zur Angabe von Tags und Eingabereihenfolge



Tags und Reihenfolge der Eingabe-Daten-Sets. Wählen Sie diese Option aus, um nur vollständige Datensätze anzuhängen.

- **Tag.** Listet die aktuellen Tag-Namen für die einzelnen Eingabedatenquellen auf. Tag-Namen bzw. **Tags** dienen zur eindeutigen Kennzeichnung der Datenverknüpfungen für das Zusammenführen oder Anhängen. Stellen Sie sich Wasser aus verschiedenen Leitungen vor, das an einem bestimmten Punkt zusammengeführt wird und ab da durch eine einzige Leitung fließt. Die Daten in IBM® SPSS® Modeler fließen in ähnlicher Weise und der Zusammenführungspunkt ist häufig eine komplexe Interaktion zwischen den verschiedenen Datenquellen. Tags bieten eine Möglichkeit zur Verwaltung der Eingaben ("Leitungen") für einen Zusammenführungs- oder Anhangknoten, sodass beim Speichern oder Trennen des Knotens die Links erhalten bleiben und leicht zu identifizieren sind.

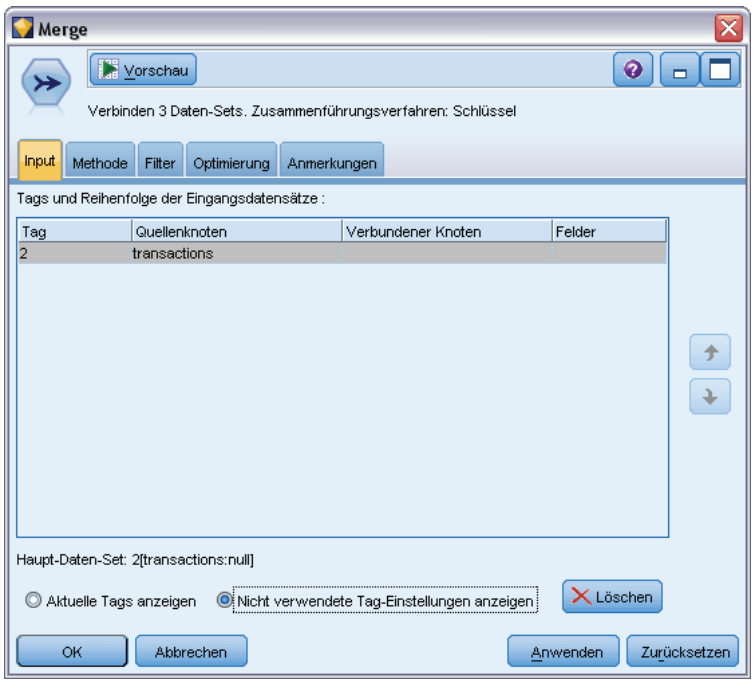
Wenn Sie weitere Datenquellen mit einem Zusammenführungs- oder Anhangknoten verbinden, werden automatisch Standard-Tags erstellt, bei denen die Reihenfolge, in der die Knoten verbunden wurden, durch Zahlen gekennzeichnet wird. Diese Reihenfolge steht in keinem Zusammenhang mit den Feldern im Ein- oder Ausgabe-Daten-Set. Sie können das Standard-Tag ändern, indem Sie in der Spalte *Tag* einen neuen Namen eingeben.

- **Quellenknoten.** Zeigt den Quellenknoten an, dessen Daten zusammengefasst werden.
- **Verbundener Knoten.** Zeigt den Knotennamen für den Knoten an, der mit dem Zusammenführungs- oder Anhangknoten verbunden ist. Häufig sind beim komplexen Data Mining mehrere Zusammenführungsvorgänge erforderlich, die denselben Quellenknoten beinhalten können. Der Name des verbundenen Knotens bietet eine Möglichkeit zur Unterscheidung
- **Felder.** Führt die Anzahl der Felder in den einzelnen Datenquellen auf.

Aktuelle Tags anzeigen. Wählen Sie diese Option aus, um Tags anzuzeigen, die aktiv vom Zusammenführungs- oder Anhangknoten verwendet werden. Anders ausgedrückt: Die aktuellen Tags kennzeichnen Links zu dem Knoten, bei denen ein Datenfluss vorliegt. In der Leitungsmetapher entsprechen die aktuellen Tags den Leitungen, in denen Wasser fließt.

Nicht verwendete Tag-Einstellungen anzeigen. Wählen Sie diese Option aus, um Tags bzw. Links anzuzeigen, die zuvor für Verbindungen mit dem Zusammenführungs- bzw. Anhangknoten verwendet wurden, die derzeit jedoch nicht mit einer Datenquelle verbunden sind. Dies entspricht leeren Leitungen, die in einem Leitungssystem noch intakt sind. Sie können diese "Leitungen" mit einer neuen Quelle verbinden oder sie entfernen. Um nicht verwendete Tags aus dem Knoten zu entfernen, klicken Sie auf Löschen. Dadurch werden alle nicht verwendeten Tags sofort gelöscht.

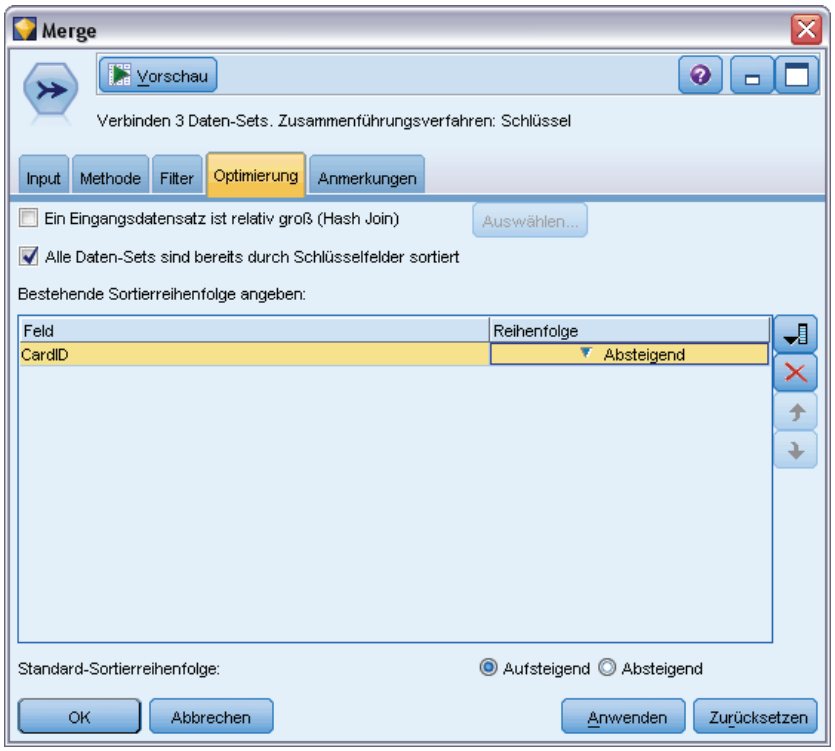
Abbildung 3-21
Entfernen nicht verwendeter Tags aus dem Knoten Zusammenführungsknoten ("Mergen")



Optimierungseinstellungen für das Zusammenführen

Es stehen zwei Optionen zur Auswahl, mit denen Sie die Daten in bestimmten Situationen effektiver zusammenführen. Diese Optionen optimieren die Zusammenführung, wenn ein Eingabe-Daten-Set deutlich größer ist als die anderen Daten-Sets oder wenn Ihre Daten bereits nach einigen oder allen Schlüsselfeldern sortiert sind, die beim Zusammenführen herangezogen werden.

Abbildung 3-22
Optimierungseinstellungen



Ein Eingabe-Daten-Set ist relativ groß. Mit dieser Option geben Sie an, dass eines der Eingabe-Daten-Sets deutlich größer ist als die anderen Daten-Sets. Die kleineren Daten-Sets werden zwischengespeichert. Bei der anschließenden Zusammenführung wird das große Daten-Set ohne Zwischenspeichern und ohne Sortieren verarbeitet. Dieser Join-Typ bietet sich in der Regel für Daten an, die ein Sternen-Schema oder einen ähnlichen Aufbau besitzen, bei dem eine große, zentrale Tabelle mit gemeinsam genutzten Daten vorliegt (z. B. bei Transaktionsdaten). Wenn Sie diese Option aktivieren, klicken Sie auf Auswählen und geben Sie das große Daten-Set an. Hierbei ist zu beachten, dass Sie nur *ein* großes Daten-Set festlegen können. Die nachstehende Tabelle bietet einen Überblick über die Join-Typen, die mit dieser Methode optimiert werden können.

Join-Typ	**Optimierung für ein großes Eingabe-Daten-Set möglich?**
Inner	Ja
Partiell...	Ja, wenn das große Daten-Set keine unvollständigen Datensätze enthält.
Full	Nein
Anti-Join	Ja, wenn das große Daten-Set die erste Eingabe bildet.

Alle Datensätze sind bereits durch Schlüsselfelder sortiert. Mit dieser Option geben Sie an, dass die Eingabedaten bereits nach mindestens einem der Schlüsselfelder sortiert sind, die beim Zusammenführen herangezogen werden sollen. Stellen Sie sicher, dass *alle* Eingabe-Daten-Sets sortiert sind.

Bestehende Sortierreihenfolge angeben. Gibt die Felder an, die bereits sortiert sind. Über das Dialogfeld "Felder auswählen" nehmen Sie Felder in die Liste auf. Sie können nur unter den Schlüsselfeldern wählen, die für die Zusammenführung verwendet werden (auf der Registerkarte "Verbinden" angegeben). Geben Sie in der Spalte *Reihenfolge* jeweils an, ob die Felder in aufsteigender oder absteigender Reihenfolge sortiert sind. Wenn Sie mehrere Felder angeben, achten Sie darauf, die Felder in der richtigen Sortierreihenfolge aufzuführen. Mit den Pfeilen rechts neben der Liste ordnen Sie die Felder in der richtigen Reihenfolge an. Wenn Sie die vorhandene Sortierreihenfolge nicht richtig angeben, tritt beim Ausführen des Streams ein Fehler auf. In der Fehlermeldung wird dabei die Nummer des Datensatzes angezeigt, bei dem die Sortierung nicht mit Ihren Angaben übereinstimmt.

Abhängig davon, ob bei der von der Datenbank verwendeten Kollationsmethode die Groß- und Kleinschreibung berücksichtigt wird, funktioniert die Optimierung möglicherweise nicht ordnungsgemäß, wenn eine oder mehrere Eingaben von der Datenbank sortiert werden. Wenn Sie beispielsweise zwei Eingaben verwenden, wobei bei der einen zwischen Groß- und Kleinschreibung unterschieden wird und bei der anderen nicht, könnten die Ergebnisse der Sortierung voneinander abweichen. Die Zusammenführungsoptimierung führt dazu, dass die Datensätze gemäß ihrer sortierten Reihenfolge verarbeitet werden. Wenn die Eingaben mittels verschiedener Kollationsmethoden sortiert wurden, meldet der Zusammenführungsknoten daher einen Fehler und zeigt die Nummer des Datensatzes an, in dem die Sortierung inkonsistent ist. Wenn alle Eingaben aus derselben Quelle stammen oder mithilfe von sich gegenseitig einschließenden Kollationen sortiert wurden, können die Datensätze erfolgreich zusammengeführt werden.

Hinweis: Die Zusammenführung kann ggf. durch Parallelverarbeitung beschleunigt werden.

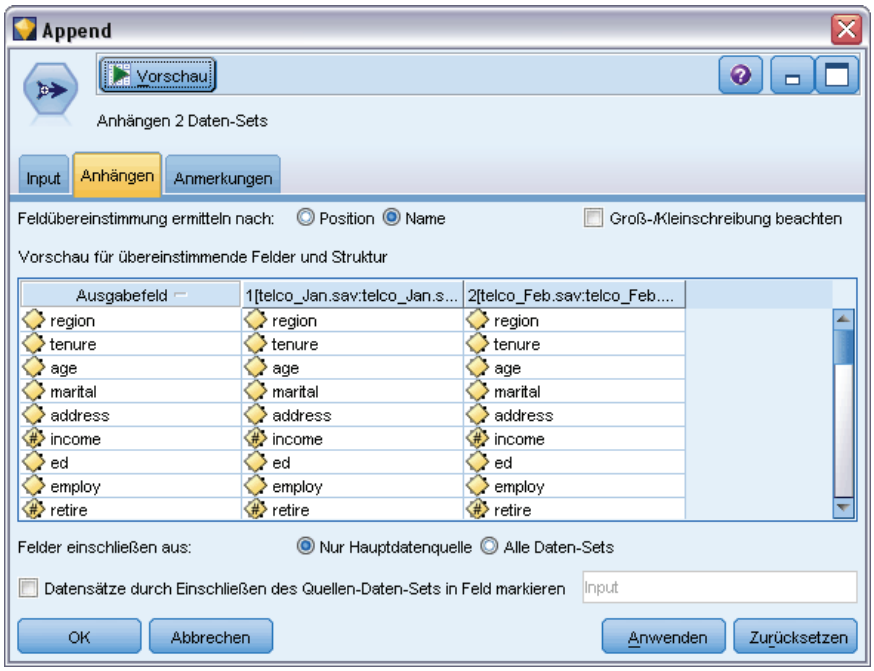
Anhangknoten

Mit Anhangknoten können Sie Sets von Datensätzen miteinander verketten. Anders als bei Zusammenführungsknoten ("Mergen"), in denen Datensätze aus verschiedenen Quellen miteinander verbunden werden, lesen Anhangknoten alle Datensätze aus einer Quelle und geben Sie nach unten im Stream weiter, bis keine mehr vorhanden sind. Anschließend werden die Datensätze aus der nächsten Quelle unter Verwendung derselben Datenstruktur (Anzahl der Datensätze, Anzahl der Felder usw.) wie bei der ersten Eingabe (Primäreingabe) gelesen. Wenn die Primärquelle mehr Felder aufweist als eine andere Eingabequelle, wird die systemdefinierte Null-Zeichenkette ($null$) für alle unvollständigen Werte verwendet.

Anhangknoten sind sinnvoll für die Kombination von Daten-Sets mit ähnlicher Struktur, aber unterschiedlichen Daten. Sie könnten beispielsweise Transaktionsdaten für verschiedene Zeiträume in verschiedenen Dateien gespeichert haben (z. B. zwei Absatzdatendateien für März und April). Angenommen, diese Dateien weisen dieselbe Struktur (dieselben Felder in derselben Reihenfolge) auf, werden sie mit dem Anhangknoten in einer großen Datei zusammengefasst, die anschließend analysiert werden kann.

Hinweis: Zum Anhängen von Dateien sind ähnliche Feldmessniveaus erforderlich. Beispiel: Einem Feld des Typs *Nominal* kann kein Feld angehängt werden, dessen Messniveau *Stetig* ist.

Abbildung 3-23
Dialogfeld des Anhangknotens mit Feldabgleichung nach Namen



Festlegen der Anhangoptionen

Feldübereinstimmung ermitteln nach. Dient zur Auswahl einer Methode für die Abgleichung der anzuhängenden Felder.

- **Position.** Wählen Sie diese Option aus, um Daten-Sets auf der Grundlage der Position der Felder in der Hauptdatenquelle anzuhängen. Bei Verwendung dieser Methode sollten Ihre Daten sortiert sein, um einen ordnungsgemäßen Anhang zu gewährleisten.
- **Name.** Wählen Sie diese Option aus, um Daten-Sets auf der Grundlage des Namens der Felder in den Eingabe-Daten-Sets anzuhängen. Wählen Sie außerdem Groß-/Kleinschreibung beachten, wenn beim Abgleichen der Feldnamen die Groß- und Kleinschreibung berücksichtigt werden soll.

Ausgabefeld. Listet die Quellenknoten auf, die mit dem Anhangknoten verbunden sind. Der erste Knoten in der Liste ist die primäre Eingabequelle. Die Felder in der Anzeige können durch Klicken auf den Spaltentitel sortiert werden. Bei dieser Sortierung wird keine tatsächliche Umordnung der Felder im Daten-Set durchgeführt.

Felder einschließen aus. Wählen Sie die Option Nur Hauptdatenquelle aus, um Ausgabefelder auf der Grundlage der Felder in der Hauptdatenquelle zu erstellen. Die Hauptdatenquelle ist die erste Eingabe, die auf der Registerkarte “Eingaben” angegeben wurde. Wählen Sie Alle Daten-Sets aus, um Ausgabefelder für alle Felder in allen Daten-Sets zu erstellen, unabhängig davon, ob ein Feld vorhanden ist, das in allen Eingabe-Datensätzen übereinstimmt.

Datensätze durch Einschließen des Quellen-Daten-Sets im Feld markieren. Wählen Sie diese Option, um ein zusätzliches Feld zur Ausgabedatei hinzuzufügen, dessen Werte das Quellen-Daten-Set für die einzelnen Datensätze anzeigen. Geben Sie im Textfeld einen Namen an. Der Standardname des Felds lautet *Eingabe*.

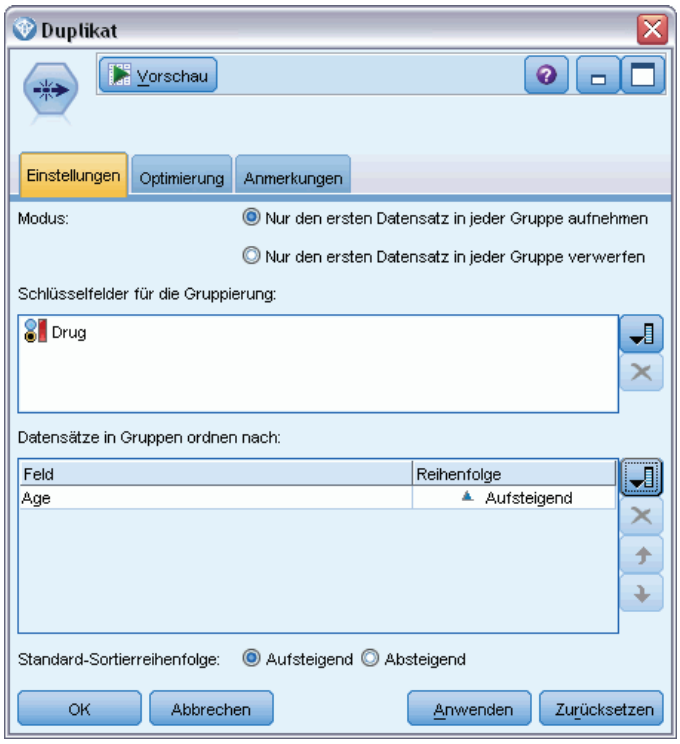
Duplikatknoten

Doppelte Datensätze in einem Daten-Set müssen entfernt werden, bevor mit dem Data Mining begonnen werden kann. In einer Marketingdatenbank beispielsweise werden einzelne Personen möglicherweise mehrfach mit unterschiedlichen Adress- oder Firmendaten aufgeführt. Mit dem Duplikatknoten können Sie nach doppelten Datensätzen in Ihrem Daten-Set suchen und diese entfernen.

Mit dem Duplikatknoten können Sie doppelte Datensätze entweder entfernen, indem jeweils der erste Datensatz an den Daten-Stream übergeben wird, oder aber nach doppelten Datensätzen suchen, indem der erste Datensatz verworfen wird und stattdessen etwaige Duplikate an den Stream übergeben werden.

Zusätzlich können Sie für jeden distinkten Schlüsselwert eine Sortierreihenfolge für die Ergebnisse festlegen. Wenn für jeden distinkten Schlüssel eine bestimmte Zeile angezeigt werden sollen, müssen Sie die Datensätze innerhalb des Duplikatknotens sortieren, anstatt einen weiter oben liegenden Sortierknoten zu verwenden (siehe “Sortieren von Datensätzen innerhalb des Duplikatknotens” weiter unten).

Abbildung 3-24
Duplikatknoten – Dialogfeld



Modalwert. Dient zur Auswahl, ob der erste Datensatz aufgenommen oder ausgeschlossen (verworfen) werden soll.

- **Nur jeweils den ersten Datensatz in jeder Gruppe aufnehmen.** Nimmt den jeweils ersten Datensatz in den Daten-Stream auf und entfernt alle Duplikate.
- **Nur jeweils den ersten Datensatz in jeder Gruppe verwerfen.** Verwirft jeweils den ersten gefundenen Datensatz und übergibt stattdessen etwaige doppelte Datensätze an den Daten-Stream. Mit dieser Option können Duplikate in den Daten *gefunden* werden, um sie später im Stream zu untersuchen.

Schlüsselfelder zur Gruppierung. Listet die Felder auf, die verwendet werden, um zu bestimmen, ob die Datensätze identisch sind. Sie verfügen über folgende Möglichkeiten:

- Zum Hinzufügen von Feldern zu dieser Liste verwenden Sie die Feldauswahl-Schaltfläche auf der rechten Seite.
- Zum Löschen von Feldern aus der Liste verwenden Sie die rote X (Lösch)-Schaltfläche.

Datensätze innerhalb von Gruppen sortieren nach. Listet die Felder auf, die verwendet werden, um zu bestimmen, wie Datensätze innerhalb jedes distinkten Schlüsselwerts sortiert werden und ob sie in auf- oder absteigender Reihenfolge sortiert werden. Sie verfügen über folgende Möglichkeiten:

- Zum Hinzufügen von Feldern zu dieser Liste verwenden Sie die Feldauswahl-Schaltfläche auf der rechten Seite.
- Zum Löschen von Feldern aus der Liste verwenden Sie die rote X (Lösch)-Schaltfläche.
- Verschieben Sie Felder mit den Schaltflächen “Nach oben” oder “Nach unten”, wenn Sie nach mehr als einem Feld sortieren.

Standard-Sortierreihenfolge. Legen Sie fest, ob Datensätze standardmäßig Aufsteigend oder Absteigend sortiert werden sollen.

Sortieren von Datensätzen innerhalb des Duplikatknotens

Mithilfe der Option Datensätze innerhalb von Gruppen sortieren nach innerhalb des Duplikatknotens können Sie für jeden distinkten Schlüssel eine bestimmte Zeile anzeigen; das Setzen eines vorangehenden Sortierknotens ist nicht nötig. Nehmen wir zum Beispiel an, wir besitzen folgende Daten über das Alter von Menschen, die verschreibungspflichtige Medikamente einnehmen:

Alter	Medikament
50	Medikament A
71	Medikament B
44	Medikament A
65	Medikament X
39	Medikament A
75	Medikament C

Alter	Medikament
72	Medikament Y
57	Medikament X
79	Medikament Y
69	Medikament C
74	Medikament B
85	Medikament Y
69	Medikament X

Um den ältesten Anwender eines jeden Medikaments zu ermitteln, würden wir den Modalwert auf “Nur jeweils den ersten Datensatz in jeder Gruppe aufnehmen” setzen, “Medikament” als Schlüsselfeld und “Alter” als Sortierfeld verwenden und eine absteigende Sortierreihenfolge auswählen. Die Reihenfolge der Eingaben hat keine Auswirkungen auf das Ergebnis, denn die Sortiereinstellungen legen fest, welche der Zeilen für ein bestimmtes Medikament angezeigt wird. Die endgültige Datenausgabe sähe also wie folgt aus:

Alter	Medikament
50	Medikament A
74	Medikament B
75	Medikament C
69	Medikament X
85	Medikament Y

Distinkte Optimierungseinstellungen

Wenn die Daten, an denen Sie arbeiten, nur eine kleine Anzahl an Datensätzen umfassen oder bereits sortiert wurden, können Sie ihre Behandlung so optimieren, dass IBM® SPSS® Modeler die Daten effizienter verarbeitet.

Hinweis: Wenn Sie entweder Eingabedaten-Set verfügt über eine kleinere Anzahl an distinkten Schlüsseln auswählen oder die SQL-Erzeugung für den Knoten verwenden, kann jede Zeile innerhalb des distinkten Schlüsselwerts angezeigt werden; um zu kontrollieren, welche Zeile innerhalb eines distinkten Schlüssels angezeigt wird, müssen Sie die Sortierreihenfolge mithilfe des Felds Datensätze innerhalb von Gruppen sortieren nach in der Registerkarte “Einstellungen” festlegen. Die Optimierungs-Optionen haben keine Auswirkungen auf die Ergebnisausgabe des Duplikatknotens, solange Sie in der Registerkarte “Einstellungen” eine Sortierreihenfolge festgelegt haben.

Abbildung 3-25
Optimierungseinstellungen

Eingabedaten-Set verfügt über eine kleinere Anzahl an distinkten Schlüsseln. Wählen Sie diese Option, wenn Sie über eine kleine Anzahl an Datensätzen oder eine kleine Anzahl an eindeutigen Werten der Schlüsselfelder oder beides verfügen. Dadurch lässt sich eventuell die Leistungsfähigkeit verbessern.

Eingabedaten-Set ist bereits nach Gruppier- und Sortierfeldern auf der Registerkarte "Einstellungen" sortiert. Wählen Sie diese Option nur aus, wenn Ihre Daten bereits nach allen Feldern sortiert sind, die auf der Registerkarte "Einstellungen" unter Datensätze innerhalb von Gruppen sortieren nach aufgelistet sind, und wenn die auf- und absteigende Sortierreihenfolge der Daten identisch ist. Dadurch lässt sich eventuell die Leistungsfähigkeit verbessern.

SQL-Erzeugung deaktivieren. Wählen Sie diese Option aus, um die SQL-Erzeugung für den Knoten zu deaktivieren.

Kapitel 4

Feldoperationsknoten

Feldoperationen – Überblick

Nach der ersten Datenuntersuchung steht in der Regel die Auswahl, die Bereinigung oder die Konstruktion von Daten als Vorbereitung für die Analyse an. Die Palette "Feldfunktionen" enthält viele Knoten, die für diese Transformation und Vorbereitung nützlich sind.

So können Sie mit einem Ableitungsknoten ein Attribut erstellen, das noch nicht in den Daten repräsentiert wird. Oder Sie können mit einem Klassierknoten automatisch Feldwerte für eine gezielte Analyse neu kodieren. Typknoten werden häufig verwendet, da sie die Möglichkeit bieten, für jedes Feld im Daten-Set ein Messniveau, Werte und eine Modellierungsrolle zuzuweisen. Diese Operationen sind nützlich für den Umgang mit fehlenden Werten und für die Downstream-Modellierung.

Die Palette "Feldfunktionen" enthält folgende Knoten:

Der Knoten "Automated Data Preparation" (ADP) kann Ihre Daten analysieren und Korrekturen identifizieren, problematische oder vermutlich überflüssige Felder ausschließen, wie erforderlich neue Attribute ableiten und die Leistung durch intelligente Prüf- und Stichprobenverfahren verbessern. Sie können den Knoten vollständig automatisiert nutzen, damit er Korrekturen wählen und anwenden kann. Sie können die Änderungen aber auch prüfen, bevor sie durchgeführt werden, und wie gewünscht akzeptieren, ablehnen oder ändern. Für weitere Informationen siehe Thema Automatisierte Datenaufbereitung auf S. 108.

Der Typknoten gibt Feldmetadaten und Eigenschaften an. Sie können beispielsweise ein Messniveau (stetig, nominal, ordinal oder Flag) für die einzelnen Felder angeben, Optionen für den Umgang mit fehlenden Werten und systemdefinierten Nullwerten festlegen, die Rolle eines Felds zu Modellierungszwecken festlegen, Feld- und Wertelabels angeben oder die Werte für ein Feld angeben. Für weitere Informationen siehe Thema Typknoten auf S. 136.

Der Filterknoten filtert (verwirft) Felder, benennt Felder um und ordnet Felder von einem Quellenknoten einem anderen zu. Für weitere Informationen siehe Thema Filtern bzw. Umbenennen von Feldern auf S. 156.

Der Ableitungsknoten ändert Datenwerte oder erstellt neue Felder aus einem oder mehreren bestehenden Feldern. Er erstellt Felder vom Typ "Formel", "Flag", "Nominal", "Status", "Anzahl" und "Bedingt". Für weitere Informationen siehe Thema Ableitungsknoten auf S. 167.

Der Ensemble-Knoten kombiniert zwei oder mehr Modell-Nuggets, um genauere Vorhersagen zu erzielen, als aus einem dieser Modelle allein gewonnen werden können. Für weitere Informationen siehe Thema Ensemble-Knoten auf S. 163.

Der Füllerknoten ersetzt Feldwerte und ändert den Speichertyp. Sie können auswählen, dass die Werte auf der Grundlage einer CLEM-Bedingung wie beispielsweise @BLANK(@FIELD) ersetzt werden sollen. Alternativ können Sie auswählen, dass alle Leerstellen oder Nullwerte mit einem bestimmten Wert ersetzt werden sollen. Füllerknoten werden häufig zusammen mit einem Typknoten verwendet, um fehlende Werte zu ersetzen. Für weitere Informationen siehe Thema Füllerknoten auf S. 179.

Der Anonymisierungsknoten ändert die Art und Weise, wie Feldnamen und -werte weiter unten im Stream dargestellt werden, und verschleiert damit die ursprünglichen Daten. Dies kann sinnvoll sein, wenn andere Benutzer in die Lage versetzt werden sollen, Modelle unter Verwendung vertraulicher Daten wie beispielsweise Kundennamen zu erstellen. Für weitere Informationen siehe Thema Anonymisierungsknoten auf S. 183.

Der Umkodierungsknoten transformiert ein Set kategorialer Werte in ein anderes. Die Umkodierung dient zur Reduzierung von Kategorien bzw. Neugruppierung von Daten für die Analyse. Für weitere Informationen siehe Thema Umkodierungsknoten auf S. 187.

Der Klassierknoten erstellt automatisch neue nominale (Set-) Felder auf der Grundlage der Werte eines oder mehrerer bestehender stetiger Felder (numerischer Bereich). Sie können beispielsweise ein stetiges Einkommensfeld in ein neues kategoriales Feld transformieren, das Einkommensgruppen als Abweichungen vom Mittelwert enthält. Nach der Erstellung von Klassen für das neue Feld können Sie einen Ableitungsknoten anhand der Trennwerte generieren. Für weitere Informationen siehe Thema Klassierknoten auf S. 192.

Mit dem Knoten “RFM-Analyse” (Recency-, Frequency-, Monetary-Analyse) können Sie quantitativ ermitteln, welche Kunden wahrscheinlich die besten sind, indem Sie untersuchen, wann sie zuletzt etwas von Ihnen erworben haben (Recency (Aktualität)), wie häufig sie eingekauft haben (Frequency (Häufigkeit)) und wie viel sie für alle Transaktionen zusammengenommen ausgegeben haben (Monetary (Geldwert)). Für weitere Informationen siehe Thema Knoten “RFM-Analyse” auf S. 204.

Der Partitionsknoten erstellt ein Partitionsfeld, das Daten in getrennte Untergruppen für die Trainings-, Test- und Validierungsphase der Modellerstellung aufteilt. Für weitere Informationen siehe Thema Partitionsknoten auf S. 208.

Der Dichotomknoten leitet mehrere Flag-Felder auf der Grundlage der kategorialen Werte ab, die für ein oder mehrere nominale Felder definiert sind. Für weitere Informationen siehe Thema Dichotomknoten auf S. 211.

Der Knoten “Neu strukturieren” wandelt ein nominales Feld oder ein Flag-Feld in eine Gruppe von Feldern um, die mit den Werten aus einem weiteren Feld ausgefüllt werden können. Beispiel: Aus einem Feld mit dem Namen *Zahlungsart*, mit den Werten *Kreditkarte*, *Bar* und *EC-Karte* werden drei neue Felder erstellt (*Kreditkarte*, *Bar*, *EC-Karte*), die jeweils den Wert der jeweiligen Zahlung enthalten. Für weitere Informationen siehe Thema Neustrukturierungsknoten auf S. 212.

Der Transponierknoten vertauscht die Daten in Zeilen und Spalten, sodass aus Datensätzen Felder und aus Feldern Datensätze werden. Für weitere Informationen siehe Thema Transponierknoten auf S. 215.

Der Zeitintervallknoten gibt Intervalle an und erstellt (bei Bedarf) Beschriftungen für die Modellierung von Zeitreihendaten. Wenn die Werte nicht gleichmäßig verteilt sind, kann der Knoten nach Bedarf Werte auffüllen oder aggregieren, um ein gleichmäßiges Intervall zwischen den Datensätzen zu erzeugen. Für weitere Informationen siehe Thema Zeitintervallknoten auf S. 219.

Der Verlaufsknoten erstellt neue Felder mit Daten aus Feldern in vorangegangenen Datensätzen. Verlaufsknoten werden am häufigsten für sequenzielle Daten, beispielsweise Zeitreihendaten, verwendet. Vor der Verwendung eines Verlaufsknotens sollten die Daten mithilfe eines Sortierknotens sortiert werden. Für weitere Informationen siehe Thema Verlaufsknoten auf S. 240.

Der Knoten “Felder ordnen” definiert die natürliche Reihenfolge, die bei der Anzeige der weiter unten im Stream liegenden Felder verwendet wird. Diese Reihenfolge betrifft die Anzeige von Feldern an unterschiedlichen Stellen, beispielsweise in Tabellen, Listen und in der Feldauswahl. Dieser Vorgang dient beispielsweise dazu, um bei der Arbeit mit umfangreichen Daten-Sets die relevanten Felder deutlicher hervorzuheben. Für weitere Informationen siehe Thema Knoten “Felder ordnen” auf S. 241.

Mehrere dieser Knoten können direkt über den von einem Data Audit-Knoten erstellten Audit-Bericht generiert werden. Für weitere Informationen siehe Thema Erzeugen von anderen Knoten zur Datenvorbereitung in Kapitel 6 auf S. 435.

Automatisierte Datenaufbereitung

Die Aufbereitung von Daten zur Analyse ist einer der wichtigsten Schritte in jedem Projekt – und gewöhnlich auch einer der zeitaufwendigsten. Die automatisierte Datenaufbereitung (ADP) übernimmt diese Aufgabe für Sie. Sie analysiert Ihre Daten und identifiziert Problemlösungen, findet problematische oder wahrscheinlich nicht nützliche Felder, leitet zum passenden Zeitpunkt neue Attribute ab und verbessert die Leistungsfähigkeit durch intelligente Screening-Methoden. Sie können den Algorithmus **vollautomatisch** verwenden und so Problemlösungen auswählen und anwenden oder Sie können ihn **interaktiv** verwenden und so die Änderungen in einer Vorschau betrachten, bevor sie vorgenommen werden, und sie gegebenenfalls akzeptieren oder ablehnen.

Mit ADP können Sie Ihre Daten schnell und einfach für die Modellerstellung aufbereiten, ohne über Vorkenntnisse der dazugehörigen statistischen Konzepte verfügen zu müssen. Modelle lassen sich damit schneller erstellen und scoren; zudem verbessert sich mit ADP die Robustheit automatisierter Modellierungsprozesse wie der Modellaktualisierung und von Champion/Challenger.

Anmerkung: Wenn die ADP ein Feld für die Analyse vorbereitet, erstellt sie ein neues Feld, das die Anpassungen oder Transformationen enthält, anstatt die bestehenden Werte und Eigenschaften des alten Felds zu ersetzen. Das alte Feld wird bei der weiteren Analyse nicht verwendet; seine Rolle wird auf “Keine” gesetzt.

Beispiel. Eine Versicherungsgesellschaft mit beschränkten Ressourcen für die Untersuchung der Versicherungsansprüche von Hauseigentümern möchte ein Modell zur Kennzeichnung verdächtiger, potenziell betrügerischer Ansprüche erstellen. Vor Erstellung des Modells bereiten sie die Daten für die Modellierung mithilfe der automatisierten Datenaufbereitung vor. Da sie die

vorgeschlagenen Transformationen zunächst überprüfen möchten, bevor die Transformationen angewendet werden, nutzen sie die automatisierte Datenaufbereitung im interaktiven Modus.

Eine Gruppe in der Kraftfahrzeugindustrie erfasst die Verkaufszahlen verschiedener Personenkraftwagen. Um starke und schwache Modelle identifizieren zu können, soll eine Beziehung zwischen den Fahrzeugverkaufszahlen und den Fahrzeugeigenschaften hergestellt werden. Zur Vorbereitung der Daten für die Analyse wird die automatisierte Datenaufbereitung verwendet. Es werden Modelle mit Daten "vor" und "nach" der Aufbereitung erstellt, um zu sehen, wie sich die Ergebnisse unterscheiden.

Abbildung 4-1
Registerkarte "Ziel" in der automatisierten Datenaufbereitung

Wie lautet Ihr Ziel? Die automatisierte Datenaufbereitung empfiehlt Schritte zur Datenaufbereitung, die sich auf die Geschwindigkeit auswirken, mit der andere Algorithmen Modelle erstellen können und die Vorhersagekraft dieser Modelle verbessern. Diese können die Transformation, Erstellung und Auswahl von Funktionen beinhalten. Das Ziel kann ebenfalls transformiert werden. Sie können die Prioritäten der Modellerstellung festlegen, auf die sich die Datenaufbereitung konzentrieren sollte.

- **Geschwindigkeit und Genauigkeit ausgleichen.** Diese Option bereitet die Daten auf und sorgt dabei für eine ausgeglichene Priorität zwischen der Geschwindigkeit, mit der Daten durch die Modellerstellung verarbeitet werden, und der Genauigkeit der Vorhersagen.
- **Geschwindigkeit optimieren.** Diese Option bereitet die Daten auf und gibt dabei der Geschwindigkeit Vorrang, mit der Daten durch Modellerstellungsalgorithmen verarbeitet werden. Wählen Sie diese Option, wenn Sie mit sehr großen Daten-Sets arbeiten oder nach einer schnellen Antwort suchen.
- **Genauigkeit optimieren.** Diese Option bereitet die Daten auf und gibt dabei der Genauigkeit der durch Modellerstellungsalgorithmen erzeugten Vorhersagen Vorrang.
- **Analyse anpassen** Wählen Sie diese Option, wenn Sie den Algorithmus auf der Registerkarte "Einstellungen" manuell ändern wollen. Beachten Sie, dass diese Einstellung automatisch ausgewählt wird, wenn Sie anschließend Änderungen auf der Registerkarte "Einstellungen" vornehmen, die mit einem der anderen Ziele nicht kompatibel sind.

Knoten-Training

Der ADP-Knoten wurde als Prozessknoten implementiert und arbeitet ähnlich wie der Typknoten; **Training** des ADP-Knotens entspricht der Instantiierung des Typknotens. Sobald die Analyse durchgeführt wurde, werden die angegebenen Transformationen ohne weitere Analyse auf die Daten angewendet, solange sich das vorgelagerte Datenmodell nicht ändert. Wenn die Verbindung zum ADP-Knoten getrennt wird, speichert dieser wie die Typ- und Filterknoten das Datenmodell und die Transformationen und muss so nicht erneut trainiert werden, wenn die Verbindung wiederhergestellt wird; dadurch können Sie ihn auf eine Untergruppe typischer Daten trainieren und anschließend kopieren oder so oft wie nötig auf Live-Daten bereitstellen.

Verwendung der Symbolleiste

Mit der Symbolleiste können Sie die Anzeige der Datenanalyse ausführen und aktualisieren sowie Knoten generieren, die Sie zusammen mit den Originaldaten verwenden können.

Abbildung 4-2
Automatisierte Datenaufbereitung – Symbolleiste

- **Erzeugen** In diesem Menü können Sie entweder einen Filter- oder einen Ableitungsknoten erzeugen. Beachten Sie, dass dieses Menü nur verfügbar ist, wenn auf der Registerkarte "Analyse" eine Analyse angezeigt wird.

  Der Filterknoten entfernt transformierte Eingabefelder. Wenn Sie den ADP-Knoten so konfigurieren, dass die Original-Eingabefelder im Daten-Set beibehalten werden, wird dadurch das Original-Eingabe-Set wiederhergestellt und Sie können das Wertfeld bezüglich der Eingaben interpretieren. Dies ist beispielsweise dann nützlich, wenn Sie eine Grafik des Wertfelds anhand mehrerer Eingaben erzeugen möchten.

  Der Ableitungsknoten kann das Original-Daten-Set und die Original-Zieleinheiten wiederherstellen. Sie können einen Ableitungsknoten nur dann erzeugen, wenn der ADP-Knoten eine Analyse enthält, die ein Bereichsziel neu skaliert (d. h. die Box-Cox-Neuskalierung ist im Feld "Eingaben & Ziel vorbereiten" ausgewählt). Sie können keinen Ableitungsknoten erzeugen, wenn das Ziel kein Bereich ist oder wenn die Box-Cox-Neuskalierung nicht ausgewählt ist. Für weitere Informationen siehe Thema Erzeugen eines Ableitungsknotens auf S. 134.
- **Ansicht** Enthält Optionen, die steuern, was auf der Registerkarte "Analyse" angezeigt wird. Dazu zählen die Steuerungen zur Bearbeitung von Grafiken sowie die Anzeigen für das Hauptfenster und verknüpfte Ansichten.
- **Vorschau** Zeigt ein Muster der Transformationen an, die auf die Eingabedaten angewendet werden.
- **Daten analysieren** Startet eine Analyse mit den aktuellen Einstellungen und zeigt die Ergebnisse in der Registerkarte "Analyse" an.
- **Analyse löschen** Löscht die bestehende Analyse (nur verfügbar, wenn eine aktuelle Analyse vorhanden ist).

Knotenstatus

Der Status des ADP-Knotens auf der Zeichenfläche von IBM® SPSS® Modeler wird entweder durch einen Pfeil oder ein Häkchen auf dem Symbol verdeutlicht, das anzeigt, ob eine Analyse durchgeführt wurde oder nicht.

Registerkarte "Felder"

Abbildung 4-3

Bevor Sie ein Modell erstellen können, müssen Sie festlegen, welche Felder als Ziele und als Eingaben verwendet werden sollen. Von wenigen Ausnahmen abgesehen, verwenden alle Modellierungsknoten die Feldinformationen des oberhalb liegenden Typknotens. Wenn Sie einen Typknoten benutzen, um Eingabe- und Zielfelder auszuwählen, brauchen Sie auf dieser Registerkarte keine Änderungen vorzunehmen.

Typknoteneinstellungen verwenden. Diese Option weist den Knoten an, die Feldinformationen von einem weiter oben liegenden Typknoten zu verwenden. Dies ist die Standardeinstellung.

Benutzerdefinierte Einstellungen verwenden. Diese Option weist den Knoten an, die hier angegebenen Feldinformationen anstelle der in einem weiter oben liegenden Typknoten angegebenen zu verwenden. Geben Sie nach Auswahl dieser Option wie erforderlich die unten stehenden Felder an.

Ziel. Wählen Sie die Zielfelder für Modelle aus, die eines oder mehrere Zielfelder benötigen. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Ziel* festlegen.

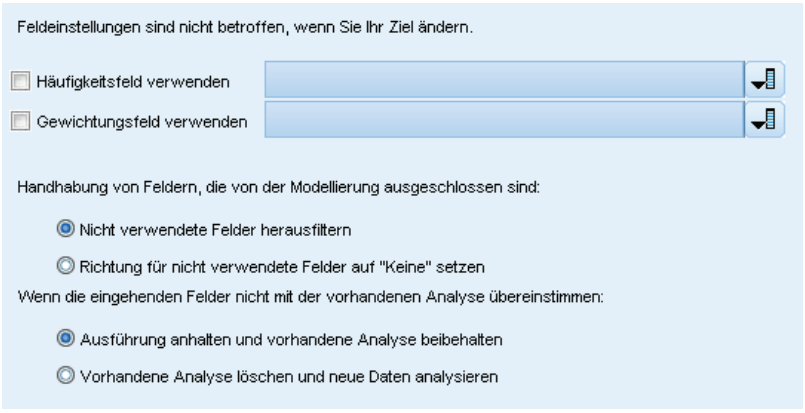
Eingaben. Wählen Sie das Eingabefeld bzw. die Eingabefelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen.

Registerkarte "Einstellungen"

Die Registerkarte "Einstellungen" enthält mehrere unterschiedliche Gruppen von Einstellungen, die Sie ändern können, um genau festzulegen, wie der Algorithmus Ihre Daten verarbeiten soll. Wenn Sie an den Standardeinstellungen Änderungen vornehmen, die mit den anderen Zielen nicht kompatibel sind, wird auf der Registerkarte "Ziel" automatisch die Option Analyse anpassen ausgewählt.

Feldeinstellungen

Abbildung 4-4
Automatisierte Datenaufbereitung – Feldeinstellungen



Häufigkeitsfeld verwenden. Mit dieser Option könen Sie ein Feld als Häufigkeitsgewichtung auswählen. Wenden Sie diese Option an, wenn die Datensätze in Ihren Trainingsdaten jeweils mehr als eine Einheit darstellen; dies ist zum Beispiel bei aggregierten Daten der Fall. Die Feldwerte sollten die Anzahl der Einheiten sein, die von jedem Datensatz repräsentiert werden.

"Feld gewichten" verwenden. Mit dieser Option könen Sie ein Feld als Fallgewichtung auswählen. Fallgewichtungen werden verwendet, um Differenzen in der Varianz zwischen den Ebenen des Ausgabefelds zu berücksichtigen.

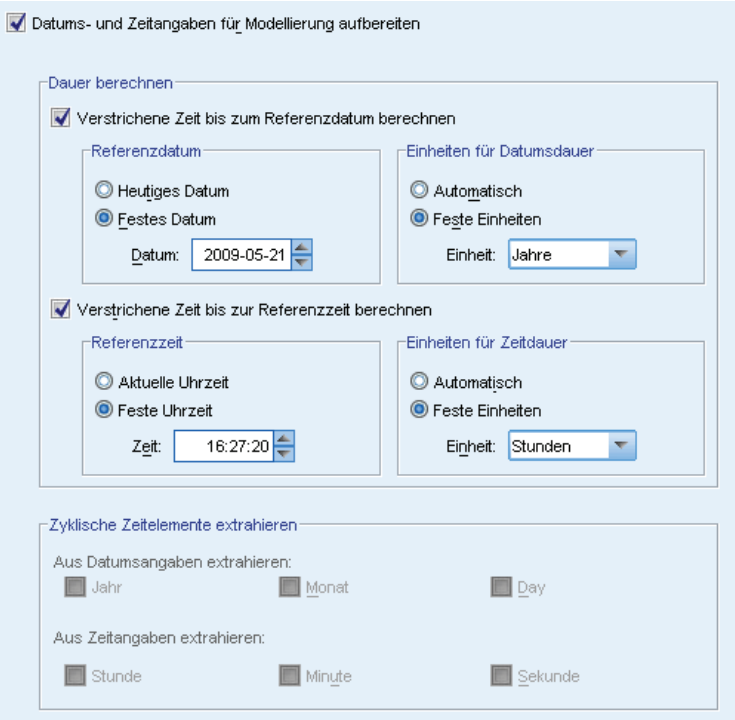
Verarbeiten von Feldern, die von der Modellierung ausgeschlossen sind. Geben Sie an, was mit ausgeschlossenen Feldern geschehen soll. Sie können wählen, ob sie aus den Daten herausgefiltert werden sollen oder einfach ihre *Rolle* auf Keine gesetzt werden soll.

Wenn die eingehenden Felder nicht mit der bestehenden Analyse übereinstimmen. Geben Sie an, was geschehen soll, falls eines oder mehrere erforderliche Eingabefelder im eingehenden Datensatz fehlen, wenn Sie einen trainierten ADP-Knoten ausführen.

- **Ausführung anhalten und bestehende Analyse beibehalten.** Dieser Befehl stoppt den Ausführungsvorgang, speichert die gegenwärtigen Analysedaten und zeigt eine Fehlermeldung an.
- **Vorhandene Analyse löschen und neue Daten analysieren.** Dadurch wird die vorhandene Analyse gelöscht, die eingehenden Daten werden analysiert und die empfohlenen Transformationen werden auf diese Daten angewendet.

Datum und Uhrzeit aufbereiten

Abbildung 4-5
Automatisierte Datenaufbereitung – Datum und Uhrzeit aufbereiten – Einstellungen



Viele Modellierungsalgorithmen sind nicht in der Lage, Datums- und Zeitangaben direkt zu behandeln; mit diesen Einstellungen können Sie neue Laufzeitdaten ableiten, die Sie in Ihren bestehenden Daten als Modelleingaben aus Datums- und Zeitangaben verwenden können. Die Felder mit Datums- und Zeitangaben müssen mit Datums- oder Zeitspeichertypen vordefiniert sein. Die ursprünglichen Datums- und Zeitfelder werden nicht als Modelleingaben nach der automatisierten Datenaufbereitung empfohlen.

Datums- und Zeitangaben für Modellierung aufbereiten. Durch Deaktivieren dieser Option werden alle anderen Datums- und Zeiteingaben deaktiviert und die Auswahl beibehalten.

Verstrichene Zeit bis zum Referenzdatum berechnen. Errechnet die Anzahl der Jahre/Monate/Tage seit einem Referenzdatum für jede Variable, die Datumsangaben enthält.

- **Referenzdatum.** Geben Sie das Datum an, ab dem die Dauer bezüglich der Datumsinformationen in den Eingabedaten berechnet wird. Durch die Auswahl von Heutiges Datum wird das aktuelle Systemdatum stets verwendet, wenn ADP ausgeführt wird. Um ein bestimmtes Datum zu verwenden, wählen Sie Festes Datum und geben Sie das erforderliche Datum ein. Das aktuelle Datum wird automatisch im Feld Festes Datum eingegeben, wenn der Knoten zum ersten Mal erstellt wird.
- **Einheiten für Datumsdauer.** Legen Sie fest, ob ADP die Einheit der Datumsdauer automatisch bestimmen soll, oder wählen Sie Feste Einheiten für Jahre, Monate oder Tage.

Verstrichene Zeit bis zur Referenzzeit berechnen. Errechnet die Anzahl der Stunden/Minuten/Sekunden seit einer Referenzzeit für jede Variable, die Uhrzeiten enthält.

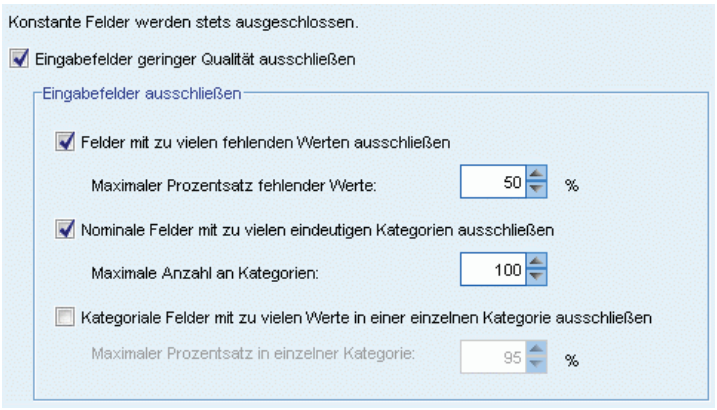
- **Referenzzeit.** Geben Sie die Zeit an, ab der die Dauer bezüglich der Zeitinformationen in den Eingabedaten berechnet wird. Durch die Auswahl von Aktuelle Uhrzeit wird die aktuelle Systemzeit stets verwendet, wenn ADP ausgeführt wird. Um eine bestimmte Uhrzeit zu verwenden, wählen Sie Feste Uhrzeit und geben Sie die erforderlichen Daten ein. Die aktuelle Uhrzeit wird automatisch im Feld Feste Uhrzeit eingegeben, wenn der Knoten zum ersten Mal erstellt wird.
- **Einheiten für Zeitdauer.** Legen Sie fest, ob ADP die Einheit der Zeitdauer automatisch bestimmen soll, oder wählen Sie Feste Einheiten für Stunden, Minuten oder Sekunden.

Zyklische Zeitelemente extrahieren. Verwenden Sie diese Einstellungen, um ein einzelnes Datums- oder Zeitfeld in ein oder mehrere Felder aufzuteilen. Wenn Sie zum Beispiel alle drei Datumskontrollkästchen auswählen, wird das Eingabedatumsfeld "1954-05-23" in drei Felder aufgeteilt: 1954, 5 und 23, wobei jedes das unter Feldnamen definierte Suffix verwendet und das ursprüngliche Datumsfeld ignoriert wird.

- **Aus Datumsangaben extrahieren.** Legen Sie für eine beliebige Datumseingabe fest, ob Sie Jahre, Monate, Tage oder eine Kombination daraus extrahieren möchten.
- **Aus Zeitangaben extrahieren.** Legen Sie für eine beliebige Zeiteingabe fest, ob Sie Stunden, Minuten, Sekunden oder eine Kombination daraus extrahieren möchten.

Felder ausschließen

Abbildung 4-6
Automatisierte Datenaufbereitung – Felder ausschließen – Einstellungen



Schlechte Datenqualität kann sich negativ auf die Genauigkeit Ihrer Vorhersagen auswirken; Sie können daher die akzeptable Qualitätsstufe für Eingabefunktionen festlegen. Alle konstanten oder 100 % an fehlenden Werten aufweisenden Felder werden automatisch ausgeschlossen.

Eingabefelder mit niedriger Qualität ausschließen. Durch Deaktivieren dieser Option werden alle anderen Befehle "Felder ausschließen" deaktiviert und die Auswahl beibehalten.

Felder mit zu vielen fehlenden Werten ausschließen. Felder mit mehr als dem angegebenen Prozentsatz an fehlenden Werten werden aus der weiteren Analyse ausgeschlossen. Geben Sie einen Wert größer oder gleich 0 ein, was dem Deaktivieren dieser Option entspricht, und einen Wert kleiner oder gleich 100, so dass die Felder mit allen fehlenden Werten automatisch ausgeschlossen werden. Der Standardwert lautet "50".

Nominale Felder mit zu vielen eindeutigen Kategorien ausschließen. Nominale Felder mit mehr als der angegebenen Anzahl an Kategorien werden aus der weiteren Analyse ausgeschlossen. Geben Sie eine positive Ganzzahl ein. Der Standardwert ist 100. Dies ist nützlich für das automatische Entfernen von Feldern aus der Modellierung, die eine datensatzeindeutige Information enthalten, wie zum Beispiel eine ID, eine Adresse oder einen Namen.

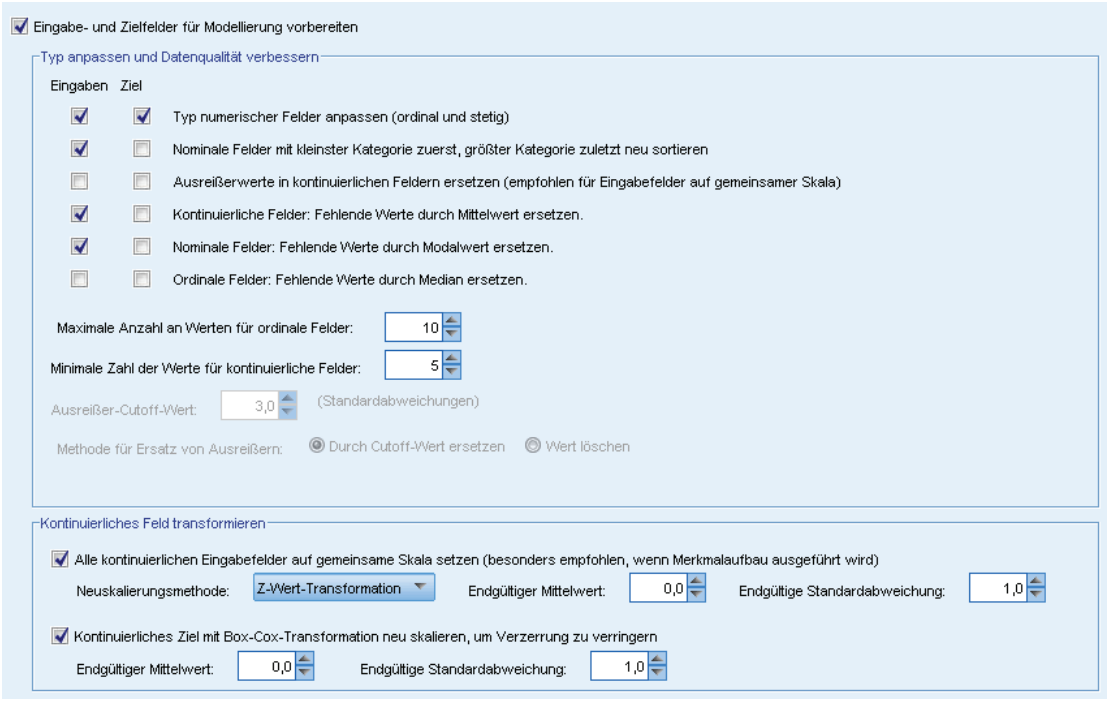
Kategoriale Felder mit zu vielen Werten in einer einzelnen Kategorie ausschließen. Ordinale und nominale Felder mit einer Kategorie, die mehr als die angegebene Prozentzahl an Datensätzen enthält, werden aus der weiteren Analyse ausgeschlossen. Geben Sie einen Wert größer oder gleich 0 ein, was dem Deaktivieren dieser Option entspricht, und einen Wert kleiner oder gleich 100, so dass konstante Felder automatisch ausgeschlossen werden. Der Standardwert ist 95.

Vorbereiten von Eingaben und Zielen

Da sich Daten nie in einem perfekten Zustand für die Verarbeitung befinden, kann es hilfreich sein, vor dem Ausführen einer Analyse einige Einstellungen anzupassen. Dazu können zum Beispiel das Ein- oder Ausschließen von Ausreißern gehören, Angaben über den Umgang mit fehlenden Werten oder das Anpassen des Typs.

Anmerkung: Wenn Sie die Werte in diesem Feld ändern, wird die Registerkarte Ziele automatisch auf die Auswahl der Option Benutzerdefinierte Analyse aktualisiert.

Abbildung 4-7
Automatisierte Datenaufbereitung - Einstellungen von Eingabe und Ziel



Vorbereiten der Eingabe- und Zielfelder für die Modellierung. Schaltet alle Felder in dem Eingabefeld entweder an oder aus.

Anpassen des Typs und Verbessern der Datenqualität. Für die Eingaben und das Ziel können mehrere Datentransformationen separat angegeben werden, da es wünschenswert sein kann, die Zielwerte nicht zu ändern. So kann zum Beispiel eine Vorhersage über das Einkommen in Dollar aussagekräftiger sein als eine Vorhersage, die als Log(Dollar) angegeben wird. Wenn das Ziel außerdem fehlende Werte aufweist, ergibt sich für die Vorhersage kein Nutzen daraus, die fehlenden Werte zu ergänzen, wogegen das Ergänzen fehlender Werte bei den Eingaben durchaus dazu führen kann, dass einige Algorithmen Informationen aufbereiten können, die anderenfalls verloren gehen würden.

Weitere Einstellungen für diese Transformationen, wie zum Beispiel der Ausreißer-Trennwert, sind sowohl für das Ziel als auch die Eingaben üblich.

Sie können die folgenden Einstellungen für entweder die Eingaben oder das Ziel oder für beides vornehmen:

- **Anpassen des Typs des numerischen Felds.** Damit können Sie bestimmen, ob die numerischen Felder mit einem Messniveau *Ordinal* auf *Stetig* konvertiert werden können oder umgekehrt. Sie können die minimalen und maximalen Schwellenwerte für die Konversion angeben.
- **Nominale Felder neu sortieren.** Mit dieser Option können Sie nominale (Set-) Felder der Reihe nach sortieren, von der kleinsten zur größten Kategorie.
- **Ersetzen von Ausreißerwerten in stetigen Feldern.** Geben Sie an, ob Ausreißer ersetzt werden sollen. Nutzen Sie diese Option in den Verbindung mit dem Verfahren zum Ersetzen von Ausreißern unten.
- **Stetige Felder: Fehlende Werte durch Mittelwert ersetzen** Mit dieser Option können Sie fehlende Werte stetiger (Bereichs-) Funktionen ersetzen.
- **Nominale Felder: Fehlende Werte durch Modalwert ersetzen.** Mit dieser Option können Sie fehlende Werte nominaler (Set-) Funktionen ersetzen.
- **Ordinale Felder: Fehlende Werte durch Median ersetzen.** Mit dieser Option können Sie fehlende Werte ordinaler (sortiertes Set) Funktionen ersetzen.

Maximale Anzahl an Werten für ordinale Felder. Geben Sie den Schwellenwert an, bei dem ordinale (sortiertes Set) Felder in stetige (Bereich) Felder umdefiniert werden sollen. Der Standardwert ist 10. Wenn also ein ordinales Feld mehr als 10 Kategorien aufweist, wird es als stetig (Bereich) umdefiniert.

Minimale Anzahl an Werten für stetige Felder. Geben Sie den Schwellenwert an, bei dem Skalenfelder oder stetige (Bereich) Felder in ordinale (sortiertes Set) Felder umdefiniert werden sollen. Der Standardwert ist 5. Wenn also ein stetiges Feld weniger als 5 Kategorien aufweist, wird es als ordinal (sortiertes Set) umdefiniert.

Ausreißer-Trennwert. Geben Sie das in Standardabweichungen gemessene Ausreißer-Trennwert-Kriterium an. Der Standardwert ist 3.

Verfahren zum Ersetzen von Ausreißern Wählen Sie aus, ob Ausreißer durch Trimmen (Setzen) auf den Trennwert ersetzt werden sollen oder ob sie gelöscht und als fehlende Werte angegeben werden sollen. Jeder als fehlender Wert eingestufte Ausreißer unterliegt den oben ausgewählten Einstellungen für die Behandlung fehlender Werte.

Alle stetigen Eingabefelder auf eine gemeinsame Skala setzen. Um stetige Eingabefelder zu normalisieren, kreuzen Sie dieses Kontrollkästchen an und wählen das Normalisierungsverfahren aus. Standardmäßig ist die Z-Wert-Transformation eingestellt, bei der Sie den Endgültigen Mittelwert angeben können, der den Standardwert 0 hat, und die Endgültige Standardabweichung, die den Standardwert 1 hat. Alternativ können Sie die Min/Max-Transformation auswählen und die minimalen und maximalen Werte angeben, die standardmäßig auf 0 beziehungsweise 100 eingestellt sind.

Dieses Feld ist besonders nützlich, wenn Sie Funktionserstellung durchführen im Funktionsbereich "Erstellen& auswählen" angeben.

Neu Skalieren eines stetigen Ziels mit einer Box-Cox-Tranformation. Um ein stetiges (Skala oder Bereich) Zielfeld zu normalisieren, kreuzen Sie dieses Kontrollkästchen an. Bei der Box-Cox-Transformation sind standardmäßig die Werte 0 für den Endgültigen Mittelwert und 1 für die Endgültige Standardabweichung eingestellt.

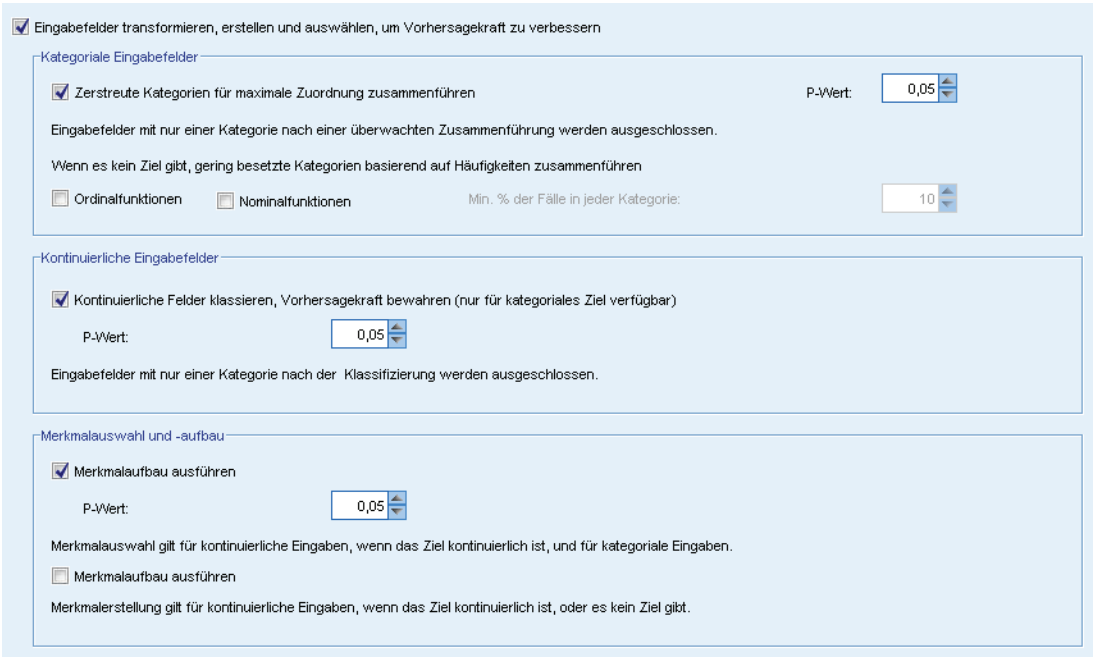
Anmerkung: Bei einer Normalisierung des Ziels wird die Dimension des Ziels transformiert. In diesem Fall müssen Sie u. U. einen Ableitungsknoten für die Anwendung einer inversen Transformation erstellen, um die transformierten Einheiten wieder in ein zur weiteren Verarbeitung erkennbares Format zu bringen. Für weitere Informationen siehe Thema Erzeugen eines Ableitungsknotens auf S. 134.

Auswahl von Erstellung und Funktion

Um die Vorhersagekraft Ihrer Daten zu verbessern, können Sie die Eingabefelder transformieren oder basierend auf den bestehenden Feldern neue erstellen.

Anmerkung: Wenn Sie die Werte in diesem Feld ändern, wird die Registerkarte Ziele automatisch auf die Auswahl der Option Benutzerdefinierte Analyse aktualisiert.

Abbildung 4-8
Automatisierte Datenaufbereitung – Transformation, Erstellung und Auswahl – Einstellungen



Transformieren, erstellen und auswählen von Eingabefeldern zur Verbessern der Vorhersagekraft. Schaltet alle Felder in dem Eingabefeld entweder an oder aus.

Dünn besetzte Kategorien zur Maximierung des Zielzusammenhangs zusammenführen. Mit dieser Option erstellen Sie ein sparsameres Modell, indem die Anzahl der zu verarbeitenden Variablen in Zusammenhang mit dem Ziel reduziert wird. Ändern Sie bei Bedarf den Wahrscheinlichkeitswert von der Standardeinstellung 0,05.

Hinweis: Wenn alle Kategorien zu einer verschmolzen werden, werden die originalen und abgeleiteten Versionen des Felds ausgeschlossen, da sie als Einflussgrößen keinen Wert haben.

Wenn kein Ziel existiert, dünn besetzte Kategorien auf der Basis folgender Häufigkeiten zusammenführen. Wenn Sie Daten verarbeiten, die kein Ziel aufweisen, können Sie auswählen, dünn besetzte Kategorien von ordinalen (sortiertes Set) oder nominalen (Set) Funktionen oder beiden zusammenzuführen. Geben Sie den minimalen Prozentsatz an Fällen oder Datensätzen in den Daten an, der die zusammenzuführenden Kategorien identifiziert. Der Standardwert ist 10.

Kategorien werden mithilfe der folgenden Regeln zusammengeführt:

- Das Zusammenführen erfolgt nicht bei binären Feldern.
- Wenn es bei der Zusammenführung nur zwei Kategorien gibt, stoppt die Zusammenführung.
- Wenn es keine originale Kategorie oder eine während des Zusammenführens erzeugte Kategorie gibt, die weniger als den angegebenen minimalen Prozentsatz an Fällen aufweist, stoppt die Zusammenführung.

Stetige Felder einteilen und gleichzeitig die Vorhersagekraft erhalten. Wenn die Daten ein kategoriales Ziel enthalten, können Sie stetige Eingaben mit starkem Zusammenhang einteilen, um die Verarbeitungsleistung zu verbessern. Ändern Sie bei Bedarf den Wahrscheinlichkeitswert für die homogenen Untergruppen von der Standardeinstellung 0,05.

Wenn in dem Klassierungsvorgang eine einzelne Klassierung für ein bestimmtes Feld durchgeführt wird, werden die Original- und eingeteilten Versionen des Felds ausgeschlossen, da sie keinen Wert als Einflussvariable aufweisen.

Anmerkung: Die Klassierung in ADP unterscheidet sich von der optimalen Klassierung, die in anderen Teilen von IBM® SPSS® Modeler verwendet wird. Bei der optimalen Klassierung werden Entropieinformationen verwendet, um eine stetige Variable in eine kategoriale Variable umzuwandeln; dazu müssen Daten sortiert und im Arbeitsspeicher abgelegt werden. ADP verwendet homogene Untergruppen zum Klassieren einer stetigen Variable, das bedeutet, dass die ADP-Klassierung keine Daten sortieren und im Arbeitsspeicher ablegen muss. Der Einsatz von homogenen Untergruppen zum Klassieren einer stetigen Variablen bedeutet, dass die Anzahl der Kategorien nach der Klassierung immer kleiner oder gleich der Anzahl der Kategorien des Ziels ist.

Funktionsauswahl durchführen. Wählen Sie diese Option, um Funktionen mit einem niedrigen Korrelationskoeffizienten zu entfernen. Ändern Sie bei Bedarf den Wahrscheinlichkeitswert von der Standardeinstellung 0,05.

Diese Option gilt nur für stetige Eingabefunktionen mit stetigem Ziel und kategoriale Eingabefunktionen.

Funktionserstellung durchführen. Wählen Sie diese Option aus, um neue Funktionen von einer Kombination aus mehreren bestehenden Funktionen abzuleiten (die in der Modellierung nicht weiter beachtet werden).

Diese Option gilt nur für stetige Eingabefunktionen mit stetigem Ziel oder Eingabefunktionen, in denen kein Ziel vorhanden ist.

Feldnamen

Abbildung 4-9
Automatisierte Datenaufbereitung – Namensfelder – Einstellungen

Zur einfachen Identifikation neuer und transformierter Funktionen erstellt ADP allgemeine neue Namen, Präfixe oder Suffixe und wendet diese an. Sie können diese Namen ändern und ihnen mehr Aussagekraft für Ihre eigenen Anforderungen und Daten geben. Andere Bezeichnungen müssen in einem nachgelagerten Typknoten angegeben werden.

Transformierte und erstellte Felder. Geben Sie die Namenserweiterungen an, die auf transformierte Ziel- und Eingabefelder angewendet werden sollen.

Beachten Sie, dass in einem String-Knoten die Einstellung leerer String-Felder je nach gewählter Behandlung nicht verwendeter Felder einen Fehler verursachen kann. Wenn Behandlung von aus der Modellierung ausgeschlossenen Feldern unter “Feldeinstellungen” auf der Registerkarte “Einstellungen” auf Nicht verwendete Felder filtern gesetzt ist, können die Namenserweiterungen für Eingaben und das Ziel auf “Nichts” gesetzt werden. Die Originalfelder werden durch Filterung ausgeschlossen und die transformierten Felder darüber gespeichert; in diesem Fall haben die transformierten Felder den gleichen Namen wie Ihr Original.

Wenn Sie jedoch Richtung nicht verwendeter Felder auf “Keine” setzen auswählen, werden leere (oder Null-) Namenserweiterungen für das Ziel und die Eingaben einen Fehler verursachen, weil Sie versuchen, doppelt vorhandene Feldnamen zu erstellen.

Geben Sie außerdem über die Einstellungen “Auswählen und erstellen” den Präfixnamen an, der auf erstellte Funktionen angewendet werden soll. Der neue Name wird erstellt, indem ein numerisches Suffix an diesen Präfix-Stammnamen angehängt wird. Das Zahlenformat hängt davon ab, wie viele neue Funktionen abgeleitet werden, zum Beispiel:

- Es werden 1-9 erstellte Funktionen benannt: Funktion1 bis Funktion9.

- 10 – 99 konstruierte Merkmale werden benannt: Funktion01 bis Funktion99.
- 100 – 999 konstruierte Merkmale werden benannt: Funktion001 bis Funktion999 usw.

So wird gewährleistet, dass die erstellten Funktionen ungeachtet ihrer Anzahl in einer vernünftigen Reihenfolge sortiert werden.

Aus Datums- und Zeitangaben berechnete Dauerzeiten. Geben Sie die Namenserweiterungen an, die auf die aus Datums- und Zeitangaben berechnete Dauer angewendet werden sollen.

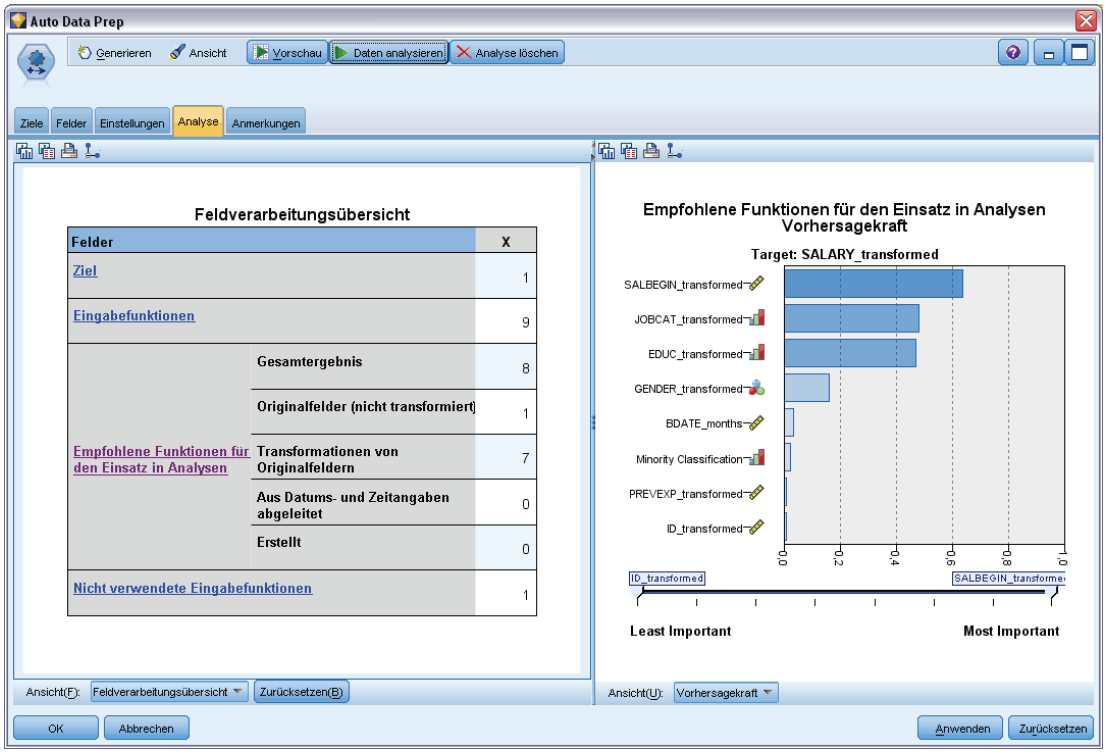
Aus Datums- und Zeitangaben extrahierte zyklische Elemente. Geben Sie die Namenserweiterungen an, die auf die aus Datums- und Zeitangaben extrahierten zyklischen Elemente angewendet werden sollen.

Registerkarte "Analyse"

▶ Wenn Sie mit den ADP-Einstellungen einschließlich aller in den Registerkarten "Ziel", "Felder" und "Einstellungen" vorgenommenen Änderungen zufrieden sind, klicken Sie auf Daten analysieren. Der Algorithmus wendet die Eingabedaten an und zeigt die Ergebnisse auf der Registerkarte "Analyse" an.

Die Registerkarte "Analyse" enthält Ausgaben in Grafik- und Tabellenform, die die Verarbeitung Ihrer Daten zusammenfassen, und zeigt Empfehlungen an, wie die Daten möglicherweise bearbeitet oder zum Scoring verbessert werden können. Anschließend können Sie diese Empfehlungen überprüfen und entweder akzeptieren oder ablehnen.

Abbildung 4-10
Registerkarte "Analyse" in der automatisierten Datenaufbereitung



Die Registerkarte “Analyse” besteht aus zwei Bereichen, der Hauptansicht im linken Bereich und der verknüpften oder Hilfsansicht im rechten Bereich. Es gibt drei Hauptansichten:

- Feldverarbeitungsübersicht (Standard). Für weitere Informationen siehe Thema Feldverarbeitungsübersicht auf S. 123.
- Felder. Für weitere Informationen siehe Thema Felder auf S. 124.
- Aktionsübersicht. Für weitere Informationen siehe Thema Aktionsübersicht auf S. 126.

Es gibt vier verknüpfte/Hilfsansichten:

- Vorhersagekraft (Standard). Für weitere Informationen siehe Thema Vorhersagekraft auf S. 127.
- Feldertabelle. Für weitere Informationen siehe Thema Feldertabelle auf S. 128.
- Felddetails. Für weitere Informationen siehe Thema Felddetails auf S. 129.
- Aktionsdetails. Für weitere Informationen siehe Thema Aktionsdetails auf S. 131.

Verknüpfungen zwischen Ansichten

In der Hauptansicht steuert unterstrichener Text in den Tabellen die Anzeige in der verknüpften Ansicht. Wenn Sie auf den Text klicken, erhalten Sie Informationen über ein bestimmtes Feld, ein Set von Feldern oder einen Verarbeitungsschritt. Der zuletzt von Ihnen ausgewählte Link wird in einer dunkleren Farbe angezeigt; dies hilft Ihnen dabei, die Verbindung zwischen den Inhalten der beiden Ansichtsbereiche zu identifizieren.

Zurücksetzen der Ansichten

Klicken Sie auf Zurücksetzen im unteren Bereich der Hauptansicht, um die ursprünglichen Empfehlungen der Analyse erneut anzuzeigen und alle in den Analyseansichten vorgenommenen Änderungen rückgängig zu machen.

Feldverarbeitungsübersicht

Abbildung 4-11
Feldverarbeitungsübersicht

Feldverarbeitungsübersicht

Felder		X
Ziel		1
Eingabefunktionen		9
Empfohlene Funktionen für den Einsatz in Analysen	Gesamtergebnis	8
	Originalfelder (nicht transformiert)	1
	Transformationen von Originalfeldern	7
	Aus Datums- und Zeitangaben abgeleitet	0
	Erstellt	0
Nicht verwendete Eingabefunktionen		1

Die Tabelle “Feldverarbeitungsübersicht” gibt Ihnen eine Momentaufnahme des projizierten Gesamteinflusses der Verarbeitung, einschließlich Änderungen des Status der Funktionen und der Anzahl der erstellten Funktionen.

Beachten Sie, dass dabei kein Modell erstellt wird und somit kein Maß oder keine Grafik der Veränderung der Gesamtvorhersagekraft vor und nach der Datenaufbereitung vorhanden ist; Sie können stattdessen Grafiken der Vorhersagekraft einzelner empfohlener Einflussvariablen anzeigen.

Die Tabelle zeigt folgende Informationen an:

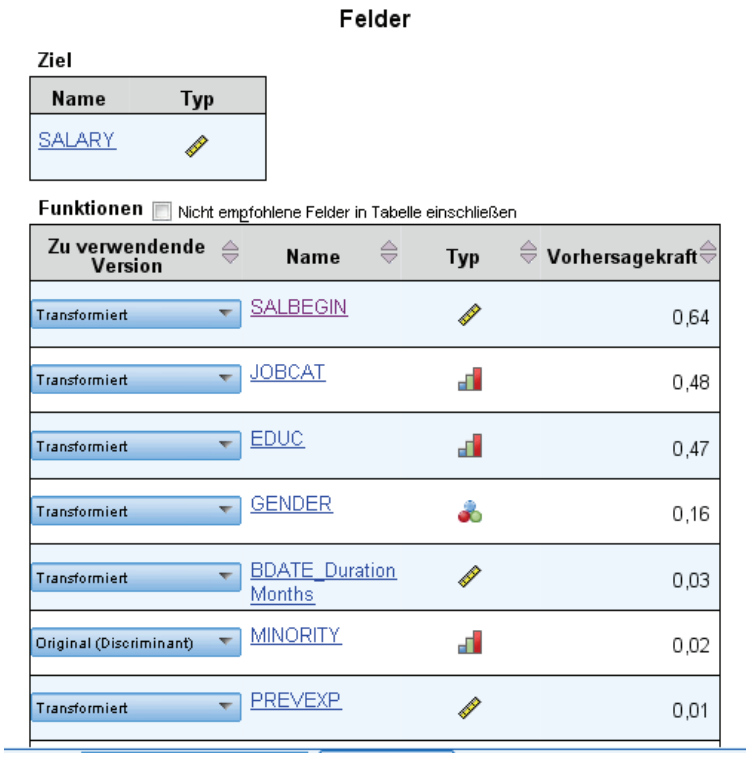
- Die Anzahl der Zielfelder.
- Die Anzahl der ursprünglichen (Eingabe-)Prädiktoren.
- Die zur Verwendung in der Analyse und Modellierung empfohlenen Prädiktoren. Dazu gehören die Gesamtzahl empfohlener Felder, die Anzahl der ursprünglichen, nicht transformierten empfohlenen Felder, die Anzahl transformierter empfohlener Felder (außer Zwischenversionen, von Datums- und Zeitprädiktoren abgeleitete Felder und erstellte Prädiktoren), die Anzahl der von Datums- und Zeitfelder abgeleiteten empfohlenen Felder sowie die Anzahl erstellter empfohlener Prädiktoren.
- Die Anzahl der Eingabeprädiktoren, die in keiner Form empfohlen werden, sei es in ihrer ursprünglichen Form, als abgeleitetes Feld oder als Eingabe in einem erstellten Prädiktor.

Klicken Sie auf die unterstrichenen Informationen unter Felder, um weitere Informationen in einer verknüpften Ansicht anzuzeigen. In der verknüpften Ansicht “Feldertabelle” erhalten Sie Informationen über Ziel, Eingabefunktionen und Nicht verwendete Eingabefunktionen. Für weitere

Informationen siehe Thema Feldertabelle auf S. 128. Empfohlene Funktionen für den Einsatz in Analysen werden in der verknüpften Ansicht “Vorhersagekraft” angezeigt. Für weitere Informationen siehe Thema Vorhersagekraft auf S. 127.

Felder

Abbildung 4-12
Fields



In der Hauptansicht “Felder” werden die verarbeiteten Felder angezeigt sowie, ob ADP diese zur Verwendung in nachgelagerten Modellen empfiehlt. Sie können die Empfehlung für jedes Feld überschreiben, zum Beispiel, um erstellte Funktionen auszuschließen oder Funktionen einzuschließen, von denen ADP empfiehlt, sie auszuschließen. Wenn ein Feld transformiert wurde, können Sie entscheiden, ob Sie die vorgeschlagene Transformation akzeptieren oder die Originalversion verwenden möchten.

Die Felderansicht besteht aus zwei Tabellen, eine für das Ziel und eine für Prädiktoren, die entweder verarbeitet oder erstellt wurden.

Tabelle “Ziel”

Die Tabelle Ziel wird nur angezeigt, wenn in den Daten ein Ziel definiert wurde.

Die Tabelle enthält zwei Spalten:

- **Name.** Dies ist der Name oder die Bezeichnung des Zielfelds. Der Originalname wird immer verwendet, auch wenn das Feld transformiert wurde.
- **Messniveau.** Hier erscheint das Symbol für das entsprechende Messniveau; fahren Sie mit der Maus über das Symbol, um eine Bezeichnung (stetig, ordinal, nominal usw.) anzuzeigen, die die Daten beschreibt.

  Wenn das Ziel transformiert wurde, gibt die Spalte Messniveau die endgültige transformierte Version an. *Anmerkung*: Transformationen für das Ziel können nicht abgeschaltet werden.

Tabelle "Prädiktoren"

Die Tabelle Prädiktoren wird immer angezeigt. Jede Zeile der Tabelle repräsentiert ein Feld. Standardmäßig sind die Zeilen nach absteigender Vorhersagekraft sortiert.

Bei gewöhnlichen Funktionen wird der Originalname immer als Zeilenname verwendet. Sowohl Original- als auch abgeleitete Versionen von Datums-/Zeitfeldern werden in der Tabelle (in getrennten Zeilen) angezeigt; die Tabelle enthält auch erstellte Prädiktoren.

Beachten Sie, dass transformierte Versionen von in der Tabelle angezeigten Feldern immer die Endversionen darstellen.

Standardmäßig werden in der Tabelle "Prädiktoren" nur empfohlene Felder angezeigt. Um die restlichen Felder anzuzeigen, wählen Sie das Feld Nicht empfohlene Felder in Tabelle einschließen über der Tabelle aus; diese Felder werden dann am Ende der Tabelle angezeigt.

Die Tabelle enthält folgende Spalten:

- **Zu verwendende Version.** Hier wird eine Dropdown-Liste angezeigt, die festlegt, ob ein Feld nachgelagert verwendet wird oder ob die vorgeschlagenen Transformationen verwendet werden sollen. Standardmäßig werden in der Dropdown-Liste die Empfehlungen wiedergegeben.

  Für gewöhnliche Prädiktoren, die transformiert wurden, stehen in der Dropdown-Liste drei Optionen zur Auswahl: Transformiert, Original und Nicht verwenden.

  Für nicht transformierte gewöhnliche Prädiktoren sind folgende Auswahlmöglichkeiten verfügbar: Original und Nicht verwenden.

  Für abgeleitete Datums-/Zeitfelder und erstellte Prädiktoren sind folgende Auswahlmöglichkeiten verfügbar: Transformiert und Nicht verwenden.

  Für Original-Datumsfelder ist die Dropdown-Liste deaktiviert und auf Nicht verwenden gesetzt.

  Anmerkung: Für Prädiktoren mit Original- und transformierten Versionen werden bei einem Wechsel zwischen den Versionen Original und Transformiert automatisch die Einstellungen Messniveau und Vorhersagekraft für diese Funktionen aktualisiert.
- **Name.** Jeder Feldname ist ein Link. Klicken Sie auf den Namen, um in der verknüpften Ansicht weitere Informationen über das Feld anzuzeigen. Für weitere Informationen siehe Thema Felddetails auf S. 129.

- **Messniveau.** Hier erscheint das Symbol für den entsprechenden Datentyp; fahren Sie mit der Maus über das Symbol, um eine Bezeichnung (stetig, ordinal, nominal usw.) anzuzeigen, die die Daten beschreibt.
- **Vorhersagekraft.** Die Vorhersagekraft wird nur für Felder angezeigt, die von ADP empfohlen werden. Diese Spalte wird nicht angezeigt, wenn kein Ziel definiert wurde. Die Vorhersagekraft reicht von 0 bis 1, wobei größere Werte "bessere" Einflussgrößen andeuten. Im Allgemeinen ist die Vorhersagekraft für den Vergleich von Einflussgrößen in einer ADP-Analyse nützlich, doch sollten Vorhersagekraft-Werte nicht in Analysen verglichen werden.

Aktionsübersicht

Abbildung 4-13
Aktionsübersicht

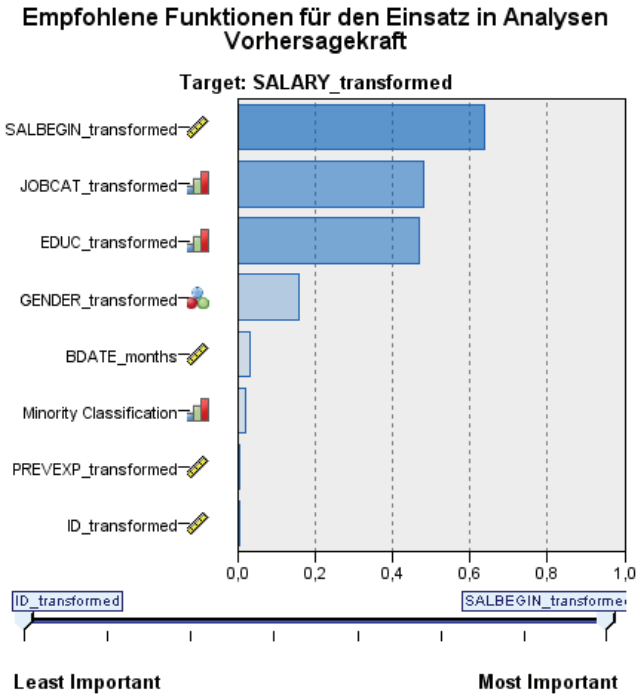
Bei jeder von der automatisierten Datenaufbereitung vorgenommenen Aktion werden Eingabeprädiktoren transformiert und/oder herausgefiltert. Felder, die in einer Aktion erhalten bleiben, werden in der nächsten verwendet. Die Felder, die bis zum letzten Schritt erhalten bleiben, werden dann für die Modellierung empfohlen, während Eingaben zu transformierten und erstellten Prädiktoren durch Filterung ausgeschlossen werden.

Die Aktionsübersicht ist eine einfache Tabelle, in der die von der ADP vorgenommenen Verarbeitungsaktionen aufgelistet sind. Klicken Sie auf den unterstrichenen Link Aktion, um in einer verknüpften Ansicht weitere Informationen über die durchgeführten Schritte anzuzeigen. Für weitere Informationen siehe Thema Aktionsdetails auf S. 131.

Anmerkung: Es werden nur die Original- und endgültigen transformierten Versionen jedes Felds angezeigt, jedoch keine während der Analyse verwendeten Zwischenversionen.

Vorhersagekraft

Abbildung 4-14
Vorhersagekraft



Wird standardmäßig bei der ersten Ausführung der Analyse angezeigt. Wenn Sie dagegen Empfohlene Prädiktoren für den Einsatz in Analysen in der Hauptansicht "Feldverarbeitungsübersicht" auswählen, zeigt das Diagramm die Vorhersagekraft der empfohlenen Prädiktoren an. Felder werden nach Vorhersagekraft sortiert, wobei das Feld mit dem höchsten Wert zuerst erscheint.

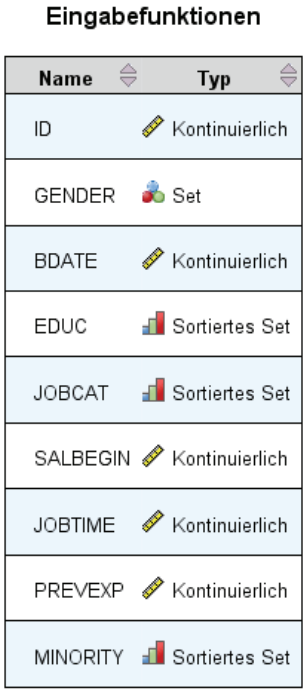
Bei transformierten Versionen gewöhnlicher Prädiktoren gibt der Feldname Ihre Suffixauswahl im Bereich "Feldnamen" auf der Registerkarte "Einstellungen" an, z. B.: *_transformiert*.

Die Symbole für das Messniveau werden nach den einzelnen Feldnamen angezeigt.

Die Vorhersagekraft jedes empfohlenen Prädiktors wird entweder aus einer linearen Regression oder einem Naive Bayes-Modell berechnet, abhängig davon, ob das Ziel stetig oder kategorial ist.

Feldertabelle

Abbildung 4-15
Feldertabelle



Die Feldertabelle wird angezeigt, wenn Sie in der Hauptansicht “Feldverarbeitungsübersicht” auf Ziel, Prädiktoren oder Nicht verwendete Prädiktoren klicken, und enthält eine einfache Tabelle, die die wichtigsten Funktionen auflistet.

Die Tabelle enthält zwei Spalten:

- **Name.** Der Name des Prädiktors.

  Für Ziele wird der Originalname oder die Originalbeschriftung des Felds verwendet, selbst wenn das Ziel transformiert wurde.

  Bei transformierten Versionen gewöhnlicher Prädiktoren gibt der Name Ihre Suffixauswahl im Bereich “Feldnamen” auf der Registerkarte “Einstellungen” an, z. B.: *_transformiert*.

  Bei aus Datums- und Zeitangaben abgeleiteten Feldern wird der Name der endgültigen transformierten Version verwendet, z. B.: *bdatum_Jahre*.

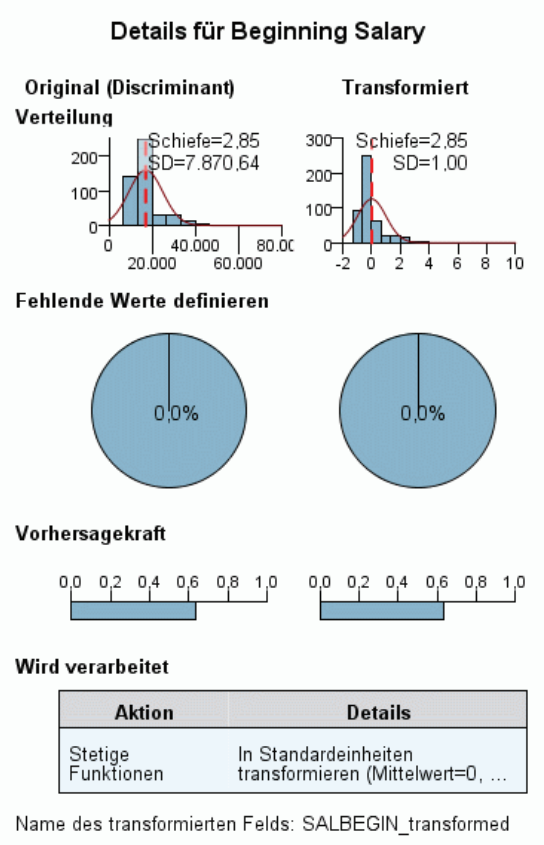
  Bei erstellten Prädiktoren wird der Name des erstellten Prädiktors verwendet, z. B.: *Prädiktor1*.

- **Messniveau.** Hier erscheint das Symbol für den entsprechenden Datentyp.

  Für das Ziel gibt das Messniveau stets die transformierte Version wieder (wenn das Ziel transformiert wurde), z. B. bei einem Wechsel von ordinal (sortiertes Set) zu stetig (Bereich, Skala) oder umgekehrt.

Felddetails

Abbildung 4-16
Felddetails



Die Ansicht “Felddetails” wird angezeigt, wenn Sie auf Name in der Hauptansicht “Felder” klicken, und enthält Informationen über Verteilung, fehlende Werte und (falls zutreffend) Vorhersagekraft-Diagramme für das ausgewählte Feld. Außerdem wird der Verarbeitungsverlauf für das Feld und der Name des transformierten Felds angezeigt (falls zutreffend).

Für jedes Diagramm-Set werden nebeneinander zwei Versionen angezeigt, um das Feld mit und ohne angewendete Transformationen zu vergleichen. Wenn keine transformierte Version des Felds vorhanden ist, wird nur ein Diagramm für die Originalversion angezeigt. Für abgeleitete Datums- und Zeitfelder sowie erstellte Prädiktoren werden die Diagramme nur für den neuen Prädiktor angezeigt.

Anmerkung: Wenn ein Feld wegen zu vieler Kategorien ausgeschlossen wurde, wird nur der Verarbeitungsverlauf angezeigt.

Verteilungsdiagramm

Die Verteilung stetiger Felder wird als Histogramm angezeigt, mit einer überlagerten Normalverteilungskurve und einer vertikalen Referenzlinie für den Mittelwert; kategoriale Felder werden als Balkendiagramm angezeigt.

Die Histogramme werden nach Standardabweichung und Schiefe bezeichnet, allerdings wird Letztere nicht angezeigt, wenn die Anzahl der Werte kleiner gleich 2 oder die Varianz des originalen Felds kleiner als 10-20 ist.

Fahren Sie mit der Maus über das Diagramm, um entweder den Mittelwert für Histogramme oder die Zählung und den Prozentsatz der Gesamtzahl der Datensätze für Kategorien in Balkendiagrammen anzuzeigen.

Diagramm fehlender Werte

Kreisdiagramme vergleichen den Prozentsatz fehlender Werte mit und ohne angewendete Transformationen; die Diagrammbeschriftungen zeigen den Prozentsatz an.

Wenn ADP die Behandlung fehlender Werte durchgeführt hat, enthält das Kreisdiagramm nach der Transformation auch den Ersatzwert als Beschriftung, d. h. den anstelle von fehlenden Werten verwendeten Wert.

Fahren Sie mit der Maus über das Diagramm, um die Zählung der fehlenden Werte und den Prozentsatz der Gesamtzahl an Datensätzen anzuzeigen.

Vorhersagekraft-Diagramme

Für empfohlene Felder zeigen Balkendiagramme die Vorhersagekraft vor und nach der Transformation an. Wenn das Ziel transformiert wurde, steht die berechnete Vorhersagekraft in Beziehung zum transformierten Ziel.

Anmerkung: Die Vorhersagekraft-Diagramme werden nicht angezeigt, wenn kein Ziel definiert wurde oder wenn Sie in der Hauptansicht auf das Ziel klicken.

Fahren Sie mit der Maus über das Diagramm, um den Wert der Vorhersagekraft anzuzeigen.

Tabelle "Verarbeitungsverlauf"

Die Tabelle zeigt, wie die transformierte Version eines Felds abgeleitet wurde. Von ADP durchgeführte Aktionen werden in der Reihenfolge ihrer Ausführung aufgelistet. Bei bestimmten Schritten wurden jedoch u. U. mehrere Aktionen für ein spezielles Feld durchgeführt.

Anmerkung: Die Tabelle wird nur für transformierte Felder angezeigt.

Die Informationen in der Tabelle erscheinen in zwei oder drei Spalten:

- **Aktion.** Der Name der Aktion. Zum Beispiel "Stetige Prädiktoren". Für weitere Informationen siehe Thema Aktionsdetails auf S. 131.

- **Details.** Die Liste der durchgeführten Verarbeitung. Zum Beispiel “Zu Standardeinheiten transformieren”.
- **Funktion.** Diese Spalte erscheint nur bei erstellten Prädiktoren und zeigt die lineare Kombination von Eingabefeldern an, z. B. 0,06*Alter + 1,21*Größe.

Aktionsdetails

Abbildung 4-17
ADP-Analyse – Aktionsdetails

Schritt 9: Stetige Funktionen

Transformation	Anzahl der Funktionen	Kriterien	
		Mittelwert	SD
In Standardeinheiten transformieren	5	0	1

Erstellung eines Funktionsbereichs	X
Erstellte Funktionen	0
Funktionen, die wegen niedrigem Zielzusammenhang ausgeschlossen wurden	1
Funktionen, die ausgeschlossen wurden, weil sie nach der Einteilung konstant waren.	0

Die verknüpfte Ansicht “Aktionsdetails” wird angezeigt, wenn Sie in der Hauptansicht “Aktionsübersicht” auf den unterstrichenen Link Aktion klicken, und enthält sowohl aktionsspezifische als auch allgemeine Informationen über jeden durchgeführten Verarbeitungsschritt. Die aktionsspezifischen Informationen erscheinen stets zuerst.

Für jede Aktion wird die Beschreibung als Titel im oberen Bereich der verknüpften Ansicht verwendet. Die aktionsspezifischen Informationen erscheinen unter dem Titel und enthalten u. U. Details zur Anzahl abgeleiteter Prädiktoren, zu umgewandelten Feldern, zu Zieltransformationen, zu zusammengeführten oder neu sortierten Kategorien und zu erstellten oder ausgeschlossenen Prädiktoren.

Bei der Verarbeitung jeder Aktion kann sich die für die Verarbeitung verwendete Anzahl an Prädiktoren ändern, wenn beispielsweise Prädiktoren ausgeschlossen oder zusammengeführt werden.

Anmerkung: Wenn eine Aktion deaktiviert oder kein Ziel angegeben wurde, erscheint eine Fehlermeldung anstelle der Aktionsdetails, wenn Sie in der Hauptansicht “Aktionsübersicht” auf die Aktion klicken.

Es gibt neun mögliche Aktionen, davon sind allerdings nicht alle notwendigerweise für jede Analyse aktiv.

Tabelle "Textfelder"

Die Tabelle zeigt folgende Anzahl:

- Entfernte leere nachstehende Werte.
- Von der Analyse ausgeschlossene Prädiktoren.

Tabelle "Datums- und Uhrzeitprädiktoren"

Die Tabelle zeigt folgende Anzahl:

- Aus Datums- und Uhrzeitprädiktoren abgeleitete Dauer.
- Datums- und Uhrzeitelemente.
- Insgesamt abgeleitete Datums- und Uhrzeitprädiktoren.

Das Referenzdatum oder die -uhrzeit wird als Fußnote angezeigt, wenn eine Datumsdauer berechnet wurde.

Tabelle "Screening von Prädiktoren"

Die Tabelle zeigt die Anzahl folgender von der Verarbeitung ausgeschlossener Prädiktoren:

- Konstanten.
- Prädiktoren mit zu vielen fehlenden Werten.
- Prädiktoren mit zu vielen Fällen in einer einzelnen Kategorie.
- Nominale Felder (Sets) mit zu vielen Kategorien.
- Insgesamt ausgeschlossene Prädiktoren.

Tabelle "Messniveau überprüfen"

Die Tabelle zeigt die Anzahl umgewandelter Felder und teilt sich wie folgt auf:

- In stetige Feldern umgewandelte ordinale Felder (sortierte Sets).
- In ordinale Felder umgewandelte stetige Felder.
- Anzahl an Umwandlungen insgesamt.

Wenn keine Eingabefelder (Ziel oder Prädiktoren) stetig oder ordinal waren, wird dies in einer Fußnote angezeigt.

Tabelle "Ausreißer"

Die Tabelle zeigt, ob und wie Ausreißer behandelt wurden.

- Entweder die Anzahl stetiger Felder, für die Ausreißer gefunden und entfernt wurden, oder die Anzahl stetiger Felder, für die Ausreißer gefunden und als fehlend eingestuft wurden, je nach Ihren Einstellungen im Feld “Eingaben & Ziel vorbereiten” auf der Registerkarte “Einstellungen”.
- Die Anzahl stetiger Felder, die ausgeschlossen wurden, weil sie nach der Ausreißer-Behandlung konstant waren.

Der Ausreißer-Trennwert wird in einer Fußnote vermerkt. Eine weitere Fußnote wird angezeigt, wenn keine Eingabefelder (Ziel oder Prädiktoren) stetig waren.

Tabelle “Fehlende Werte”

Die Tabelle zeigt die Anzahl an Feldern, in denen fehlende Werte ersetzt wurden, und teilt sich wie folgt auf:

- Ziel. Diese Zeile wird nicht angezeigt, wenn kein Ziel angegeben wurde.
- Prädiktoren. Dies teilt sich weiter auf in Anzahl an “nominal (Set)”, “ordinal (sortiertes Set)” und “stetig”.
- Die gesamte Anzahl ersetzter fehlender Werte.

Tabelle “Ziel”

Die Tabelle zeigt wie folgt, ob das Ziel tranformiert wurde:

- Box-Cox-Transformation in Normalverteilung. Dies teilt sich weiter in Spalten auf, die die angegebenen Kriterien (Mittelwert und Standardabweichung) und Lambda zeigen.
- Zielkategorien zur Verbesserung der Stabilität neu sortiert.

Tabelle “Kategoriale Prädiktoren”

Die Tabelle zeigt folgende Anzahl kategorialer Prädiktoren:

- Wessen Kategorien wurden zur Verbesserung der Stabilität in aufsteigender Reihenfolge neu sortiert.
- Wessen Kategorien wurden zur Maximierung des Zielzusammenhangs zusammengeführt.
- Wessen Kategorien wurden zur Behandlung dünn besetzter Kategorien zusammengeführt.
- Wegen niedrigem Zielzusammenhang ausgeschlossen.
- Ausgeschlossen, weil nach der Zusammenführung konstant.

Wenn es keine kategorialen Prädiktoren gab, wird dies durch eine Fußnote vermerkt.

Tabelle “Stetige Prädiktoren”

Es gibt zwei Tabellen. Die erste zeigt eine der folgenden Transformationen:

- Zu Standardeinheiten transformierte Prädiktorwerte. Zusätzlich werden hier die Anzahl transformierter Prädiktoren, der angegebene Mittelwert und die Standardabweichung angezeigt.

- Einem gemeinsamen Bereich zugeordnete Prädiktorwerte. Zusätzlich werden hier die Anzahl der mithilfe der min./max. Transformation transformierten Prädiktoren sowie die angegebenen Mindest- und Höchstwerte angezeigt.
- Klassierte Prädiktorwerte und die Anzahl klassierter Prädiktoren.

Die zweite Tabelle enthält Informationen über die Erstellung von Prädiktorbereichen, die als Anzahl folgender Prädiktoren angezeigt werden:

- Erstellt.
- Wegen niedrigem Zielzusammenhang ausgeschlossen.
- Ausgeschlossen, weil nach der Klassierung konstant.
- Ausgeschlossen, weil nach der Erstellung konstant.

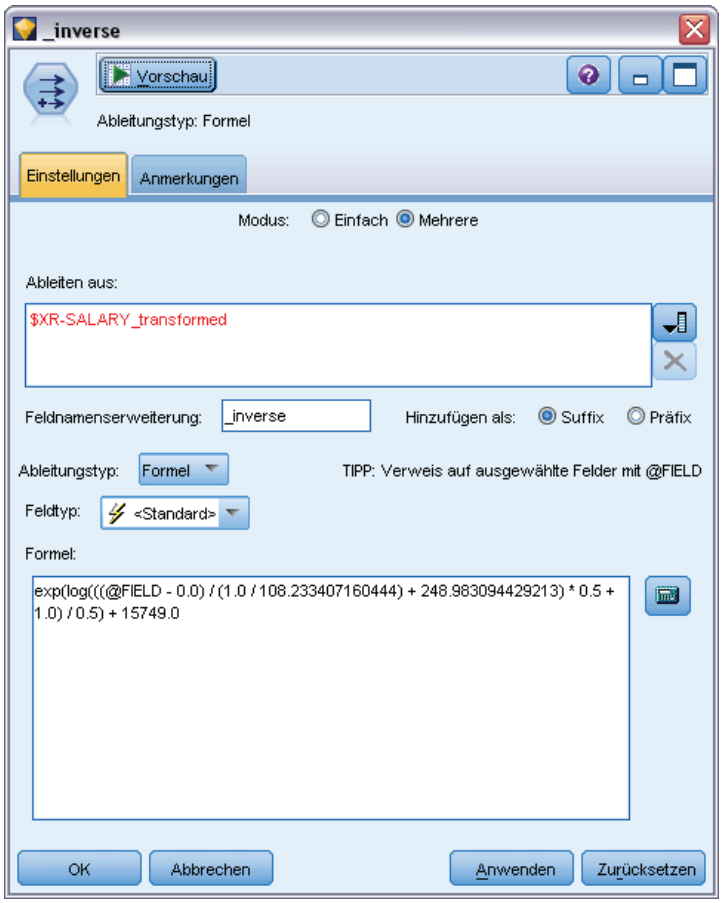
Wenn keine stetigen Prädiktoren eingegeben wurden, wird dies durch eine Fußnote vermerkt.

Erzeugen eines Ableitungsknotens

Wenn Sie einen Ableitungsknoten erstellen, wendet dieser die inverse Zieltransformation auf das Wertfeld an. Standardmäßig gibt der Knoten den Namen des Wertfelds ein, das mithilfe eines Automodeler-Knotens (zum Beispiel Auto Classifier oder Auto Numeric) oder dem Ensemble-Knoten erstellt werden würde. Wenn ein metrisches (Bereichs-)Ziel transformiert wurde, wird das Wertfeld in transformierten Einheiten angezeigt, zum Beispiel log($) anstelle von $. Um die Ergebnisse interpretieren und verwenden zu können, müssen Sie den vorhergesagten Wert wieder in das ursprüngliche metrische Maß zurückkonvertieren.

Anmerkung: Sie können einen Ableitungsknoten nur dann erzeugen, wenn der ADP-Knoten eine Analyse enthält, die ein Bereichsziel neu skaliert (d. h. die Box-Cox-Neuskalierung ist im Feld “Eingaben & Ziel vorbereiten” ausgewählt). Sie können keinen Ableitungsknoten erzeugen, wenn das Ziel kein Bereich ist oder wenn die Box-Cox-Neuskalierung nicht ausgewählt ist.

Abbildung 4-18
Aus dem Knoten "Automatisierte Datenaufbereitung" erzeugter Ableitungsknoten



Der Ableitungsknoten wird im Mehrfachmodus erstellt und verwendet @FIELD im Ausdruck, damit Sie das transformierte Ziel gegebenenfalls hinzufügen können. Es können zum Beispiel folgende Informationen verwendet werden:

- Zielfeldname: Antwort
- Feldname des transformierten Ziels: Antwort_transformiert
- Wertfeldname: $XR-Antwort_transformiert

Der Ableitungsknoten würde folgendes neues Feld erstellen: $XR-Antwort_transformiert_invers.

Anmerkung: Wenn Sie keinen Automodeler- oder Ensemble-Knoten verwenden, müssen Sie den Ableitungsknoten so bearbeiten, dass dieser das korrekte Wertfeld für Ihr Modell transformiert.

Normalisierte stetige Ziele

Wenn Sie das Kontrollkästchen Stetiges Ziel mit einer Box-Cox-Transformation neu skalieren im Feld "Eingaben & Ziel vorbereiten" auswählen, wird dadurch standardmäßig das Ziel transformiert und Sie können ein neues Feld erstellen, das für Ihre Modellerstellung als Ziel fungiert. Wenn zum

Beispiel Ihr ursprüngliches Ziel *Antwort* war, heißt das neue Ziel *Antwort_transformiert*; dem ADP-Knoten nachgelagerte Modelle nehmen dieses Ziel automatisch auf.

Dies kann jedoch je nach ursprünglichem Ziel Probleme verursachen. Wenn das Ziel zum Beispiel *Alter* war, werden die Werte des neuen Ziels nicht *Jahre*, sondern eine transformierte Version *Jahre* sein. Somit können Sie die Werte nicht betrachten und interpretieren, da sie nicht in erkennbaren Einheiten vorliegen. In diesem Fall können Sie eine inverse Transformation anwenden, durch die Ihre transformierten Einheiten wieder in ihr ursprünglich gewünschtes Format zurückkonvertiert werden. Gehen Sie dazu wie folgt vor:

- Klicken Sie zunächst auf Daten analysieren, um die ADP-Analyse auszuführen, und wählen Sie dann *Ableitungsknoten* im Menü *Erzeugen* aus.
- Setzen Sie den Ableitungsknoten nach Ihrem Nugget auf der Modellzeichenfläche.

Der Ableitungsknoten stellt die ursprünglichen Dimensionen des Wertfelds wieder her, sodass die Vorhersage in den ursprünglichen Werten *Jahre* erscheint.

Der Ableitungsknoten transformiert standardmäßig das durch ein Automodeler- oder Ensemble-Modell erzeugte Wertfeld. Wenn Sie ein einzelnes Modell erstellen, müssen Sie den Ableitungsknoten so bearbeiten, dass dieser aus Ihrem tatsächlichen Wertfeld ableitet. Wenn Sie Ihr Modell evaluieren möchten, sollten Sie das transformierte Ziel dem Feld Ableiten aus im Ableitungsknoten hinzufügen. Dadurch wird die gleiche inverse Transformation auf das Ziel angewendet und jeder nachgelagerte Evaluierungs- oder Analyse-Knoten wird die transformierten Daten korrekt verwenden, vorausgesetzt, Sie stellen diese Knoten so ein, dass sie Feldnamen anstelle von Metadaten verwenden.

Wenn Sie zudem den Originalnamen wiederherstellen möchten, können Sie einen Filter-Knoten verwenden, um das eventuell noch vorhandene ursprüngliche Zielfeld zu entfernen, und die Ziel- und Wertfelder umbenennen.

Typknoten

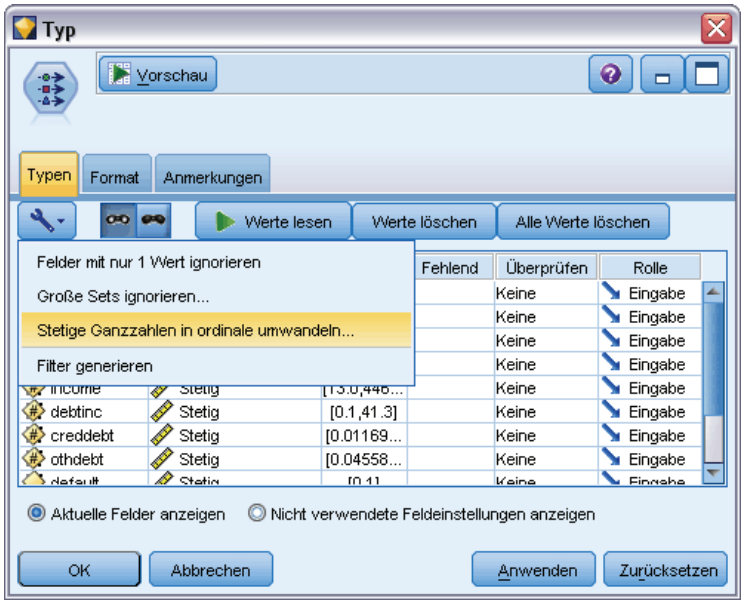
Die Feldeigenschaften können in einem Quellenknoten oder in einem separaten Typknoten angegeben werden. Die Funktionsweise ist bei beiden Knoten ähnlich. Folgende Eigenschaften stehen zur Verfügung:

- **Feld.** Doppelklicken Sie auf einen beliebigen Feldnamen, um Werte- und Feldbeschriftungen für Daten in IBM® SPSS® Modeler anzugeben. So können aus IBM® SPSS® Statistics beispielsweise importierte Feldmetadaten hier angezeigt oder geändert werden. Auf ähnliche Weise können Sie auch neue Beschriftungen für Felder und ihre Werte erstellen. Die Beschriftungen, die Sie hier angeben, werden überall in SPSS Modeler angezeigt, je nach der von Ihnen im Dialogfeld “Stream-Eigenschaften” getroffenen Auswahl.
- **Messung.** Dies ist das Messniveau, das zur Beschreibung der Eigenschaften von Daten in einem bestimmten Feld verwendet wird. Wenn alle Details eines Felds bekannt sind, wird es als **vollständig instanziiert** bezeichnet. Für weitere Informationen siehe Thema Messniveaus auf S. 138.

Anmerkung: Das Messniveau eines Felds ist etwas anderes als sein Speichertyp, der angibt, ob die Daten als Zeichenkette, ganze Zahl, reelle Zahl, Datum, Zeit oder Zeitstempel gespeichert werden sollen.

- **Werte.** In dieser Spalte können Sie Optionen zum Lesen von Datenwerten aus dem Daten-Set auswählen oder die Option Angeben verwenden, um Messniveaus und Werte in einem separaten Dialogfeld anzugeben. Sie können auch Felder übergeben, ohne ihre Werte zu lesen. Für weitere Informationen siehe Thema Datenwerte auf S. 143.
- **Fehlend.** Wird verwendet, um anzugeben, wie fehlende Werte für das Feld behandelt werden. Für weitere Informationen siehe Thema Fehlende Werte definieren auf S. 149.
- **Überprüfen.** In dieser Spalte können Sie Optionen festlegen, um sicherzustellen, dass die Feldwerte den angegebenen Werten oder Bereichen entsprechen. Für weitere Informationen siehe Thema Überprüfen von Typenwerten auf S. 149.
- **Rolle.** Wird verwendet, um Modellierungsknoten mitzuteilen, ob es sich bei Feldern um Eingabefelder (Prädiktorfelder) oder Zielfelder (vorhergesagte Felder) für einen Maschinenlernprozess handelt. Beides und Keine sind auch verfügbare Rollen, zusammen mit Partition, das ein Feld bezeichnet, das für die Aufteilung von Datensätzen in separate Stichproben zu Training-, Test- und Validierungszwecken verwendet wird. Der Wert Aufteilung gibt an, dass für jeden möglichen Wert des Felds separate Modelle erstellt werden. Für weitere Informationen siehe Thema Festlegen der Feldrolle auf S. 150.

Abbildung 4-19
Typknotenoptionen



Mehrere andere Optionen können im Fenster “Typknoten” angegeben werden:

- Mithilfe der Schaltfläche Felder mit nur 1 Wert ignorieren im Menü “Extras” können Sie festlegen, dass Felder mit nur einem Wert ignoriert werden sollen, sobald ein Typknoten als Instanz erstellt wurde (entweder über Ihre Spezifikationen, aus eingelesenen Werten oder durch die Ausführung des Streams). Bei Auswahl von “Felder mit nur 1 Wert ignorieren” werden Felder mit nur einem Wert automatisch ignoriert.

- Mithilfe der Schaltfläche Große Sets ignorieren im Menü “Extras” knnen Sie festlegen, dass große Sets ignoriert werden sollen, sobald ein Typknoten als Instanz erstellt wurde. Bei Auswahl von “Große Sets ignorieren” werden automatisch Sets mit sehr vielen Mitgliedern ignoriert.
- Mithilfe der Schaltfläche Stetige Ganzzahlen in Ordinalzahlen umwandeln im Menü “Extras” können Sie Ganzzahlenbereiche in Sets umwandeln, sobald ein Typknoten als Instanz erstellt wurde. Für weitere Informationen siehe Thema Stetige Daten umwandeln auf S. 141.
- Mithilfe der entsprechenden Schaltfläche im Menü “Extras” können Sie einen Filterknoten zum Verwerfen der ausgewählten Felder generieren.
- Mithilfe der Umschalttaste mit der Sonnenbrille können Sie den Standard für alle Felder auf “Lesen” oder “Übergeben” setzen. Die Registerkarte “Typen” im Quellenknoten übergibt Felder standardmäßig, während der Typknoten standardmäßig Werte liest.
- Mit der Schaltfläche Werte löschen können Sie die in diesem Knoten vorgenommenen Änderungen an den Feldwerten löschen (nicht übernommene Werte) und die Werte aus den aufwärts liegenden Operationen erneut lesen. Diese Operation dient zum Zurücksetzen von Änderungen, die Sie für bestimmte, aufwärts liegende Felder vorgenommen haben.
- Mit der Schaltfläche Alle Werte löschen können Sie die Werte für **alle** in den Knoten eingelesenen Felder zurücksetzen. Diese Option setzt die Spalte *Werte* für alle Felder effektiv auf **Lesen**. Mit dieser Option können Sie die Werte für alle Felder zurücksetzen und die Werte und Typen aus den aufwärts liegenden Operationen erneut lesen.
- Über das Kontextmenü (Kopieren) können Sie Attribute aus einem Feld in ein anderes kopieren. Für weitere Informationen siehe Thema Kopieren von Typattributen auf S. 152.
- Mithilfe der Option Nicht verwendete Feldeinstellungen anzeigen können Sie Typeinstellungen für Felder anzeigen, die nicht mehr in den Daten vorliegen oder die zuvor mit diesem Typknoten verbunden waren. Dies ist sinnvoll, wenn Sie einen Typknoten für Daten-Sets, die sich geändert haben, erneut verwenden möchten.

Messniveaus

Das Messniveau (früher “Datentyp” oder “Verwendung” genannt) beschreibt die Nutzung der Datenfelder in IBM® SPSS® Modeler. Das Messniveau kann auf der Registerkarte “Typen” eines Quellenknotens oder Typknotens festgelegt werden. Beispiel: Sie möchten das Messniveau für ein Feld ganzer Zahlen mit den Werten 1 und 0 auf *Flag* setzen. Das bedeutet normalerweise, dass 1=*True* und 0=*False* ist.

Speicherung versus Messung. Das Messniveau eines Felds unterscheidet sich dessen Speichertyp, der angibt, ob die Daten als Zeichenkette, ganze Zahl, reelle Zahl, Datum, Zeit oder Zeitstempel gespeichert werden sollen. Während die Datentypen mithilfe eines Typknotens an jeder beliebigen Stelle im Stream geändert werden können, muss der Speichertyp beim Lesen der Daten in SPSS Modeler stets an der Quelle festgelegt werden (kann jedoch später mithilfe einer Konvertierungsfunktion geändert werden). Für weitere Informationen siehe Thema Festlegen von Feldspeichertyp und Formatierung in Kapitel 2 auf S. 32.

Bei einigen Modellierungsknoten werden die zulässigen Messniveautypen für die Eingabe- und Zielfelder durch Symbole auf der Registerkarte “Felder” angegeben.

Messniveau-Symbole

Symbol	Messniveau
	Default
	Stetig
	Kategorial
	Flag
	Nominal
	Ordinal
	Typlos

Die folgenden Messniveaus stehen zur Verfügung:

- **Standard.** Daten, deren Speichertyp und Werte unbekannt sind (z. B. weil sie noch nicht gelesen wurden), werden als <Standard> angezeigt.
- **Stetig.**Wird zur Beschreibung numerischer Werte verwendet, beispielsweise des Bereichs 0–100 oder 0,75–1,25. Ein stetiger Wert kann eine ganze Zahl, eine reelle Zahl oder ein Datum/eine Uhrzeit sein.
- **Kategorial.** Wird für Zeichenkettenwerte verwendet, wenn eine exakte Anzahl unterschiedlicher Werte nicht bekannt ist. Dies ist ein Datentyp **ohne Instanz**, was bedeutet, dass nicht alle möglichen Informationen über Speicherung und Verwendung der Daten bereits bekannt sind. Nach dem Lesen der Daten ist das Messniveau *Flag*, *Nominal* oder *Ohne Typ*, abhängig von der maximalen Anzahl an Mitgliedern für nominale Felder, die im Dialogfeld "Stream-Eigenschaften" angegeben wurde.
- **Flag.** Wird für Daten mit zwei verschiedenen Werten verwendet, die auf das Vorhandensein bzw. Nichtvorhandensein eines Merkmals hinweisen, beispielsweise true und false, Yes und No oder 0 und 1. Die verwendeten Werte können abweichen, aber ein Wert muss immer als "wahr" und der andere als "falsch" festgelegt sein. Die Daten können als Text, ganze Zahl, reelle Zahl, Datum/Uhrzeit oder Zeitstempel dargestellt sein.
- **Nominal.** Wird verwendet, um Daten mit mehreren unterschiedlichen Werten zu beschreiben, von denen jeder als Mitglied eines Sets behandelt wird, beispielsweise small/medium/large (klein/mittel/groß). Nominale Daten können jeden beliebigen Speichertyp aufweisen – "Numerisch", "Zeichenkette" oder "Datum/Uhrzeit". Hinweis: Durch das Setzen des Messniveaus auf *Nominal* werden nicht automatisch die Werte auf Zeichenkettenspeicherung geändert.
- **Ordinal.** Wird zur Beschreibung von Daten mit mehreren unterschiedlichen Werten verwendet, die eine natürliche Reihenfolge aufweisen. Gehaltskategorien oder Zufriedenheitsbewertungen beispielsweise können als ordinale Daten klassifiziert werden. Die Reihenfolge wird durch die natürliche Sortierfolge der Datenelemente definiert. So ist 1, 3, 5 die Standardsortierreihenfolge für eine Menge von ganzen Zahlen, während HIGH, LOW, NORMAL (aufsteigende alphabetische Reihenfolge) die Reihenfolge für eine Menge von Zeichenketten ist. Mit dem ordinalen Messniveau können Sie eine Menge kategorialer Daten als ordinale Daten festlegen - zum Zwecke der Visualisierung, Modellerstellung und

zum Export in andere Anwendungen, beispielsweise IBM® SPSS® Statistics, die ordinale Daten als gesonderten Typ erkennen. Sie können ein ordinales Feld überall dort verwenden, wo sich ein nominales Feld verwenden lässt. Außerdem können Felder jedes beliebigen Speichertyps (“Reelle Zahl”, “Ganze Zahl”, “Zeichenkette”, “Datum”, “Zeit” usw.) als “ordinal” definiert werden.

- **Ohne Typ.** Wird für Daten verwendet, die keinem der oben angegebenen Typen entsprechen, für Felder mit einem einzelnen Wert bzw . für nominale Daten, in denen das Set mehr Mitglieder als das definierte Maximum enthält. Dieser Typ ist auch sinnvoll in Fällen, in denen das Messniveau ansonsten ein Set mit zu vielen Mitgliedern wäre (beispielsweise eine Kontonummer). Bei Auswahl von Ohne Typ für ein Feld wird die Rolle automatisch auf Keine festgelegt, wobei Datensatz-ID die einzige Alternative ist. Die Standard-Maximalgröße für Sets liegt bei 250 eindeutigen Werten. Diese Zahl kann in der Registerkarte “Optionen” des Dialogfelds “Stream-Eigenschaften” (Zugriff über das Menü “Extras”) angepasst oder deaktiviert werden.

Sie können die Messniveaus wahlweise manuell festlegen oder auch die Daten durch die Software einlesen und dann das Messniveau auf der Grundlage der eingelesenen Werte automatisch bestimmen lassen.

Alternativ können Sie bei mehreren stetigen Datenfeldern, die als kategoriale Daten behandelt werden sollen, eine Option auswählen, um sie umzuwandeln. Für weitere Informationen siehe Thema Stetige Daten umwandeln auf S. 141.

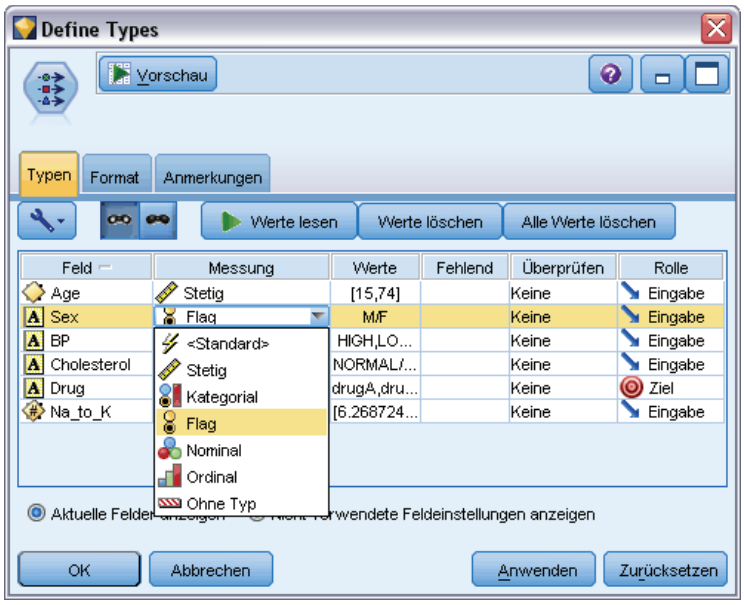
So verwenden Sie die automatische Typfestlegung:

- Setzen Sie in einem Typknoten oder auf der Registerkarte “Typen” eines Quellenknotens die Spalte *Werte* für die gewünschten Felder auf <Lesen>. Dadurch werden die Metadaten für alle abwärts liegenden Knoten verfügbar. Mit den Sonnenbrillen-Schaltflächen im Dialogfeld können Sie schnell und einfach alle Felder auf <Lesen> oder <Übergeben> setzen.
- Klicken Sie auf Werte lesen, um sofort die Werte aus der Datenquelle zu lesen.

So legen Sie das Messniveau für ein Feld manuell fest:

- Wählen Sie ein Feld in der Tabelle aus.
- Wählen Sie in der Dropdown-Liste in der Spalte *Messung* ein Messniveau für das Feld aus.
- Alternativ können Sie mit Strg-A oder Strg-Klicken mehrere Felder auswählen, bevor Sie in der Dropdown-Liste ein Messniveau auswählen.

Abbildung 4-20
Manuelles Festlegen der Messniveaus

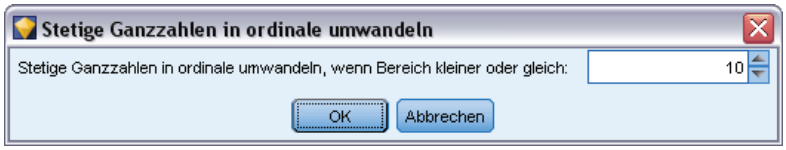


Stetige Daten umwandeln

Die Behandlung von kategorialen Daten als stetige Daten kann schwerwiegende Auswirkungen auf die Qualität eines Modells haben, besonders wenn es sich dabei um das Zielfeld handelt. So könnte beispielsweise ein Regressionsmodell anstelle eines binären Modells erzeugt werden. Um dies zu vermeiden, können Sie Ganzzahlenbereiche in kategoriale Typen wie *Ordinal* oder *Flag* umwandeln.

- Wählen Sie aus dem Menüfeld “Operationen und Generieren” (mit dem Werkzeugsymbol) die Option Stetige Ganzzahlen in Ordinalzahlen umwandeln. Das Dialogfeld “Umwandlungswerte” wird angezeigt.

Abbildung 4-21
Das Dialogfeld “Umwandlungswerte”



- Geben Sie die Größe des Bereichs an, der automatisch umgewandelt wird. Dies gilt für jeden Bereich bis zu der von Ihnen angegebenen Größe.

- Klicken Sie auf OK. Die betroffenen Bereiche werden in *Flag* oder *Ordinal* umgewandelt und auf der Registerkarte “Typen” des Typknotens angezeigt.

Ergebnisse der Umwandlung

- Wenn ein *stetiges* Feld mit Speichertyp "Ganze Zahl" in *Ordinal* umgewandelt wird, werden die unteren und oberen Werte erweitert, damit alle ganzzahligen Werte mit einbezogen werden. Wenn der Bereich die Werte 1 und 5 umfasst, besteht das Werteset aus 1, 2, 3, 4 und 5.
- Wenn das *stetige* Feld in *Flag* umgewandelt wird, werden der obere und untere Wert zum Wahr- und Falsch-Wert des Flag-Feldes.

Was ist Instanziierung?

Instanziierung ist der Prozess des Lesens oder Angebens von Informationen, beispielsweise des Speichertyps und der Werte für ein Datenfeld. Zur Optimierung der Systemressourcen handelt es sich bei der Instanziierung um einen benutzergesteuerten Prozess – Sie weisen die Software an, Datenwerte zu lesen, indem Sie die entsprechenden Optionen auf der Registerkarte "Typen" in einem Quellenknoten angeben bzw. indem Sie Daten durch einen Typknoten laufen lassen.

- Daten mit unbekannten Typen werden auch als **ohne Instanz** bezeichnet. Daten mit unbekanntem Speichertyp und unbekannten Werten werden in der Spalte *Messung* der Registerkarte "Typen" als <Standard> angezeigt.
- Wenn ein Teil der Informationen über den Speichertyp eines Felds vorliegt, beispielsweise "Zeichenkette" oder "Numerisch", werden die betreffenden Daten als **teilweise instanziiert** bezeichnet. Kategorial bzw. Stetig sind teilweise instanziierte Messniveaus. Kategorial beispielsweise gibt an, dass das Feld symbolisch ist, jedoch nicht bekannt ist, ob es nominal, ordinal oder Flag ist.
- Wenn alle Details über einen Typ bekannt sind, einschließlich der Werte, wird ein **vollständig instanziiertes** Messniveau – nominal, ordinal, Flag, stetig – in dieser Spalte angezeigt. *Hinweis*: Der Typ *stetig* wird sowohl für teilweise als auch für vollständig instanziierte Datenfelder verwendet. Bei stetigen Daten kann es sich entweder um ganze Zahlen oder um reelle Zahlen handeln.

Während der Ausführung eines Daten-Streams mit einem Typknoten werden Typen ohne Instanz sofort auf der Grundlage der ursprünglichen Datenwerte teilweise instanziiert. Sobald alle Daten den Knoten durchlaufen haben, werden alle Daten vollständig instanziiert, es sei denn, einige Werte wurden auf <Übergeben> gesetzt. Wenn die Ausführung unterbrochen wird, bleiben die Daten teilweise instanziiert. Sobald die Registerkarte "Typen" als Instanz erstellt wurde, sind die Werte eines Felds an dieser Stelle im Stream statisch. Das bedeutet, dass Änderungen weiter oben im Stream sich nicht auf die Werte eines bestimmten Felds auswirken, selbst wenn der Stream erneut ausgeführt wird. Um die Werte auf der Grundlage neuer Daten oder weiterer Bearbeitungen zu ändern bzw. zu aktualisieren, müssen Sie sie auf der Registerkarte "Typen" bearbeiten oder den Wert für ein Feld auf <Lesen > oder <Lesen +> setzen.

Zeitpunkt der Instanziierung

Im Allgemeinen ist bei nicht allzu großen Daten-Sets und wenn keine Felder später im Stream hinzugefügt werden sollen, eine Instanziierung am Quellenknoten die praktischste Methode. Die Instanziierung in einem separaten Typknoten ist jedoch in folgenden Fällen ratsam:

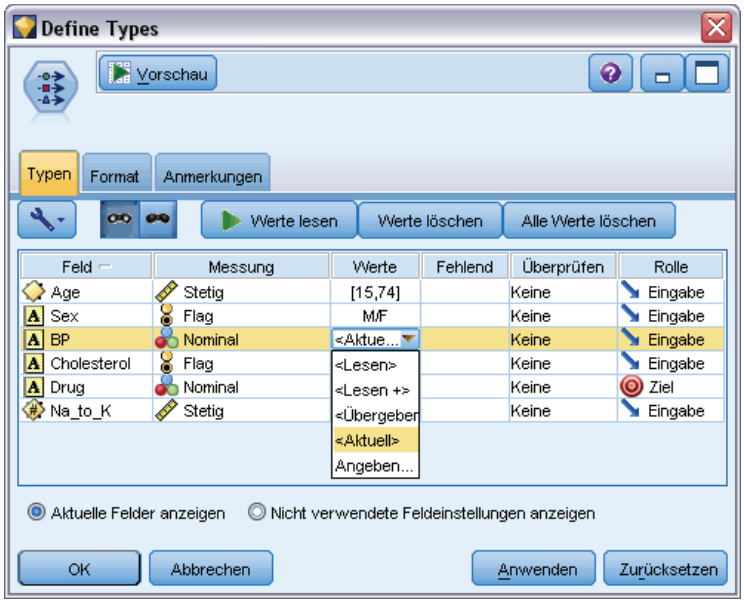
- Das Daten-Set ist groß und der Stream filtert eine Untergruppe vor dem Typknoten.

- Im Stream wurden Daten gefiltert.
- Im Stream wurden Daten zusammengeführt oder angehängt.
- Während der Verarbeitung werden neue Datenfelder abgeleitet.

Datenwerte

Mithilfe der Spalte *Werte* der Registerkarte "Typen" können Sie die Werte automatisch aus den Daten einlesen oder Sie können Messniveaus und Werte in einem separaten Dialogfeld angeben.

Abbildung 4-22
Auswählen der Methoden für das Lesen, Übergeben oder Angeben von Datenwerten.



Die in dieser Dropdown-Liste verfügbaren Optionen enthalten folgende Anweisungen für die automatische Typfestlegung:

Option	Funktion
<Read>	Die Daten werden gelesen, wenn der Knoten ausgeführt wird.
<Lesen +>	Die Daten werden gelesen und an die aktuellen Daten angehängt (sofern vorhanden).
<Pass>	Es werden keine Daten gelesen.
<Current>	Aktuelle Werte werden beibehalten.
Angeben...	Ein separates Dialogfeld wird gestartet, in dem Sie Werte und Optionen für Messniveaus angeben können.

Durch Ausführen eines Typknotens oder Klicken auf Werte lesen werden Werte aus Ihrer Datenquelle auf der Grundlage ihrer Auswahl automatisch einem Typ zugewiesen und gelesen. Diese Werte können auch mithilfe der Option "Angeben" oder durch Doppelklicken in eine Zelle in der Spalte *Feld* manuell angegeben werden.

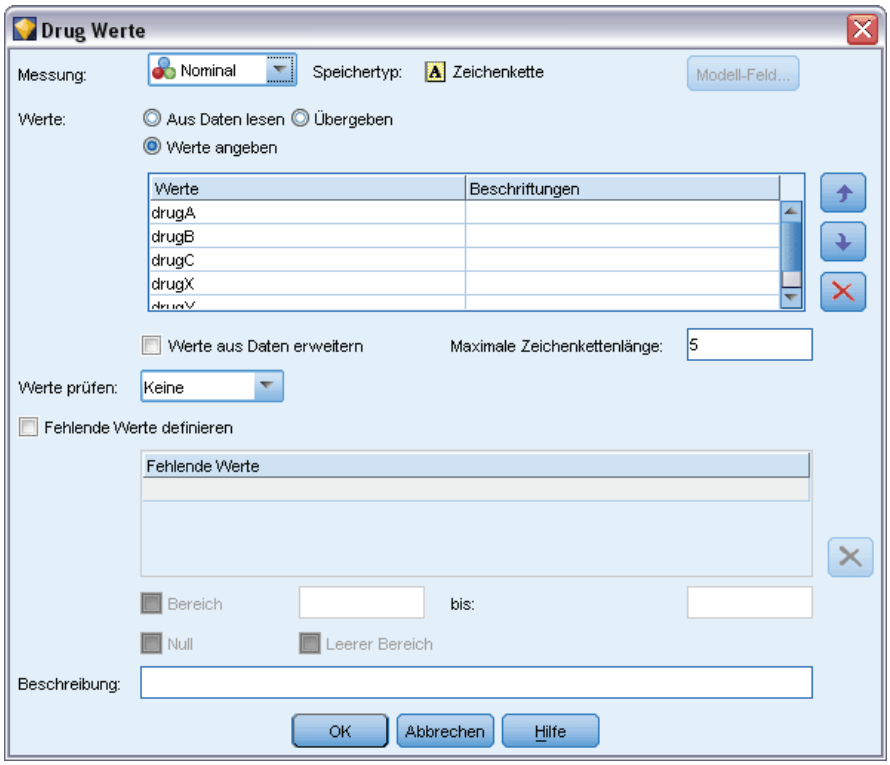
Nachdem Sie die Änderungen für die Felder im Typknoten vorgenommen haben, können Sie die Werteinformationen mithilfe der folgenden Schaltflächen in der Symbolleiste des Dialogfelds zurücksetzen:

- Mit der Schaltfläche Werte löschen können Sie die in diesem Knoten vorgenommenen Änderungen an den Feldwerten löschen (nicht übernommene Werte) und die Werte aus den aufwärts liegenden Operationen erneut lesen. Diese Operation dient zum Zurücksetzen von Änderungen, die Sie für bestimmte, aufwärts liegende Felder vorgenommen haben.
- Mit der Schaltfläche Alle Werte löschen können Sie die Werte für **alle** in den Knoten eingelesenen Felder zurücksetzen. Diese Option setzt die Spalte *Werte* für alle Felder effektiv auf **Lesen**. Mit dieser Option können Sie die Werte für alle Felder zurücksetzen und die Werte und Messniveaus aus den aufwärts liegenden Operationen erneut lesen.

Verwenden des Dialogfelds "Werte"

Durch Klicken auf die Spalte *Werte* oder *Fehlend* der Registerkarte "Typen" wird eine Dropdown-Liste mit vordefinierten Werten angezeigt. Durch Auswahl der Option *Angeben* wird ein gesondertes Dialogfeld geöffnet, in dem Sie die Optionen zum Lesen, Angeben, Beschriften und Behandeln der Werte für das ausgewählte Feld festlegen können.

Abbildung 4-23
Festlegen der Optionen für Datenwerte



Viele der Steuerelemente sind für alle Datentypen gleich. Diese gemeinsamen Steuerelemente werden hier erörtert.

Messung. Zeigt das aktuell ausgewählte Messniveau an. Sie können die Einstellung so ändern, wie die Daten verwendet werden sollen. Beispiel: Wenn ein Feld mit dem Titel *Tag_der_Woche* Zahlen enthält, die für die einzelnen Tage stehen, können Sie dieses in nominale Daten ändern, um einen Verteilungsknoten zu erstellen, der jede Kategorie einzeln untersucht.

Speichertyp. Zeigt den Speichertyp an, sofern dieser bekannt ist. Speichertypen werden vom gewählten Messniveau nicht beeinflusst. Zum Ändern des Speichertyps können Sie die Registerkarte "Daten" in den Quellenknoten "Datei (fest)" und "Datei (var.)" oder eine Konvertierungsfunktion in einem Füllerknoten verwenden.

Modell-Feld. Bei Feldern, die beim Scoren eines Modell-Nuggets generiert wurden, können auch Details zum Modell-Feld angezeigt werden. Dazu gehören der Name des Zielfelds sowie die Rolle des Felds bei der Modellierung (ob es sich um einen vorhergesagten Wert, eine Wahrscheinlichkeit, Neigung usw. handelt).

Werte. Dient zur Auswahl einer Methode zur Bestimmung der Werte für das ausgewählte Feld. Die hier vorgenommene Auswahl setzt alle Optionen außer Kraft, die Sie zuvor in der Spalte *Werte* des Dialogfelds "Typknoten" vorgenommen haben. Zu den Auswahlmöglichkeiten zum Lesen von Werten gehören:

- **Aus Daten lesen.** Wählen Sie diese Option aus, um Daten lesen zu lassen, wenn der Knoten ausgeführt wird. Diese Option entspricht <Lesen>.
- **Übergeben.** Wählen Sie diese Option aus, wenn keine Daten für das aktuelle Feld gelesen werden sollen. Diese Option entspricht <Übergeben>.
- **Werte angeben.** Die Optionen hier werden zur Angabe von Werten und Beschriftungen für das ausgewählte Feld verwendet. In Verbindung mit der Werteprüfung ermöglicht diese Option die Angabe von Werten auf der Grundlage Ihrer Kenntnisse über das aktuelle Feld. Diese Option aktiviert eindeutige Steuerelemente für jeden Feldtyp. Die Optionen für Werte und Beschriftungen werden einzeln in den nachfolgenden Themenabschnitten behandelt. *Hinweis*: Werte und Beschriftungen können nicht für ein Feld, dessen Messniveau *Ohne Typ* oder <*Standard*> lautet.
- **Werte aus Daten erweitern.** Wählen Sie diese Option, um die aktuellen Daten mit den hier eingegebenen Werten zu ergänzen. Wenn beispielsweise *Feld_1* den Bereich (0,10) aufweist und Sie den Wertebereich (8,16) eingeben, wird der Bereich durch Hinzufügen von 16 erweitert, wobei der ursprüngliche Mindestwert nicht verändert wird. Der neue Bereich ist also (0,16). Durch Auswahl dieser Option wird die Option für die automatische Typfestlegung auf <Lesen+> gesetzt.

Werte prüfen. Dient zur Auswahl einer Methode, mit der erzwungen wird, dass die Werte den angegebenen stetigen, Flag- oder nominalen Werten entsprechen. Diese Option entspricht der Spalte *Überprüfen* im Dialogfeld "Typknoten" und die hier vorgenommenen Einstellungen setzen die Einstellungen im Dialogfeld außer Kraft. In Verbindung mit der Option "Werte angeben" ermöglicht die Wertprüfung den Abgleich der Werte in den Daten mit den erwarteten Werten. Beispiel: Wenn Sie die Werte als "1, 0" angeben und anschließend die Option Verwerfen verwenden, können Sie alle Datensätze verwerfen, die andere Werte aufweisen als 1 oder 0.

Fehlende Werte definieren. Wählen Sie diese Option aus, um die unten angegebenen Steuerelemente zu aktivieren, mit denen Sie fehlende Werte oder Leerzeichen in Ihren Daten angeben können.

- **Tabelle fehlender Werte.** Ermöglicht die Definition bestimmter Werte (z. B. 99 oder 0) als Leerstellen. Der Wert sollte für den Speichertyp des Felds geeignet sein.
- **Bereich.** Wird verwendet, um den Bereich fehlender Werte anzugeben, beispielsweise der Altersbereich 1–17 oder älter als 65. Wenn ein Begrenzungswert leer gelassen wird, bleibt der Bereich ohne Begrenzung. Beispiel: Wenn als Untergrenze 100 angegeben wird, jedoch keine Obergrenze, werden alle Werte größer oder gleich 100 als fehlend definiert. Die Begrenzungswerte werden mit eingeschlossen, d. h., bei einer Untergrenze von 5 und einer Obergrenze von 10 sind 5 und 10 in der Bereichsdefinition enthalten. Ein Bereich fehlender Werte kann für jeden Speichertyp definiert werden, einschließlich "Datum/Uhrzeit" und "Zeichenkette" (in diesem Fall wird die alphabetische Sortierung verwendet, um zu bestimmen, ob ein Wert im Bereich liegt).
- **Null/Leerer Bereich.** Außerdem können Sie systemdefinierte **Nullen** (in den Daten als $null$ angezeigt) und **leere Bereiche** (Zeichenkettenwerte ohne sichtbare Zeichen) als Leerstellen angeben. Beachten Sie, dass leere Zeichenketten zum Zweck der Analyse im Typknoten als leere Bereiche behandelt werden, auch wenn sie intern anders gespeichert und in bestimmten Fällen anders behandelt werden.

Hinweis: Um Leerstellen als nicht definiert oder $null$ zu kodieren, sollten Sie den Füllerknoten verwenden.

Beschreibung. Verwenden Sie dieses Textfeld zur Angabe einer Feldbeschriftung. Diese Beschriftungen werden an verschiedenen Stellen angezeigt, beispielsweise in Diagrammen, Tabellen, Ausgaben und Modellbrowsern, je nach der im Dialogfeld "Stream-Eigenschaften" getroffenen Auswahl.

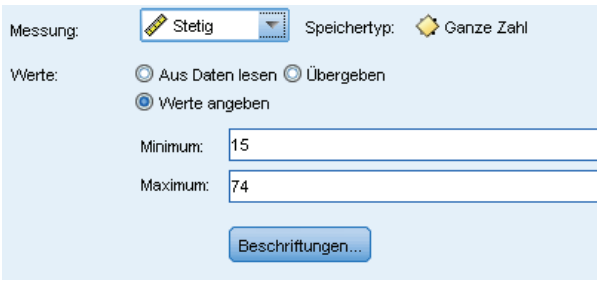
Angabe von Werten und Beschriftungen für stetige Daten

Das *stetige* Messniveau wird für numerische Felder verwendet. Es gibt drei Speichertypen für stetige Daten:

- Reelle Zahl
- Ganzzahl
- Datum/Uhrzeit

Dasselbe Dialogfeld wird zur Bearbeitung aller stetigen Felder verwendet; der Speichertyp wird nur als Referenz angezeigt.

Abbildung 4-24
Optionen zur Angabe von stetigen Werten und deren Beschriftungen



Angabe von Werten

Folgende Steuerelemente stehen nur bei stetigen Feldern zur Verfügung und werden zur Angabe von Wertebereichen verwendet:

Minimum. Geben Sie eine Untergrenze für den Wertebereich ein.

Maximum. Geben Sie eine Obergrenze für den Wertebereich ein.

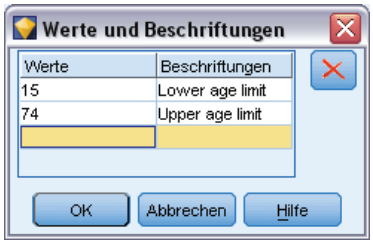
Angabe von Beschriftungen

Sie können Beschriftungen für jeden Wert eines Bereichsfelds angeben. Klicken Sie auf die Schaltfläche Beschriftungen, um ein gesondertes Dialogfeld zur Angabe der Wertelabels zu öffnen.

Unterdialogfeld "Werte und Beschriftungen"

Durch Klicken auf Beschriftungen im Dialogfeld "Werte" für ein Bereichsfeld wird ein neues Dialogfeld geöffnet, in dem Sie Beschriftungen für jeden Wert im Bereich angeben können.

Abbildung 4-25
Bereitstellen von Beschriftungen (optional) für Bereichswerte



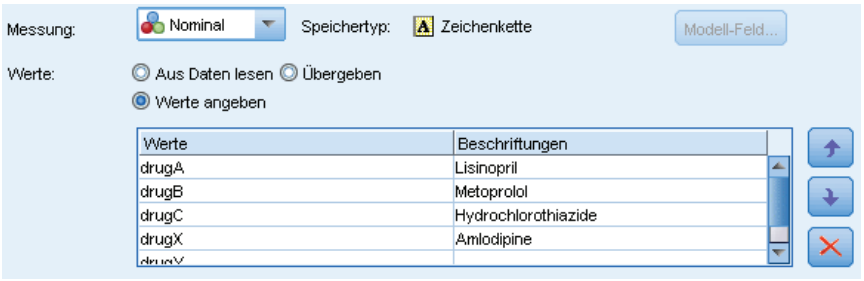
Mit den Spalten *Werte* und *Beschriftungen* in dieser Tabelle können Sie Wert-/Beschriftungspaare definieren. Die derzeit definierten Paare werden hier angezeigt. Sie können neue Beschriftungspaare hinzufügen, indem Sie in eine leere Zelle klicken und einen Wert und seine zugehörige Beschriftung eingeben. *Hinweis*: Durch das Hinzufügen von Wert-Wertebeschriftungs-Paaren in dieser Tabelle werden keine neuen Werte zum Feld hinzugefügt. Stattdessen werden einfach Metadaten für den Feldwert erstellt.

Die im Typknoten angegebenen Beschriftungen werden an verschiedenen Stellen angezeigt (als QuickInfo, Ausgabebeschriftungen usw.), je nach der im Dialogfeld "Stream-Eigenschaften" getroffenen Auswahl.

Angabe von Werten und Beschriftungen für nominale und ordinale Daten

Nominale (Set-) und ordinale Messniveaus (sortiertes Set) zeigen an, dass die Datenwerte diskret als Set-Mitglieder verwendet werden. Als Speichertypen für Sets sind "Zeichenkette", "Ganze Zahl", "Reelle Zahl" und "Datum/Uhrzeit" möglich.

Abbildung 4-26
Optionen zur Angabe von nominalen Werten und Beschriftungen



Folgende Steuerelemente stehen nur bei nominalen und ordinalen Feldern zur Verfügung und werden zur Angabe von Werten und Beschriftungen verwendet:

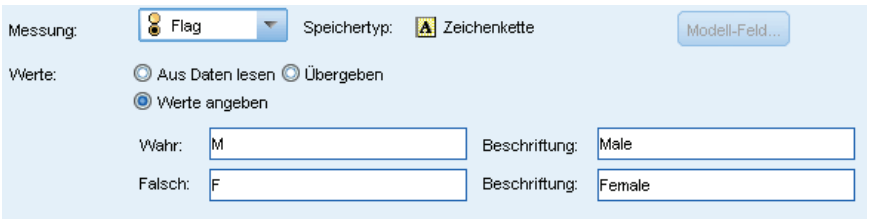
Werte. Mit der Spalte *Werte* in der Tabelle können Sie Werte auf der Grundlage Ihrer Kenntnisse über das aktuelle Feld angeben. Mithilfe dieser Tabelle können Sie erwartete Werte für das Feld eingeben und dann mit der Dropdown-Liste "Werte überprüfen" testen, ob das Daten-Set diesen Werten entspricht. Mit den Pfeilschaltflächen und der Löschschaltfläche können Sie bestehende Werte bearbeiten sowie Werte neu sortieren und löschen.

Beschriftungen. In der Spalte *Beschriftungen* können Sie Beschriftungen für jeden Wert im Set angeben. Diese Beschriftungen werden an verschiedenen Stellen angezeigt, beispielsweise in Diagrammen, Tabellen, Ausgaben und Modellbrowsern, je nach der im Dialogfeld "Stream-Eigenschaften" getroffenen Auswahl.

Angabe von Werten für ein Flag

Flag-Felder werden zur Anzeige von Daten verwendet, die zwei unterschiedliche Werte aufweisen. Als Speichertypen für Flags sind "Zeichenkette", "Ganze Zahl", "Reelle Zahl" und "Datum/Uhrzeit" möglich.

Abbildung 4-27
Optionen für die Angabe der Werte von Flag-Feldern.



Wahr. Dient zur Angabe eines Flag-Werts für das Feld, in dem die Bedingung erfüllt ist.

Falsch. Dient zur Angabe eines Flag-Werts für das Feld, in dem die Bedingung nicht erfüllt ist.

Beschriftungen. Dient zur Angabe von Beschriftungen für die einzelnen Werte im Flag-Feld. Diese Beschriftungen werden an verschiedenen Stellen angezeigt, beispielsweise in Diagrammen, Tabellen, Ausgaben und Modellbrowsern, je nach der im Dialogfeld "Stream-Eigenschaften" getroffenen Auswahl.

Fehlende Werte definieren

Die Spalte Fehlend der Registerkarte “Typen” zeigt an, ob der Umgang mit fehlenden Werten für ein Feld definiert wurde. Folgende möglichen Optionen stehen zur Auswahl:

On (*). Zeigt an, dass der Umgang mit fehlenden Werten für dieses Feld definiert ist. Dies kann durch einen abwärts gelegenen Füllerknoten oder durch eine ausdrückliche Angabe mithilfe der “Angabe”-Option (siehe unten) erfolgen.

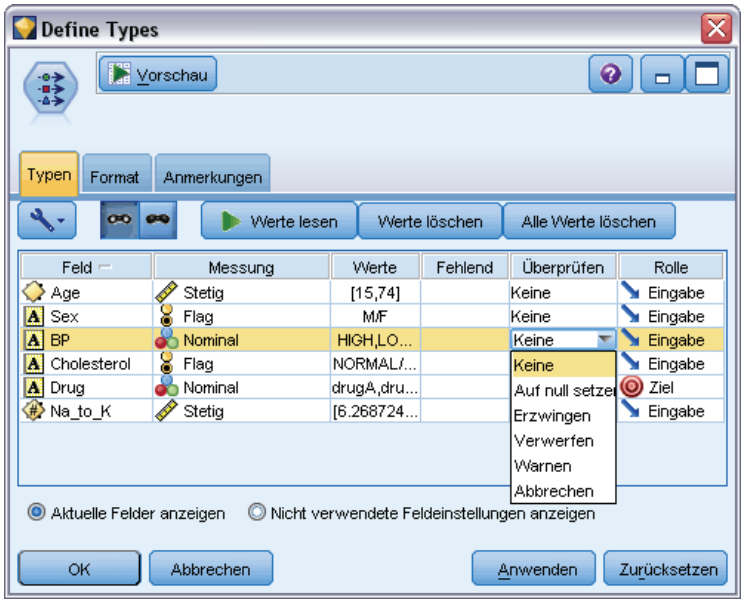
Off. Für das Feld ist kein Umgang mit fehlenden Werten definiert.

Angeben. Wählen Sie diese Option, um ein Dialogfeld anzuzeigen, in dem Sie explizite Werte festlegen können, die als fehlende Werte für dieses Feld betrachtet werden.

Überprüfen von Typenwerten

Durch Aktivierung der Option “Überprüfen” für die einzelnen Felder werden alle Werte im betreffenden Feld untersucht, um zu ermitteln, ob sie den aktuellen Typeneinstellungen bzw. den im Dialogfeld “Werte angeben” angegebenen Werten entsprechen. Dies ist sinnvoll bei der Bereinigung von Daten-Sets und zur Verringerung der Größe eines Daten-Sets innerhalb einer einzelnen Operation.

Abbildung 4-28
Auswählen von Überprüfungsoptionen für das ausgewählte Feld



Durch die Einstellung der Spalte *Überprüfung* im Dialogfeld “Typknoten” wird bestimmt, was geschieht, wenn ein Wert außerhalb der Typengrenzen gefunden wird. Die Überprüfungseinstellungen für ein Feld können Sie mithilfe der Dropdown-Liste für das betreffende Feld in der Spalte *Überprüfen* ändern. Um die Überprüfungseinstellungen für alle Felder festzulegen, klicken Sie in die Spalte *Feld* und drücken Sie Strg-A. Verwenden Sie anschließend die Dropdown-Liste für die Felder in der Spalte *Überprüfen*.

Folgende Überprüfungseinstellungen stehen zur Verfügung:

Keine. Die Werte werden ohne Überprüfung übergeben. Dies ist die Standardeinstellung.

Auf null setzen. Ändert Werte außerhalb der Grenzen auf den systemdefinierten Nullwert ($null$).

Erzwingen. Felder mit vollständig instanziierten Messniveaus werden auf Werte überprüft, die außerhalb der angegebenen Bereiche liegen. Nicht spezifizierte Werte werden mithilfe der folgenden Regeln in einen zulässigen Wert für das betreffende Messniveau konvertiert:

- Bei Flags werden alle Werte, die nicht "wahr" oder "falsch" sind, in den Wert "falsch" konvertiert.
- Bei Sets (nominal oder ordinal) werden alle unbekannten Werte in das erste Mitglied der Werte des Sets konvertiert.
- Zahlen, deren Wert die Obergrenze eines Bereichs überschreitet, werden durch den Wert der Obergrenze ersetzt.
- Zahlen, deren Wert die Untergrenze eines Bereichs unterschreitet, werden durch den Wert der Untergrenze ersetzt.
- Nullwerten in einem Bereich wird der Wert des Mittelpunkts für den betreffenden Bereich zugewiesen.

Verwerfen. Wenn unzulässige Werte gefunden werden, wird der gesamte Datensatz verworfen.

Warnen. Die Anzahl der unzulässigen Elemente wird gezählt und im Dialogfeld "Stream-Eigenschaften" gemeldet, nachdem alle Daten gelesen wurden.

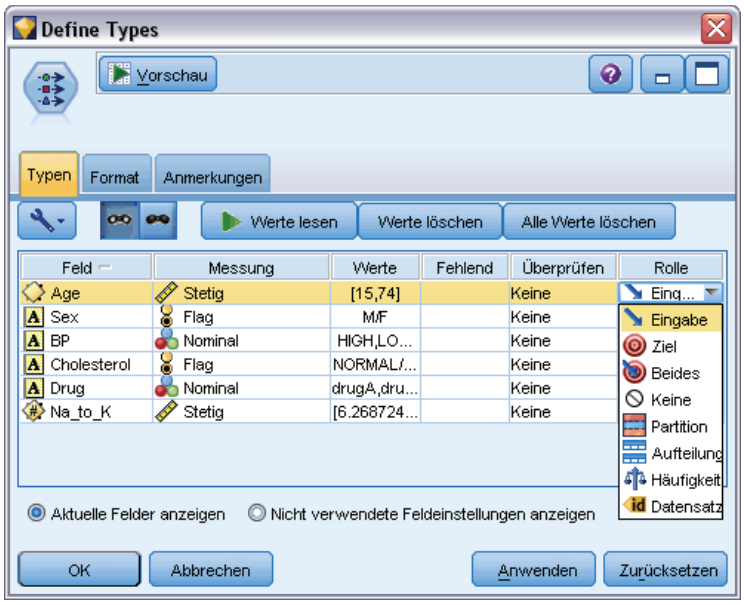
Abbrechen. Beim ersten unzulässigen Wert, der gefunden wird, wird die Ausführung des Streams abgebrochen. Der Fehler wird im Dialogfeld "Stream-Eigenschaften" gemeldet.

Festlegen der Feldrolle

Die Rolle eines Felds gibt an, wie es bei der Modellerstellung verwendet werden soll, beispielsweise ob es sich bei einem Feld um eine Eingabe oder um ein Ziel (das vorhergesagte Element) handelt.

Hinweis: Die Rollen Partition, Häufigkeit und Datensatz-ID können jeweils nur auf ein einziges Feld angewendet werden.

Abbildung 4-29
Festlegen der Feldrollenoptionen für den Typknoten



Folgende Rollen stehen zur Verfügung:

Eingabe. Das Feld wird als Eingabe für das Maschinenlernen verwendet (Prädiktorfeld).

Ziel. Das Feld wird als Ausgabe bzw. Ziel für das Maschinenlernen verwendet (eines der Felder, die das Modell vorherzusagen versucht).

Beides. Das Feld wird vom A Priori-Knoten sowohl als Prädiktor als auch als Ziel verwendet. Alle anderen Modellierungsknoten ignorieren das Feld.

Keine. Dieses Feld wird vom Maschinenlernen ignoriert. Felder, deren Messniveau auf Ohne Typ gesetzt wurde, werden in der Spalte *Rolle* automatisch auf Keine gesetzt.

Partition. Gibt ein Feld an, das zur Partitionierung der Daten in getrennte Stichproben für Trainings, Test- und (optional) Validierungszwecken verwendet wird. Dieses Feld muss ein instanziierter Set-Typ mit zwei oder drei möglichen Werten sein (wie im Dialogfeld "Feldwerte" definiert). Der erste Wert steht für die Trainings-Stichprobe, der zweite für die Test-Stichprobe und der dritte (sofern vorhanden) für die Validierungs-Stichprobe. Alle weiteren Werte werden ignoriert und Flag-Felder können nicht verwendet werden. Um die Partition in einer Analyse zu verwenden, muss auf der Registerkarte "Modelloptionen" des entsprechenden Modellerstellungs- oder Analyseknotens die Partitionierung aktiviert sein. Datensätze mit Nullwerten für das Partitionsfeld werden aus der Analyse ausgeschlossen, wenn die Partitionierung aktiviert ist. Wenn mehrere Partitionsfelder im Stream definiert wurden, muss in jedem Modellierungsknoten, der die Partitionierung verwendet, auf der Registerkarte "Felder" ein einzelnes Partitionsfeld ausgewählt werden. Wenn in Ihren Daten noch kein geeignetes Feld vorhanden ist, können Sie mithilfe eines Partitionierungs- oder Ableitungsknotens eines erstellen. Für weitere Informationen siehe Thema Partitionsknoten auf S. 208.

Aufteilen. (Nur nominale, ordinale und Flag-Felder.) Legt fest, dass ein Modell für jeden möglichen Wert des Felds erstellt werden soll.

Häufigkeit. (Nur numerische Felder.) Durch Festlegen dieser Rolle kann der Feldwert als Häufigkeitsgewichtungsfaktor für den Datensatz verwendet werden. Diese Funktion wird nur von C&R-Baum, CHAID-, QUEST- und linearen Modellen unterstützt; alle anderen Knoten ignorieren diese Rolle. Die Häufigkeitsgewichtung wird mithilfe der Option Häufigkeitsgewichtung anwenden auf der Registerkarte "Felder" der Modellierungsknoten aktiviert, die diese Funktion unterstützen.

Datensatz-ID. Das Feld wird als eindeutiger Bezeichner für einen Datensatz verwendet. Diese Funktion wird von den meisten Knoten ignoriert. Sie wird jedoch von linearen Modellen unterstützt und ist für die IBM Netezza-Knoten zum In-Database Mining erforderlich.

Kopieren von Typattributen

Die Attribute eines Typs, wie beispielsweise Werte, Überprüfungsoptionen und fehlende Werte, können problemlos zwischen Feldern kopiert werden:

- Klicken Sie mit der rechten Maustaste auf das Feld, dessen Attribute kopiert werden sollen.
- Wählen Sie im Kontextmenü die Option Kopieren aus.
- Klicken Sie mit der rechten Maustaste auf die Felder, deren Attribute geändert werden sollen.
- Wählen Sie Kontextmenü die Option Inhalte einfügen aus. *Hinweis*: Sie können mehrere Felder mittels Strg-Klicken oder über die Option Felder auswählen aus dem Kontextmenü auswählen.

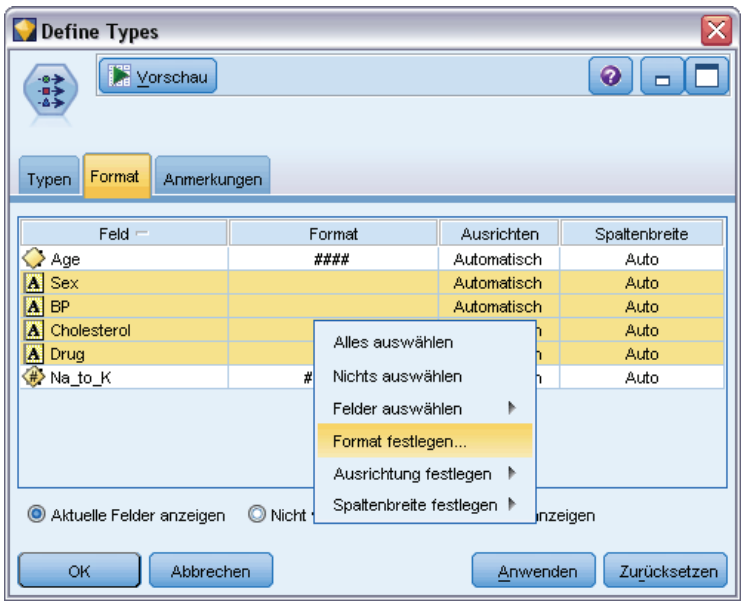
  Es wird ein neues Dialogfeld geöffnet, in dem Sie die speziellen Attribute auswählen können, die eingefügt werden sollen. Beim Einfügen in mehrere Felder gelten die hier ausgewählten Optionen für alle Zielfelder.

  Fügen Sie die folgenden Attribute ein. Wählen Sie das entsprechende Element aus der unten stehenden Liste aus, um Attribute aus einem Feld in ein anderes einzufügen.

  - **Typ.** Wählen Sie diese Option aus, um das Messniveau einzufügen.
  - **Werte.** Wählen Sie diese Option aus, um die Feldwerte einzufügen.
  - **Fehlende Werte.** Wählen Sie diese Option aus, um die Einstellungen für fehlende Werte einzufügen.
  - **Überprüfen.** Wählen Sie diese Option aus, um die Überprüfungsoptionen einzufügen.
  - **Rolle.** Wählen Sie diese Option aus, um die Rolle eines Felds einzufügen.

Feldformat – Registerkarte "Einstellungen"

Abbildung 4-30
Typknoten – Registerkarte "Format"



Die Registerkarte "Format" an den Tabellen- und Typknoten zeigt eine Liste der aktuellen oder nicht verwendeten Felder sowie Formatierungsoptionen für die einzelnen Felder. Im Folgenden finden Sie eine Beschreibung der einzelnen Spalten in der Feldformatierungstabelle:

Feld. Hier wird der Name des ausgewählten Felds angezeigt.

Format. Durch Doppelklicken auf eine Zelle in dieser Spalte können Sie die Formatierung für die einzelnen Felder anhand des eingeblendeten Dialogfelds angeben. Für weitere Informationen siehe Thema Festlegen der Feldformatierungsoptionen auf S. 154. Die hier angegebene Formatierung setzt die in den allgemeinen Stream-Eigenschaften angegebene Formatierung außer Kraft.

Hinweis: Die Knoten "Statistics-Export" und "Statistics-Ausgabe" exportieren *.sav*-Dateien, die Formatierungen für die einzelnen Felder in ihren Metadaten enthalten. Wenn das Format für die einzelnen Felder nicht vom Format der IBM® SPSS® Statistics-*.sav* -Datei unterstützt wird, verwendet der Knoten das SPSS Statistics-Standardformat.

Ausrichten. In dieser Spalte können Sie angeben, wie die Werte innerhalb der Tabellenspalte ausgerichtet werden sollen. Die Standardeinstellung ist Automatisch, was bedeutet, dass Symbolwerte links und numerische Werte rechts ausgerichtet werden. Sie können die Standardeinstellung durch Auswahl von Links, Rechts oder Mitte außer Kraft setzen.

Spaltenbreite. Standardmäßig wird die Spaltenbreite anhand der Werte des Felds automatisch berechnet. Um die automatische Berechnung der Spaltenbreite außer Kraft zu setzen, klicken Sie auf eine Tabellenzelle und wählen Sie mithilfe der Dropdown-Liste eine neue Breite aus. Zur Eingabe von benutzerdefinierten Breiten, die hier nicht aufgeführt sind, öffnen Sie das Unterdialogfeld "Feldformat", indem Sie auf eine Tabellenzelle in der Spalte *Feld* oder *Format* doppelklicken. Alternativ können Sie mit der rechten Maustaste auf eine Zelle klicken und Format festlegen auswählen.

Aktuelle Felder anzeigen. Standardmäßig wird im Dialogfeld die Liste der derzeit aktiven Felder angezeigt. Um die Liste der nicht verwendeten Felder anzuzeigen, wählen Sie Nicht verwendete Feldeinstellungen anzeigen.

Kontextmenü. Das Kontextmenü für diese Registerkarte bietet verschiedene Optionen für Auswahl und Einstellungsaktualisierung.

- **Alles auswählen.** Wählt alle Felder aus.
- **Nichts auswählen.** Hebt die Auswahl auf.
- **Felder auswählen.** Wählt Felder anhand von Typ- oder Speichertypeigenschaften aus. Zur Auswahl stehen: Kategorialen Wert auswählen, Stetigen Wert auswählen (numerisch), Element ohne Typ auswählen, Zeichenketten auswählen, Zahlen auswählen sowie Datum/Uhrzeit auswählen. Für weitere Informationen siehe Thema Messniveaus auf S. 138.
- **Format festlegen.** Öffnet ein Unterdialogfeld, in dem für die einzelnen Felder Optionen für Datum, Uhrzeit und Dezimaltrennzeichen angegeben werden können.
- **Ausrichtung festlegen.** Legt die Ausrichtung für die ausgewählten Felder fest. Zur Auswahl stehen: Automatisch, Mitte, Links und Rechts.
- **Spaltenbreite festlegen.** Legt die Feldbreite für ausgewählte Felder fest. Geben Sie Automatisch an, um die Breite aus den Daten einzulesen. Alternativ können Sie die Feldbreite auf folgende Werte festlegen: 5, 10, 20, 30, 50, 100 oder 200.

Festlegen der Feldformatierungsoptionen

Die Feldformatierung wird in einem Unterdialogfeld angegeben, das über die Registerkarte “Format” in den Typ- und Tabellenknoten aufgerufen werden kann. Wenn Sie vor dem Öffnen dieses Dialogfelds mehrere Felder ausgewählt haben, werden die Einstellungen aus dem ersten Feld in der Auswahl für alle Felder verwendet. Wenn Sie auf OK klicken, nachdem Sie hier Angaben gemacht haben, werden diese Einstellungen für alle Felder übernommen, die auf der Registerkarte “Format” ausgewählt wurden.

Abbildung 4-31
Festlegen der Formatierungsoptionen für mindestens ein Feld.

Die folgenden Optionen stehen für die einzelnen Felder zur Verfügung. Viele dieser Einstellungen können auch im Dialogfeld “Stream-Eigenschaften” angegeben werden. Alle auf der Feldebene vorgenommenen Einstellungen haben Vorrang gegenüber der für den Stream angegebenen Standardeinstellung.

Datumsformat. Wählen Sie ein Datumsformat aus, das für die Datumsspeicherfelder verwendet werden soll oder wenn Zeichenketten von den CLEM-Datumsfunktionen als Datumsangaben interpretiert werden.

Zeitformat. Wählen Sie ein Zeitformat aus, das für die Zeitspeicherfelder verwendet werden soll oder wenn Zeichenketten von den CLEM-Zeitfunktionen als Zeitangaben interpretiert werden.

Zahlenanzeigeformat. Sie können aus den Anzeigeformaten Standard (####.###), Wissenschaftlich (#.###E+##) und Währung ($###.##) wählen.

Dezimaltrennzeichen. Wählen Sie als Dezimaltrennzeichen entweder Komma (,) oder Punkt (.) aus.

Symbol für Zifferngruppierung. Wählen Sie bei Zahlenanzeigeformaten aus, welches Symbol zur Gruppierung der Werte verwendet werden soll (z. B. der Punkt in *3.000,00*). Folgende Optionen stehen zur Auswahl: “Keine”, “Punkt”, “Komma”, “Leerzeichen” und “Durch Ländereinstellung definiert” (in diesem Fall wird der Standardwert für die aktuelle Ländereinstellung verwendet).

Dezimalstellen (Standard, wissenschaftlich, Währung, Export). Gibt bei Zahlenanzeigeformaten an, wie viele Dezimalstellen bei der Anzeige, beim Drucken bzw. beim Export reeller Zahlen verwendet werden sollen. Diese Option wird getrennt für jedes Anzeigeformat angegeben. Das Exportformat gilt nur bei Feldern mit dem Speichertyp “Reelle Zahl”.

Ausrichten. Gibt an, wie die Werte innerhalb der Spalte ausgerichtet werden sollen. Die Standardeinstellung ist Automatisch, was bedeutet, dass Symbolwerte links und numerische Werte rechts ausgerichtet werden. Sie können die Standardeinstellung durch Auswahl von “Links”, “Rechts” oder “Mitte” außer Kraft setzen.

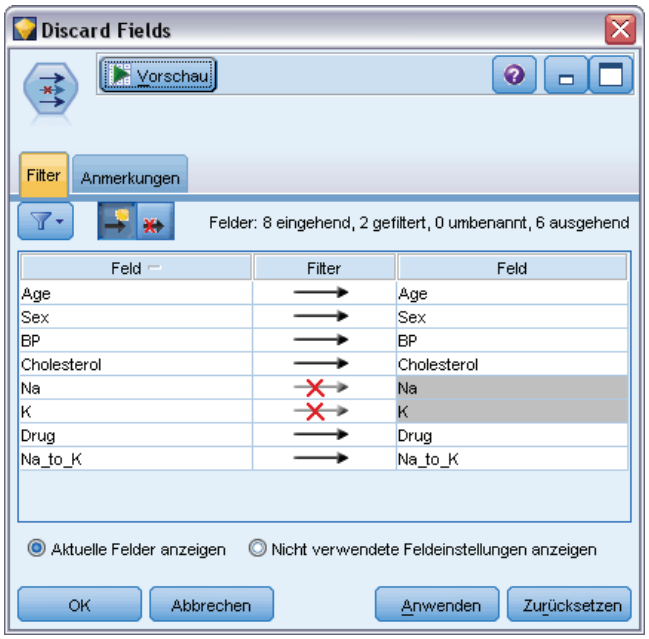
Spaltenbreite. Standardmäßig wird die Spaltenbreite anhand der Werte des Felds automatisch berechnet. Mit den Pfeilen rechts neben dem Listenfeld können Sie eine benutzerdefinierte Breite in Fünferschritten angeben.

Filtern bzw. Umbenennen von Feldern

Das Umbenennen und Ausschließen von Feldern ist an jedem beliebigen Punkt in einem Stream möglich. Beispiel: Bei einer medizinischen Studie ist möglicherweise der Kaliumspiegel (Daten der Feldebene) der Patienten (Daten der Datensatzebene) nicht relevant. Daher können Sie das Feld *K* (Kalium) herausfiltern. Dies ist mithilfe eines gesonderten Filterknotens oder mithilfe der Registerkarte "Filter" in einem Quellen- oder Ausgabeknoten möglich. Die Funktionen sind immer dieselben, unabhängig davon, von welchem Knoten aus der Zugriff erfolgt.

- An Quellenknoten wie beispielsweise "Datei (var.)", "Datei (fest)", "Statistics-Datei" und "XML-Datei" können Sie Felder beim Einlesen der Daten in IBM® SPSS® Modeler umbenennen oder filtern.
- Mit einem Filterknoten können Sie Felder an jeder Stelle des Streams umbenennen oder filtern.
- Von den Knoten "Statistics-Export", "Statistics-Transformation", "Statistics-Modell" und "Statistics-Ausgabe" können Sie Felder filtern oder umbenennen, die den IBM® SPSS® Statistics-Benennungsstandards entsprechen. Für weitere Informationen siehe Thema Umbenennen oder Filtern von Feldern für IBM SPSS Statistics in Kapitel 8 auf S. 516.
- Mit der Registerkarte "Filter" in einem der oben angegebenen Knoten können Sie Mehrfachantworten-Sets definieren bzw. bearbeiten. Für weitere Informationen siehe Thema Bearbeiten von Mehrfachantworten-Sets auf S. 160.
- Schließlich können Sie mit einem Filterknoten Felder aus einem Quellenknoten einem anderen Quellenknoten zuweisen.

Abbildung 4-32
Festlegen der Optionen für den Filterknoten



Festlegen der Filteroptionen

Die auf der Registerkarte “Filter” verwendete Tabelle zeigt den Namen der einzelnen Felder, die in den Knoten eintreten, sowie den Namen jedes Felds, das den Knoten verlässt. Mit den Optionen in diese Tabelle können Sie Felder umbenennen oder herausfiltern, die doppelt vorhanden oder für die Operationen weiter hinten im Stream nicht erforderlich sind.

- **Feld.** Zeigt die Eingabefelder aus den aktuell verbundenen Datenquellen an.
- **Filter.** Zeigt den Filterstatus aller Eingabefelder an. Gefilterte Felder weisen in dieser Spalte ein rotes “X” auf, das darauf hinweist, dass das Feld nicht an die späteren Operationen im Stream übergeben wird. Klicken Sie in die Spalte *Filter* für ein bestimmtes Feld, um die Filterfunktion zu aktivieren bzw. zu deaktivieren. Außerdem können Sie mit der Auswahl durch Umschalt-Klicken Optionen für mehrere Felder gleichzeitig auswählen.
- **Feld.** Zeigt die Felder an, die den Filterknoten verlassen. Doppelte Namen werden in roter Farbe angezeigt. Durch Klicken auf diese Spalte und Eingabe eines neuen Namens können Sie Feldnamen bearbeiten. Alternativ können Sie Felder entfernen, indem Sie doppelte Felder durch Klicken in die Spalte *Filter* deaktivieren.

Alle Spalten in der Tabelle können durch Klicken auf den Spaltentitel sortiert werden.

Aktuelle Felder anzeigen. Wählen Sie diese Option aus, um die Felder für Daten-Sets anzuzeigen, die aktiv mit dem Filterknoten verbunden sind. Diese Option wird standardmäßig ausgewählt und ist die häufigste Methode der Verwendung von Filterknoten.

Nicht verwendete Feldeinstellungen anzeigen. Wählen Sie diese Option aus, um die Felder für Daten-Sets anzuzeigen, die zu einem früheren Zeitpunkt (jetzt jedoch nicht mehr) mit dem Filterknoten verbunden waren. Diese Option wird vor allem beim Kopieren von Filterknoten

aus einem Stream in einen anderen oder beim Speichern und erneuten Laden von Filterknoten eingesetzt.

Schaltfläche "Filter" – Menü

Klicken Sie auf die Schaltfläche "Filter" links oben im Dialogfeld, um auf ein Menü zuzugreifen, das eine Reihe von Verknüpfungen und anderen Optionen enthält.

Abbildung 4-33
Optionen im Filtermenü

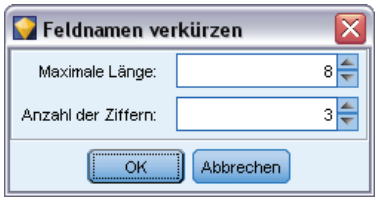
Sie haben folgende Möglichkeiten:

- Alle Felder entfernen.
- Alle Felder einschließen.
- Alle Felder umschalten.
- Duplikate entfernen. *Hinweis*: Bei Auswahl dieser Option werden alle Vorkommnisse des mehrfach vorhandenen Namens entfernt, einschließlich des ersten.
- Feldnamen und Mehrfachantworten-Sets umbenennen, sodass sie anderen Anwendungen entsprechen. Für weitere Informationen siehe Thema Umbenennen oder Filtern von Feldern für IBM SPSS Statistics in Kapitel 8 auf S. 516.
- Feldnamen verkürzen.
- Dient zur Anonymisierung der Namen von Feldern und Mehrfachantworten-Sets.
- Eingabefeldnamen verwenden.
- Mehrfachantworten-Sets bearbeiten. Für weitere Informationen siehe Thema Bearbeiten von Mehrfachantworten-Sets auf S. 160.
- Standard-Filterstatus festlegen.

Außerdem können Sie mit den Pfeilschaltflächen oben im Dialogfeld festlegen, ob Felder standardmäßig eingeschlossen oder verworfen werden sollen. Dies ist sinnvoll für große Daten-Sets, bei denen nur einige Felder weiter unten im Stream verwendet werden sollen. Wählen Sie beispielsweise nur die beizubehaltenden Felder aus und legen Sie fest, dass alle anderen Felder verworfen werden sollen, anstatt die zu verwerfenden Felder einzeln auszuwählen.

Verkürzen von Feldnamen

Abbildung 4-34
Dialogfeld "Feldnamen verkürzen"



Im Menü der Schaltfläche "Filter" (links oben auf der Registerkarte "Filter") können Sie auswählen, dass Feldnamen abgeschnitten werden sollen.

Maximale Länge. Dient zur Angabe einer Anzahl von Zeichen zur Begrenzung der Länge von Feldnamen.

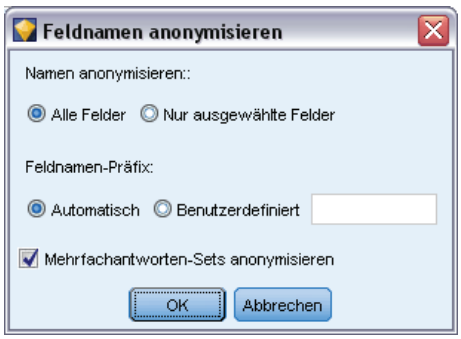
Anzahl der Ziffern. Wenn Feldnamen nach dem Verkürzen nicht mehr eindeutig sind, werden sie noch weiter verkürzt und durch Hinzufügen von Ziffern zum Namen unterschieden. Sie können angeben, wie viele Ziffern verwendet werden sollen. Mithilfe der Tabellenpfeile können Sie die Zahl einstellen.

Beispiel: Die unten stehende Tabelle illustriert das Verkürzen von Feldnamen in einem medizinischen Daten-Set mithilfe der Standardeinstellungen ("Maximale Länge" = 8 und "Anzahl der Ziffern" = 2).

Feldnamen	**Verkürzte Feldnamen**
Patienteneingabe 1	Patien01
Patienteneingabe 2	Patien02
Herzfrequenz	Herzfreq
BP	BP

Anonymisieren von Feldnamen

Abbildung 4-35
Dialogfeld "Feldnamen anonymisieren"



Feldnamen können aus jedem Knoten anonymisiert werden, der die Registerkarte "Filter" enthält. Klicken Sie dazu links oben auf das Menü der Schaltfläche "Filter" und wählen Sie die Option Feldnamen anonymisieren aus. Anonymisierte Feldnamen bestehen aus einem Zeichenkettenpräfix sowie einem eindeutigen Wert auf numerischer Basis.

Namen anonymisieren: Wählen Sie Nur ausgewählte Felder, um nur die Namen der Felder zu anonymisieren, die bereits auf der Registerkarte "Felder" ausgewählt wurden. Die Standardvorgabe lautet Alle Felder; dabei werden alle Feldnamen anonymisiert.

Feldnamens-Präfix. Das Standardpräfix für anonymisierte Feldnamen lautet anon_. Falls Sie ein anderes Präfix verwenden möchten, wählen Sie die Option Benutzerdefiniert und geben Sie das gewünschte Präfix ein.

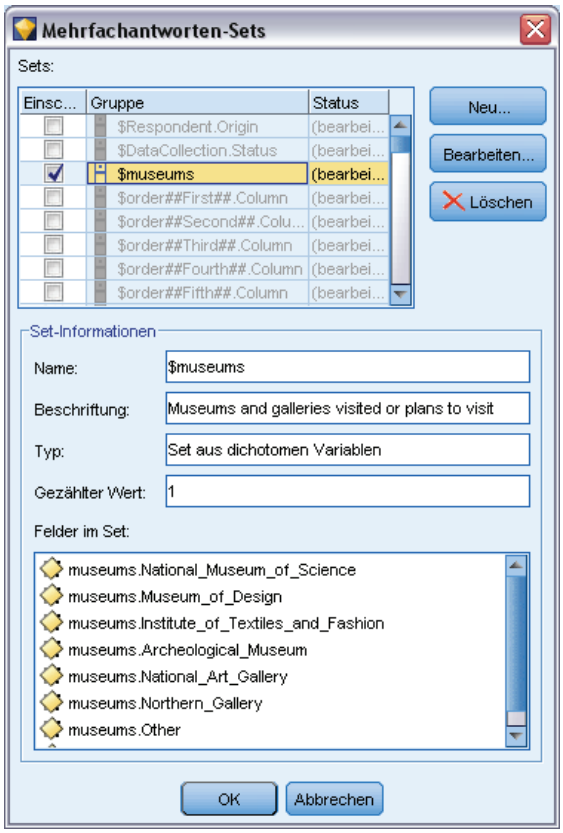
Mehrfachantworten-Sets anonymisieren. Anonymisiert die Namen von Mehrfachantworten-Sets auf dieselbe Weise wie Felder. Für weitere Informationen siehe Thema Bearbeiten von Mehrfachantworten-Sets auf S. 160.

Um die ursprünglichen Feldnamen wiederherzustellen, wählen Sie im Schatlflächenmenü "Filter" die Option Eingabefeldnamen verwenden.

Bearbeiten von Mehrfachantworten-Sets

Mehrfachantworten-Sets können aus jedem Knoten hinzugefügt bzw. bearbeitet werden, der die Registerkarte "Filter" enthält. Klicken Sie dazu links oben auf das Menü der Schaltfläche "Filter" und wählen Sie die Option Mehrfachantworten-Sets bearbeiten aus.

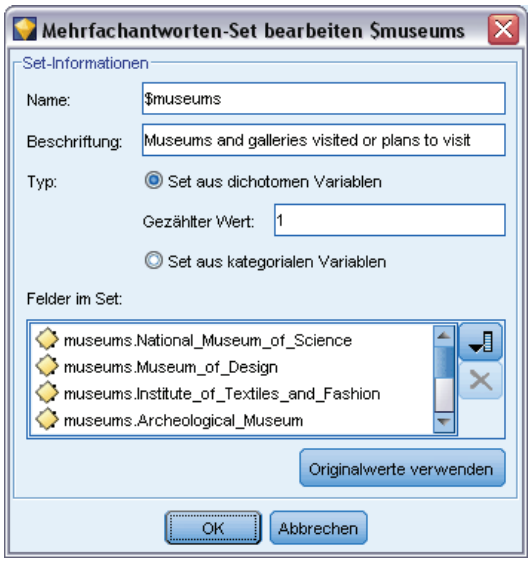
Abbildung 4-36
Dialogfeld "Mehrfachantworten-Sets"



Mehrfachantworten-Sets dienen zur Aufzeichnung von Daten, die für jeden Fall mehrere Werte aufweisen können. Dies ist beispielsweise der Fall, wenn die Teilnehmer an einer Umfrage gefragt werden, welche Museen sie besucht haben oder welche Zeitschriften sie lesen. Mehrfachantworten-Sets können mithilfe eines Data Collection-Quellenknotens oder eines Statistikdatei-Quellenknotens in IBM® SPSS® Modeler importiert werden und in SPSS Modeler mithilfe eines Filterknotens definiert werden.

- Klicken Sie auf Neu, um ein neues Mehrfachantworten-Set zu erstellen, oder klicken Sie auf Bearbeiten, um ein bestehendes Set zu bearbeiten.

Abbildung 4-37
Bearbeiten von Mehrfachantworten-Sets



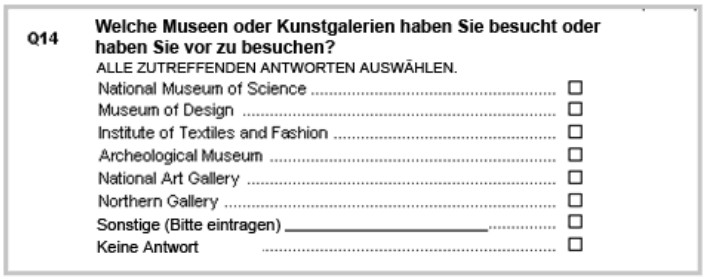
Name und Beschriftung. Gibt den Namen und die Beschreibung für das Set an.

Typ. Fragen mit Mehrfachantworten können auf zwei verschiedene Weisen verarbeitet werden:

- **Set aus dichotomen Variablen.** Für jede mögliche Antwort wird ein separates Flag-Feld erstellt. Bei 10 Zeitschriften werden also 10 Flag-Felder erstellt, die jeweils Werte wie 0 und 1 für *wahr* bzw. *falsch* aufweisen können. Unter "Gezählter Wert" können Sie angeben, welcher Wert als "wahr" gezählt werden soll. Diese Methode ist sinnvoll, wenn die Befragten die Möglichkeit haben sollen, alle zutreffenden Optionen auszuwählen.
- **Set aus kategorialen Variablen.** Für jede Antwort wird ein nominales Feld mit der maximalen Anzahl an Antwortmöglichkeiten für den Befragten erstellt. Jedes nominale Feld weist Werte für die möglichen Antworten auf, wie beispielsweise *1* für *Spiegel*, *2* für *Focus* und *3* für *Bunte*. Diese Methode ist dann am sinnvollsten, wenn Sie die Anzahl der Antwortmöglichkeiten einschränken möchten, beispielsweise, wenn die Befragten die drei Zeitschriften angeben sollen, die sie am häufigsten lesen.

Felder im Set. Mithilfe der Symbole auf der rechten Seite können Sie Felder hinzufügen bzw. entfernen.

Abbildung 4-38
Frage mit Mehrfachantworten



Kommentare

- Alle Felder in einem Mehrfachantworten-Set müssen denselben Speichertyp aufweisen.
- Es muss zwischen den Sets und den darin enthaltenen Feldern unterschieden werden. So werden beispielsweise durch das Löschen eines Sets nicht die darin enthaltenen Felder gelöscht, sondern lediglich die Verknüpfungen zwischen diesen Feldern. Das Set ist oberhalb vom Löschpunkt weiterhin sichtbar, nicht jedoch weiter unten im Stream.
- Wenn Felder mithilfe eines Filterknotens (unmittelbar auf der Registerkarte oder durch Auswahl der Optionen Umbenennen für IBM® SPSS® Statistics, Verkürzen oder Anonymisieren im Filtermenü) umbenannt werden, werden alle Verweise auf diese Felder in Mehrfachantworten-Sets ebenfalls aktualisiert. Felder in einem Mehrfachantworten-Set, die vom Filterknoten verworfen werden, werden jedoch nicht aus dem Mehrfachantworten-Set entfernt. Derartige Felder sind zwar nicht mehr im Stream sichtbar, werden jedoch weiterhin vom Mehrfachantworten-Set referenziert. Dies ist beispielsweise beim Export zu berücksichtigen.

Ensemble-Knoten

Der Ensemble-Knoten kombiniert zwei oder mehr Modell-Nuggets, um genauere Vorhersagen zu erzielen, als aus einem dieser Modelle allein gewonnen werden können. Durch die Kombination der Vorhersagen aus mehreren Modellen können Beschränkungen in einzelnen Modellen vermieden werden, was zu einer höheren Gesamtgenauigkeit führt. Auf diese Weise kombinierte Modelle bringen normalerweise eine mindestens ebenso gute Leistung wie die besten Einzelmodelle und sind häufig sogar noch besser.

Diese Kombination von Knoten geschieht automatisch in den automatisierten Modellierungsknoten “Automatischer Klassifizierer”, “Auto-Numerisch” und “Autom. Cluster”.

Nach der Verwendung eines Ensemble-Knotens können Sie mithilfe eines Analyse- oder Evaluationsknotens die Genauigkeit der kombinierten Ergebnisse mit den Ergebnissen der einzelnen als Eingabe verwendeten Modelle vergleichen. Hierbei darf die Option Von Ensemble-Modellen generierte Felder herausfiltern auf der Registerkarte “Einstellungen” des Ensemble-Knotens nicht ausgewählt sein.

Ausgabefelder (OutputFields)

Jeder Ensemble-Knoten generiert ein Feld mit den kombinierten Scores. Der Name beruht auf dem angegebenen Zielfeld und trägt das Präfix *$XF_*, *$XS_* oder *$XR_*, je nach Messniveau des Felds (Flag, nominal (Set) bzw. stetig (Bereich)). Beispiel: Wenn das Ziel ein Flag-Feld mit dem Namen *Antwort* ist, erhält das Ausgabefeld den Namen *$XF_Antwort*.

Konfidenz- bzw. Neigungsfelder. Bei Flag- und nominalen Feldern werden zusätzliche Konfidenz- bzw. Neigungsfelder, die auf der Ensemble-Methode beruhen, wie in der folgenden Tabelle beschrieben:

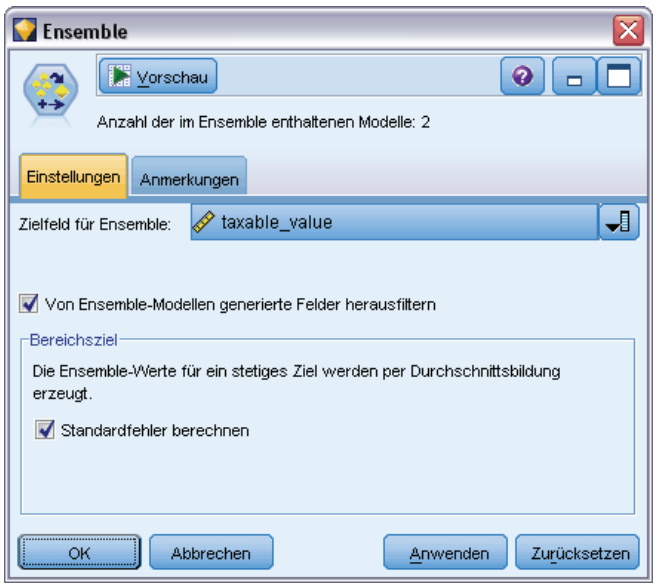
Ensemble-Methode	Feldname
Voting Nach Konfidenz gewichtetes Voting Nach Rohneigung gewichtetes Voting Nach korrigierter Neigung gewichtetes Voting Höchste Konfidenz hat Vorrang	*$XFC_<Feld>*
Durchschnittliche Rohneigung	*$XFRP_<Feld>*
Durchschnittliche korrigierte Neigung	*$XFAP_<Feld>*

Ensemble-Knoten – Einstellungen

Zielfeld für Ensemble. Dient zur Auswahl eines einzelnen Felds, das von zwei oder mehr weiter oben im Stream gelegenen Modellen als Ziel verwendet wird. Die weiter oben im Stream gelegenen Modelle können Ziele vom Typ "Flag", "nominal" oder "stetig" verwenden, aber mindestens zwei der Modelle müssen dasselbe Ziel verwenden, damit eine Kombination der Scores möglich ist.

Von Ensemble-Modellen generierte Felder herausfiltern. Entfernt alle zusätzlichen Felder aus der Ausgabe, die von den Einzelmodellen generiert wurden, die in den Ensemble-Knoten eingespeist werden. Aktivieren Sie dieses Kontrollkästchen, wenn Sie ausschließlich am kombinierten Score aus allen Eingabemodellen interessiert sind. Diese Option muss deaktiviert sein, wenn Sie beispielsweise einen Analyseknoten oder einen Evaluationsknoten verwenden möchten, um die Genauigkeit des kombinierten Score mit der Genauigkeit bei den einzelnen Eingabemodellen zu vergleichen.

Abbildung 4-39
Ensemble-Knoten mit stetigem Feld als Ziel



Die verfügbaren Einstellungen hängen vom Messniveau des Felds ab, das als Ziel ausgewählt ist.

Stetige Ziele

Bei stetigen Zielen wird der Durchschnitt aus den Scores gebildet. Dies ist die einzige verfügbare Methode für die Kombination von Scores.

Bei der Generierung von Durchschnitts-Scores oder Schätzungen verwendet der Ensemble-Knoten eine Standardfehlerberechnung, um den Unterschied zwischen den gemessenen oder geschätzten Werten und den wahren Werten zu berechnen und um anzuzeigen, wie hoch die Übereinstimmung dieser Schätzungen war. Standardfehlerberechnungen werden für neue Modelle standardmäßig generiert; Sie können das Kontrollkästchen jedoch für existierende Modelle deaktivieren, wenn sie beispielsweise neu generiert werden sollen.

Kategoriale Ziele.

Bei kategorialen Zielen werden mehrere Methoden unterstützt, darunter **Voting** (Abstimmung). Dabei wird zusammengerechnet, wie häufig jeder mögliche vorhergesagte Wert ausgewählt wurde; der Wert mit der höchsten Gesamtsumme wird dann verwendet. Beispiel: Wenn drei von fünf Modellen *Ja* vorhersagen und die anderen beiden *Nein*, dann gewinnt *Ja* mit 3 zu 2 "Stimmen". Alternativ können die Stimmen beim Voting auf der Grundlage des Konfidenz- oder Neigungswerts der einzelnen Vorhersagen **gewichtet** werden. Die Gewichte werden dann summiert und es wird wiederum der Wert mit dem höchsten Gesamtergebnis ausgewählt. Die Konfidenz für die endgültige Vorhersage ist die Summe der Gewichte für den Siegerwert dividiert durch die Anzahl der im Ensemble enthaltenen Modelle.

Abbildung 4-40
Ensemble-Knoten mit nominalem Feld als Ziel

Ausschließlich kategoriale Felder. Für Flag- und nominale Felder werden folgende Methoden unterstützt:

- Voting
- Nach Konfidenz gewichtetes Voting
- Höchste Konfidenz hat Vorrang

Abbildung 4-41
Ensemble-Knoten mit Flag-Feld als Ziel

Nur Flag-Felder. Wenn ausschließlich Flag-Felder vorliegen, steht außerdem eine Reihe von Methoden zur Verfügung, die auf Neigung beruhen:

- Nach Rohneigung gewichtetes Voting
- Nach korrigierter Neigung gewichtetes Voting
- Durchschnittliche Rohneigung
- Durchschnittliche korrigierte Neigung

Gleichstand beim Voting. Bei Voting-Methoden können Sie auswählen, wie Gleichstände aufgelöst werden sollen.

- **Zufallsauswahl.** Einer der gebundenen Werte (Werte mit Gleichstand) wird nach dem Zufallsprinzip ausgewählt.
- **Höchste Konfidenz.** Der gebundene Wert, der mit der höchsten Konfidenz vorhergesagt wurde, gewinnt. Beachten Sie, dass es sich hierbei nicht unbedingt um die höchste Konfidenz aller vorhergesagten Werte handelt.
- **Rohneigung oder korrigierte Neigung (nur bei Flag-Feldern).** Der gebundene Wert, der mit der höchsten absoluten Neigung vorhergesagt wurde. Dabei berechnet sich die absolute Neigung wie folgt:

```
abs(0.5 - propensity) *
2
```

Oder, bei korrigierter Neigung:

```
abs(0.5 - adjusted propensity) * 2
```

Ableitungsknoten

Eine der leistungsstärksten Funktionen in IBM® SPSS® Modeler ist die Möglichkeit, Datenwerte zu ändern und neue Felder aus bestehenden Daten abzuleiten. Bei längeren Data Mining-Projekten werden zumeist mehrere Ableitungen durchgeführt, beispielsweise die Extraktion einer Kunden-ID aus einer Zeichenkette mit Webprotokolldaten oder das Erstellen eines Kundenkapitalwerts auf der Basis von Transaktionsdaten und demografischen Daten. Alle diese Transformationen können mit einer Reihe von Feldoperationsknoten durchgeführt werden.

Mehrere Knoten bieten die Möglichkeit zur Ableitung neuer Felder:

Der Ableitungsknoten ändert Datenwerte oder erstellt neue Felder aus einem oder mehreren bestehenden Feldern. Er erstellt Felder vom Typ “Formel”, “Flag”, “Nominal”, “Status”, “Anzahl” und “Bedingt”. Für weitere Informationen siehe Thema Ableitungsknoten auf S. 167.

Der Umkodierungsknoten transformiert ein Set kategorialer Werte in ein anderes. Die Umkodierung dient zur Reduzierung von Kategorien bzw. Neugruppierung von Daten für die Analyse. Für weitere Informationen siehe Thema Umkodierungsknoten auf S. 187.

Der Klassierknoten erstellt automatisch neue nominale (Set-) Felder auf der Grundlage der Werte eines oder mehrerer bestehender stetiger Felder (numerischer Bereich). Sie können beispielsweise ein stetiges Einkommensfeld in ein neues kategoriales Feld transformieren, das Einkommensgruppen als Abweichungen vom Mittelwert enthält. Nach der Erstellung von Klassen für das neue Feld können Sie einen Ableitungsknoten anhand der Trennwerte generieren. Für weitere Informationen siehe Thema Klassierknoten auf S. 192.

Der Dichotomknoten leitet mehrere Flag-Felder auf der Grundlage der kategorialen Werte ab, die für ein oder mehrere nominale Felder definiert sind. Für weitere Informationen siehe Thema Dichotomknoten auf S. 211.

Der Knoten “Neu strukturieren” wandelt ein nominales Feld oder ein Flag-Feld in eine Gruppe von Feldern um, die mit den Werten aus einem weiteren Feld ausgefüllt werden können. Beispiel: Aus einem Feld mit dem Namen *Zahlungsart*, mit den Werten *Kreditkarte*, *Bar* und *EC-Karte* werden drei neue Felder erstellt (*Kreditkarte*, *Bar*, *EC-Karte*), die jeweils den Wert der jeweiligen Zahlung enthalten. Für weitere Informationen siehe Thema Neustrukturierungsknoten auf S. 212.

Der Verlaufsknoten erstellt neue Felder mit Daten aus Feldern in vorangegangenen Datensätzen. Verlaufsknoten werden am häufigsten für sequenzielle Daten, beispielsweise Zeitreihendaten, verwendet. Vor der Verwendung eines Verlaufsknotens sollten die Daten mithilfe eines Sortierknotens sortiert werden. Für weitere Informationen siehe Thema Verlaufsknoten auf S. 240.

Verwenden des Ableitungsknotens

Mithilfe des Ableitungsknotens können Sie sechs Typen neuer Felder aus einem oder mehreren Feldern erstellen:

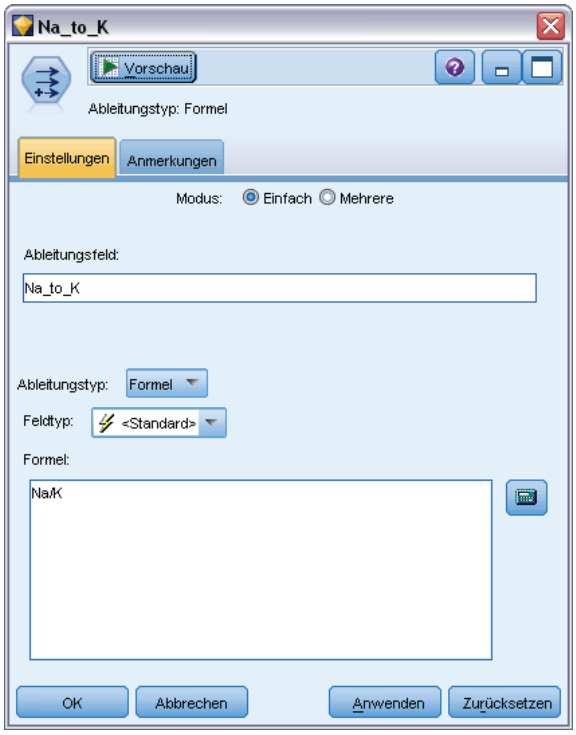
- **Formel.** Das neue Feld ist das Ergebnis eines beliebigen CLEM-Ausdrucks.
- **Flag.** Bei dem neuen Feld handelt es sich um ein Flag, das für eine angegebene Bedingung steht.
- **Nominal.** Bei dem neuen Feld handelt es sich um ein nominales Feld, was bedeutet, dass es eine Gruppe angegebener Werte als Mitglieder besitzt.
- **Status.** Das neue Feld weist einen von zwei Statuswerten auf. Der Wechsel zwischen diesen Statuswerten wird durch eine angegebene Bedingung ausgelöst.
- **Häufigkeiten.** Dieses neue Feld gibt an, wie oft eine Bedingung wahr war.
- **Bedingt.** Das neue Feld gibt den Wert eines von zwei Ausdrücken an, je nach dem Wert einer Bedingung.

Jeder dieser Knoten enthält ein Reihe von speziellen Optionen im Dialogfeld “Ableitungsknoten”. Diese Optionen werden in den nachfolgenden Themenabschnitten erörtert.

Festlegen der Grundoptionen für den Ableitungsknoten

Oben im Dialogfeld für Ableitungsknoten steht eine Reihe von Optionen zur Verfügung, mit denen Sie den Typ des von Ihnen benötigten Ableitungsknotens auswählen können.

Abbildung 4-42
Dialogfeld “Ableitungsknoten”



Modus. Wählen Sie Einfach oder Mehrere, je nachdem, ob Sie ein Feld oder mehrere Felder ableiten möchten. Bei Auswahl von Mehrere ändert sich das Dialogfeld. Es enthält nun Optionen für mehrere Ableitungsfelder.

Ableitungsfeld. Geben Sie bei einfachen Ableitungsknoten den Namen des Felds an, das Sie ableiten und zu den einzelnen Datensätzen hinzufügen möchten. Der Standardname lautet "Ableiten*N*". Dabei steht *N* für die Anzahl der Ableitungsknoten, die Sie bisher während der aktuellen Sitzung erstellt haben.

Ableitungstyp. Wählen Sie in der Dropdown-Liste einen Typ für den Ableitungsknoten aus, beispielsweise "Formel" oder "Nominal". Für jeden Typ wird auf der Grundlage der im typenspezifischen Dialogfeld angegebenen Bedingungen ein neues Feld erstellt.

Bei Auswahl einer Option in der Dropdown-Liste wird eine neue Gruppe von Steuerelementen zum Hauptdialogfeld hinzugefügt, die von den Eigenschaften jedes Ableitungsknotentyps abhängen.

Feldtyp. Wählen Sie ein Messniveau für den neu abgeleiteten Knoten aus, beispielsweise "Stetig", "Kategorial" oder "Flag". Diese Option haben alle Arten von Ableitungsknoten gemeinsam.

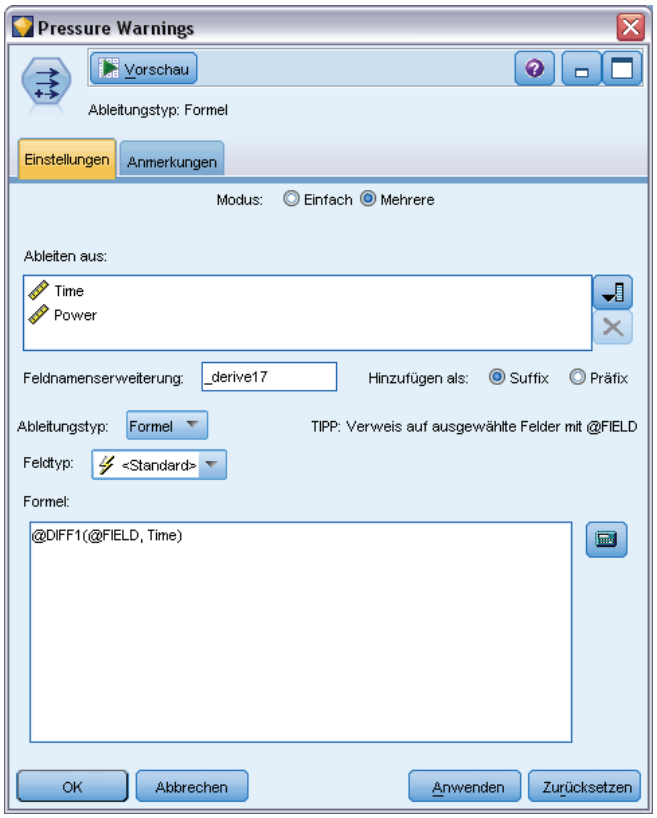
Hinweis: Für die Ableitung neuer Felder müssen häufig besondere Funktionen oder mathematische Ausdrücke verwendet werden. Um Ihnen die Erstellung dieser Ausdrücke zu erleichtern, steht im Dialogfeld für alle Typen von Ableitungsknoten ein Expression Builder zur Verfügung, mit dem Sie die Regeln überprüfen können und der außerdem eine vollständige Liste der CLEM-Ausdrücke bietet.

Ableiten mehrerer Felder

Wenn Sie den Modus innerhalb eines Ableitungsknotens auf Mehrere setzen, können Sie anhand derselben Bedingung innerhalb desselben Knotens mehrere Felder ableiten. Diese Funktion spart Zeit, wenn identische Transformationen für mehrere Felder im Daten-Set durchgeführt werden sollen. Beispiel: Wenn Sie ein Regressionsmodell erstellen möchten, das auf der Grundlage des Anfangsgehalts und der bisherigen Berufserfahrung das aktuelle Gehalt vorhersagt, kann es sinnvoll sein, für alle drei schiefen Variablen eine Log-Transformation durchzuführen. Anstatt für jede Transformation einen neuen Ableitungsknoten hinzuzufügen, können Sie dieselbe Funktion gleichzeitig zu allen Feldern hinzufügen. Wählen Sie einfach alle Felder aus, aus denen ein neues Feld abgeleitet werden soll, und geben Sie dann den Ableitungsausdruck mithilfe der Funktion @FIELD innerhalb der Feldklammern ein.

Hinweis: Die Funktion @FIELD ist ein wichtiges Tool zur gleichzeitigen Ableitung mehrerer Felder. Damit können Sie auf den Inhalt des aktuellen Felds bzw. der aktuellen Felder Bezug nehmen, ohne den genauen Feldnamen angeben zu müssen. Ein CLEM-Ausdruck, der zur Anwendung einer Log-Transformation auf mehrere Felder verwendet wird, ist beispielsweise log(@FIELD).

Abbildung 4-43
Ableiten mehrerer Felder



Folgende Optionen werden zum Dialogfeld hinzugefügt, wenn Sie den Modus Mehrere auswählen:

Ableiten aus. Verwenden Sie die Feldauswahl-Schaltfläche zur Auswahl von Feldern, aus denen neue Felder abgeleitet werden sollen. Für jedes ausgewählte Feld wird ein Ausgabefeld generiert. *Hinweis*: Die ausgewählten Felder müssen nicht denselben Speichertyp aufweisen. Allerdings schlägt der Ableitungsvorgang fehl, wenn die Bedingung nicht für *alle* Felder gültig ist.

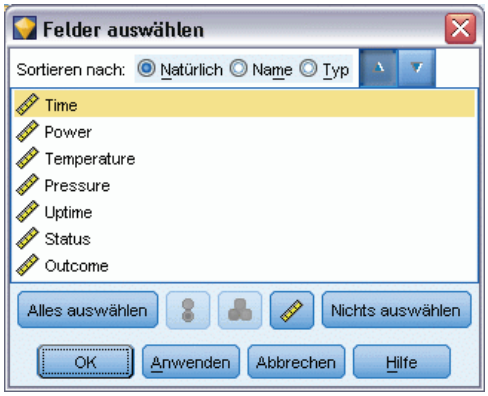
Feldnamenerweiterung. Geben Sie die Erweiterung ein, die zu den neuen Feldnamen hinzugefügt werden soll. Beispiel: Bei einem neuen Feld, das den Logarithmus von *Aktuelles Gehalt* enthält, könnten Sie den Feldnamen mit *log_* erweitern, wodurch sich *log_Aktuelles Gehalt* ergibt. Mit den Optionsfeldern können Sie auswählen, ob die Erweiterung als Präfix (am Anfang) oder als Suffix (am Ende) des Feldnamens eingefügt werden soll. Der Standardname lautet "Ableiten*N*". Dabei steht *N* für die Anzahl der Ableitungsknoten, die Sie bisher während der aktuellen Sitzung erstellt haben.

Wie beim Ableitungsknoten im Einzelmodus müssen Sie jetzt einen Ausdruck erstellen, der zur Ableitung eines neuen Felds verwendet wird. Je nach dem Typ der ausgewählten Ableitungsoperation steht eine Reihe von Optionen zum Erstellen einer Bedingung zur Verfügung. Diese Optionen werden in den nachfolgenden Themenabschnitten erörtert. Um einen Ausdruck zu erstellen, können Sie einfach Eingaben in den Formularfeldern vornehmen oder durch Klicken auf die Taschenrechner-Schaltfläche den Expression Builder verwenden. Denken Sie daran, die Funktion @FIELD zu verwenden, wenn es um Bearbeitungen in mehreren Feldern geht.

Auswählen mehrerer Felder

Bei allen Knoten, die Operationen in mehreren Eingabefeldern durchführen, wie “Ableiten” (Mehrfachmodus), “Aggregieren”, “Sortieren”, “Multidiagramm” und “Zeitplot”, können Sie mithilfe des Dialogfelds “Felder auswählen” schnell und einfach mehrere Felder auswählen.

Abbildung 4-44
Auswählen mehrerer Felder



Sortieren nach. Sie können die verfügbaren Felder für die Anzeige sortieren. Dazu stehen folgende Optionen zur Verfügung:

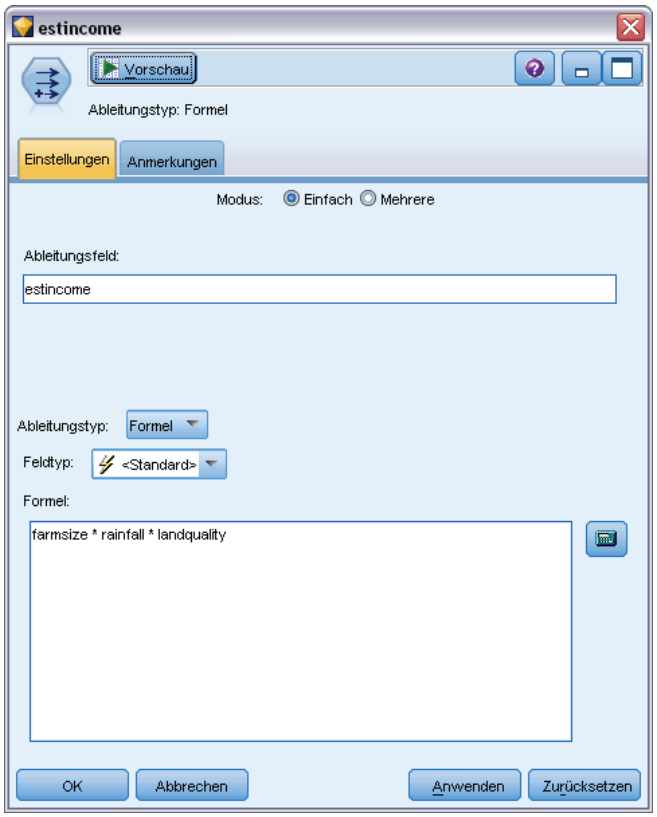
- **Natürlich.** Zeigt die Felder in der Reihenfolge an, in der Sie über den Daten-Stream an den aktuellen Knoten übergeben wurden.
- **Name.** Die Felder werden für die Anzeige alphabetisch sortiert.
- **Typ.** Die Felder werden in der Anzeige nach Messniveau sortiert. Diese Option ist bei der Auswahl von Feldern mit einem bestimmten Messniveau nützlich.

Sie können die Felder in der Liste einzeln auswählen oder mithilfe von Umschalt-Klicken bzw. Strg-Klicken mehrere Felder gleichzeitig auswählen. Außerdem können Sie mit den Schaltflächen unter der Liste Gruppen von Feldern anhand ihres Messniveaus oder alle Felder in der Tabelle auswählen bzw. die Auswahl aller Felder aufheben.

Festlegen der Formel-Ableitungsoptionen

Formel-Ableitungsknoten erstellen ein neues Feld für jeden Datensatz im Daten-Set auf der Grundlage der Ergebnisse eines CLEM-Ausdrucks. Beachten Sie, dass dieser Ausdruck nicht bedingt sein kann. Um Werte auf der Grundlage eines bedingten Ausdrucks abzuleiten, verwenden Sie den Typ “Flag” oder “Bedingt” des Ableitungsknotens.

Abbildung 4-45
Festlegen der Optionen für einen Formel-Ableitungsknoten

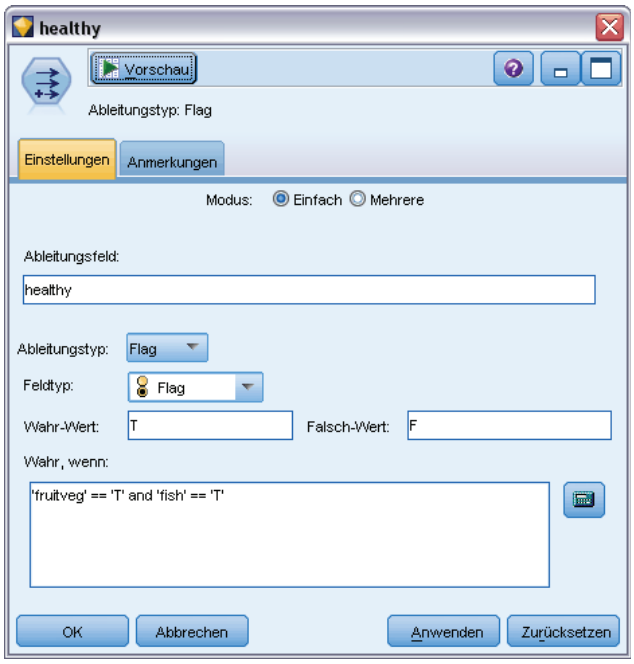


Formel. Geben Sie mithilfe der CLEM-Sprache eine Formel an, um einen Wert für das neue Feld abzuleiten.

Festlegen der Flag-Ableitungsoptionen

Flagableitungsknoten werden verwendet, um eine bestimmte Bedingung anzugeben, beispielsweise hohen Blutdruck oder Inaktivität auf dem Kundenkonto. Ein Flag-Feld wird für jeden Datensatz erstellt und wenn die Bedingung “Wahr” erfüllt ist, wird der Flag-Wert für “Wahr” in das Feld eingetragen.

Abbildung 4-46
Ableiten eines Flag-Felds



Wahr-Wert. Dient zur Angabe einen Werts, der für Datensätze, die die unten angegebene Bedingung erfüllen, in das Flag-Feld aufgenommen werden soll. Der Standardwert lautet “T”.

Falsch-Wert. Dient zur Angabe eines Werts, der für Datensätze, die die unten angegebene Bedingung *nicht* erfüllen, in das Flag-Feld aufgenommen werden soll. Der Standardwert lautet “F”.

Wahr, wenn. Dient zur Angabe einer CLEM-Bedingung zur Evaluierung bestimmter Werte jedes Datensatzes und zur Zuweisung eines Wahr- oder Falsch-Werts (oben definiert) für den Datensatz. Beachten Sie: Datensätzen wird bei nichtfalschen numerischen Werten der Wahr-Wert zugewiesen.

Hinweis: Wenn eine leere Zeichenkette ausgegeben werden soll, sollten Sie öffnende und schließende Anführungszeichen ohne etwas dazwischen eingeben, d. h. `""`. Leere Zeichenketten werden beispielsweise häufig als Falsch-Wert verwendet, damit die Wahr-Werte deutlicher in der Tabelle sichtbar sind. Anführungszeichen sollten außerdem verwendet werden, wenn ein Zeichenkettenwert gewünscht wird, der ansonsten als Zahl behandelt werden würde.

Beispiel

In Versionen von IBM® SPSS® Modeler vor 12.0 wurden Mehrfachantworten in ein einzelnes Feld importiert. Die Werte wurden dabei durch Kommas getrennt.

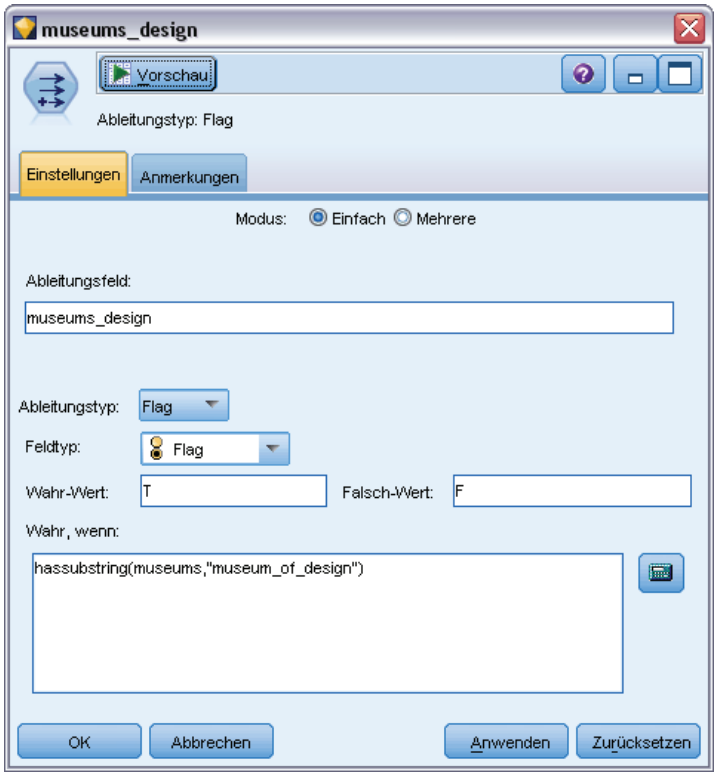
museums
museum_of_design,institute_of_textiles_and_fashion
museum_of_design
archeological_museum

museums
$null$
national_art_gallery,national_museum_of_science,other

Um diese Daten für die Analyse vorzubereiten, können Sie mithilfe der Funktion hassubstring ein gesondertes Flag-Feld für jede Antwort mit einem Ausdruck der folgenden Art erstellen:

```
hassubstring(museums,"museum_of_design")
```

Abbildung 4-47
Ableiten eines Flag-Felds mithilfe der Funktion "hassubstring"



Festlegen der Set-Ableitungsoptionen

Set-Ableitungsknoten dienen zur Ausführung eines Sets von CLEM-Bedingungen, um zu ermitteln, welche Bedingung die einzelnen Datensätze erfüllen. Wenn eine Bedingung für jeden Datensatz erfüllt ist, wird ein Wert (der angibt, welches Bedingungs-Set erfüllt war) in das neue, abgeleitete Feld eingetragen.

Abbildung 4-48
Verwenden eines Set-Ableitungsknotens

Standardwert. Geben Sie einen Wert an, der im neuen Feld verwendet werden soll, wenn keine der Bedingungen erfüllt ist.

Feld setzen auf. Dient zur Angabe eines Werts, der in das neue Feld eingetragen werden soll, wenn eine bestimmte Bedingung erfüllt ist. Jedem Wert in der Liste ist eine Bedingung zugeordnet, die in der benachbarten Spalte anzugeben ist.

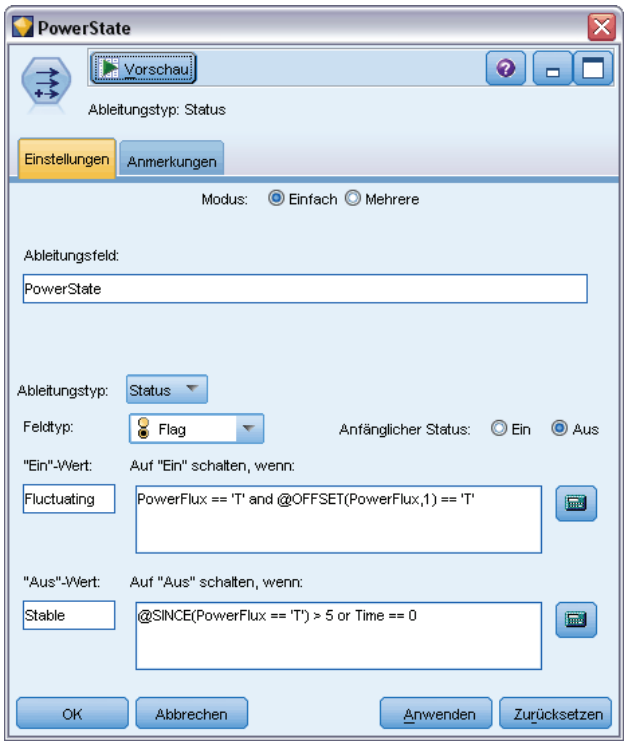
Wenn diese Bedingung wahr ist. Dient zur Angabe einer Bedingung für jedes Mitglied im Set-Feld. Mit dem Expression Builder können Sie eine Auswahl aus den verfügbaren Funktionen und Feldern treffen. Mit den Pfeilschaltflächen und der Löschschaltfläche können Sie Bedingungen neu ordnen bzw. entfernen.

Bei Bedingungen werden die Werte eines bestimmten Felds im Daten-Set gestestet. Beim Testen der einzelnen Bedingungen werden die oben angegebenen Werte dem neuen Feld zugewiesen, um anzuzeigen, welche Bedingung erfüllt wurde. Wenn keine der Bedingungen erfüllt wurde, wird der Standardwert verwendet.

Festlegen der Status-Ableitungsoptionen

Status-Ableitungsknoten weisen eine gewisse Ähnlichkeit mit Flag-Ableitungsknoten auf. Flag-Knoten setzen Werte abhängig von der Erfüllung einer *einzelnen* Bedingung für den aktuellen Datensatz fest, Status-Ableitungsknoten dagegen können die Werte eines Felds abhängig davon ändern, wie es *zwei unabhängige* Bedingungen erfüllt. Das bedeutet, dass sich der Wert ändert (Schalten auf "Ein" bzw. "Aus"), je nachdem, ob die Bedingung erfüllt ist.

Abbildung 4-49
Verwenden eines Status-Ableitungsknotens



Anfänglicher Status. Dient zur Auswahl, ob jedem Datensatz des neuen Felds ursprünglich der Wert Ein oder Aus zugewiesen werden soll. Beachten Sie: Dieser Wert kann sich ändern, wenn die einzelnen Bedingungen erfüllt werden.

“Ein”-Wert. Dient zur Angabe des Werts für das neue Feld, wenn die Bedingung für “Ein” erfüllt ist.

Auf “Ein” schalten, wenn. Dient zur Angabe einer CLEM-Bedingung, die den Wert auf “Ein” ändert, wenn die Bedingung wahr ist. Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder zu öffnen.

“Aus”-Wert. Dient zur Angabe des Werts für das neue Feld, wenn die Bedingung für “Aus” erfüllt ist.

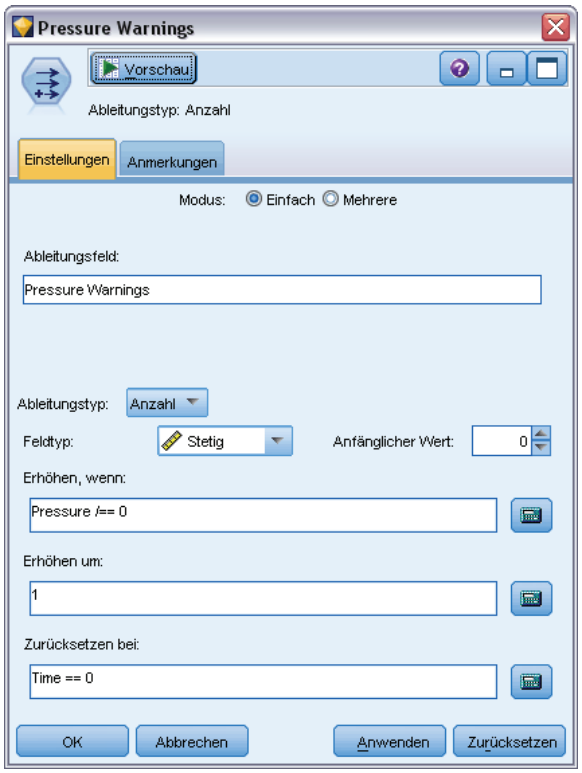
Auf “Aus” schalten, wenn. Dient zur Angabe einer CLEM-Bedingung, die den Wert auf “Aus” ändert, wenn die Bedingung falsch ist. Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder zu öffnen.

Hinweis: Um eine leere Zeichenkette anzugeben, sollten Sie öffnende und schließende Anführungszeichen ohne etwas dazwischen eingeben, d. h. `""`. Anführungszeichen sollten außerdem verwendet werden, wenn ein Zeichenkettenwert gewünscht wird, der ansonsten als Zahl behandelt werden würde.

Festlegen der Anzahl-Ableitungsoptionen

Anzahl-Ableitungsknoten werden verwendet, um eine Reihe von Bedingungen auf die Werte eines numerischen Felds im Daten-Set anzuwenden. Wenn die einzelnen Bedingungen erfüllt sind, erhöht sich der Wert des abgeleiteten Anzahlfelds um ein festgelegtes Inkrement. Diese Art von Ableitungsknoten ist sinnvoll für Zeitreihendaten.

Abbildung 4-50
Optionen für "Anzahl" im Dialogfeld "Ableitungsknoten"



Anfänglicher Wert. Legt einen Wert fest, der bei der Ausführung für das neue Feld verwendet wird. Beim anfänglichen Wert muss es sich um eine numerische Konstante handeln. Mithilfe der Pfeilschaltflächen können Sie den Wert erhöhen oder verringern.

Erhöhen, wenn. Dient zur Angabe der CLEM-Bedingung, bei deren Erfüllung der abgeleitete Wert anhand der in "Erhöhen um" angegebenen Zahl geändert wird. Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder zu öffnen.

Erhöhen um. Dient zur Festlegung des zur Erhöhung der Anzahl verwendeten Werts. Sie können entweder eine numerische Konstante oder das Ergebnis eines CLEM-Ausdrucks verwenden.

Zurücksetzen bei. Dient zur Angabe einer Bedingung, bei deren Erfüllung der abgeleitete Wert auf den anfänglichen Wert zurückgesetzt wird. Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder zu öffnen.

Festlegen der "Bedingt"-Ableitungsoptionen

"Bedingt"-Ableitungsknoten verwenden eine Reihe von Wenn-Dann-Sonst-Anweisungen zur Ableitung des Werts des neuen Felds.

Abbildung 4-51
Verwenden eines "Bedingt"-Ableitungsknotens

Wenn. Dient zur Angabe einer CLEM-Bedingung, die bei Ausführung für jeden Datensatz evaluiert wird. Wenn die Bedingung wahr (bzw. bei Zahlen: nichtfalsch) ist, wird dem neuen Feld der unten durch den Dann-Ausdruck angegebene Wert zugewiesen. Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder zu öffnen.

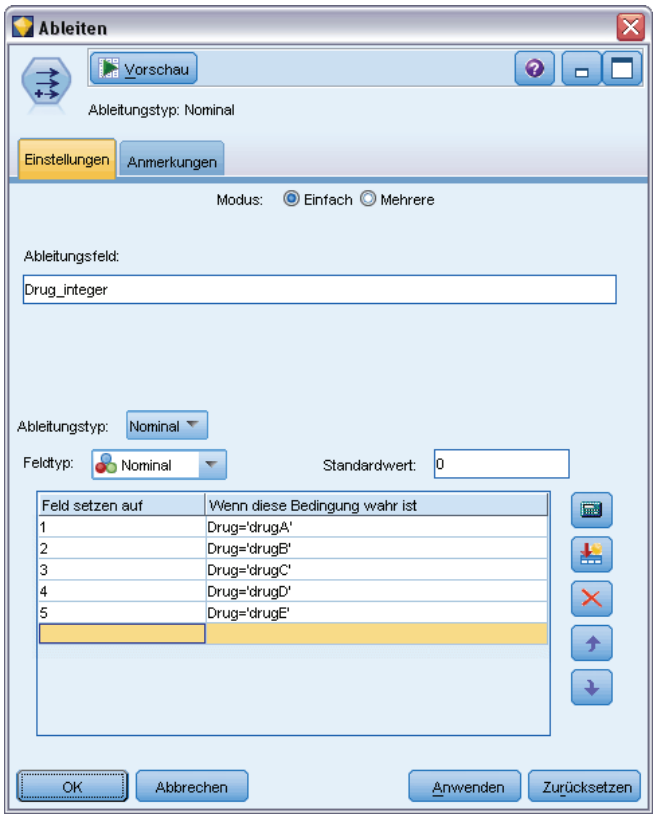
Dann. Dient zur Angabe eines Werts bzw. CLEM-Ausdrucks, der für das neue Feld gilt, wenn die oben stehende Wenn-Anweisung wahr (bzw. nichtfalsch) ist. Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder zu öffnen.

Sonst. Dient zur Angabe eines Werts bzw. CLEM-Ausdrucks, der für das neue Feld gilt, wenn die oben stehende Wenn-Anweisung falsch ist. Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder zu öffnen.

Umkodieren von Werten mit dem Ableitungsknoten

Mit Ableitungsknoten können auch Werte umkodiert werden, beispielsweise durch Konvertieren eines Zeichenkettenfelds mit kategorialen Werten in ein numerisches nominales (Set-) Feld.

Abbildung 4-52
Umkodieren von Zeichenkettenwerten

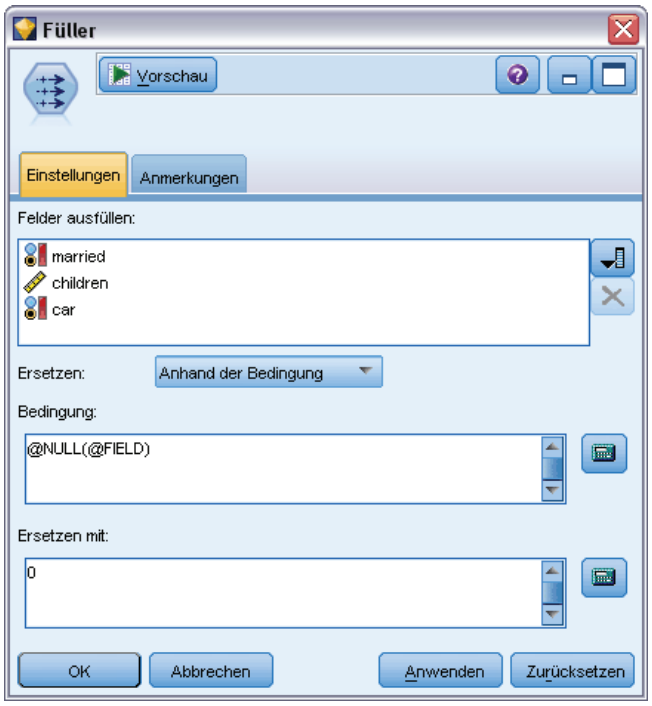


- Wählen Sie unter “Ableitungstyp” den entsprechenden Feldtyp aus (z. B. “Nominal” oder “Flag”).
- Legen Sie die Bedingungen für die Umkodierung der Werte fest. Beispielsweise können Sie den Wert auf 1 setzen, wenn Drug='drugA', auf 2, wenn Drug='drugB', usw.

Füllerknoten

Füllerknoten werden verwendet, um Feldwerte zu ersetzen und den Speichertyp zu ändern. Sie können auswählen, dass die Werte auf der Grundlage einer angegebenen CLEM-Bedingung ersetzt werden sollen, beispielsweise @BLANK(FIELD). Alternativ können Sie auswählen, dass alle Leerstellen oder Nullwerte mit einem bestimmten Wert ersetzt werden sollen. Füllerknoten werden zum Ersetzen fehlender Werte häufig in Verbindung mit dem Typknoten verwendet. Beispielsweise können Sie Leerstellen mit dem Mittelwert eines Felds ausfüllen, indem Sie einen Ausdruck wie @GLOBAL_MEAN angeben. Dieser Ausdruck füllt alle Leerzeichen mit dem durch einen Globalwerteknoten berechneten Mittelwert.

Abbildung 4-53
Dialogfeld "Füllerknoten"



Felder ausfüllen. Mit der Feldauswahl-Schaltfläche rechts neben dem Textfeld können Sie Felder aus den Daten-Sets auswählen, deren Werte untersucht und ersetzt werden. Standardmäßig werden die Werte in Abhängigkeit von den unten angegebenen Ausdrücken "Bedingung" und "Ersetzen mit" ersetzt. Sie können jedoch auch eine Alternative Ersetzungsmethode auswählen. Verwenden Sie dazu die unten stehenden Ersetzungsoptionen.

Hinweis: Bei der Auswahl mehrerer Felder für die Ersetzung mit einem benutzerdefinierten Wert müssen alle Feldtypen ähnliche sein (alle numerisch oder alle symbolisch).

Ersetzen. Hier können Sie auswählen, mit welcher der folgenden Methoden die Werte der ausgewählten Felder ersetzt werden sollen:

- **Anhand der Bedingung.** Diese Option aktiviert das Feld "Bedingung" und Expression Builder, damit Sie einen Ausdruck erstellen können, der als Bedingung für die Ersetzung mit dem angegebenen Wert verwendet werden kann.
- **Immer.** Ersetzt alle Werte für das ausgewählte Feld. Beispielsweise können Sie mit dieser Option den Speichertyp für "income" mit folgendem CLEM-Ausdruck in eine Zeichenkette konvertieren: (to_string(income)).
- **Leere Werte.** Ersetzt alle benutzerdefinierten leeren Werte im ausgewählten Feld. Die Standardbedingung @BLANK(@FIELD) wird zur Auswahl von Leerstellen verwendet. *Hinweis*: Mit der Registerkarte "Typen" im Quellenknoten oder mit einem Typknoten können Sie Leerstellen definieren.

- **Nullwerte.** Ersetzt alle systemdefinierten Nullwerte im ausgewählten Feld. Die Standardbedingung @NULL(@FIELD) wird zur Auswahl von Nullwerten verwendet.
- **Leere Werte und Nullwerte.** Ersetzt sowohl leere Werte als auch systemdefinierte Nullen im ausgewählten Feld. Diese Option ist hilfreich, wenn Sie sich nicht sicher sind, ob Nullen als fehlende Werte definiert sind oder nicht.

Bedingungen. Diese Option ist verfügbar, wenn Sie die Option Anhand der Bedingung ausgewählt haben. In diesem Textfeld können Sie einen CLEM-Ausdruck zur Evaluierung der ausgewählten Felder angeben. Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder zu öffnen.

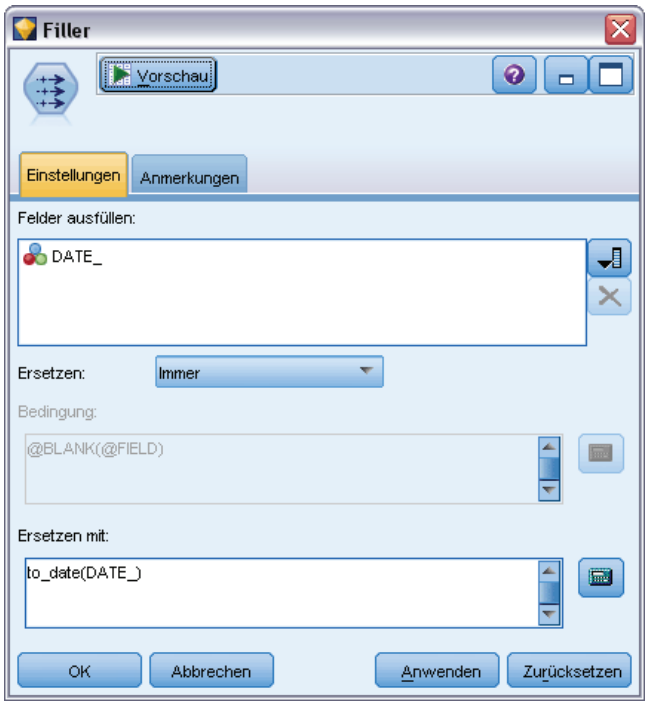
Ersetzen mit. Dient zur Angabe eines CLEM-Ausdrucks, um den ausgewählten Feldern einen neuen Wert zuzuweisen. Außerdem können Sie den Wert durch einen Nullwert ersetzen, indem Sie undef in das Textfeld eingeben. Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder zu öffnen.

Hinweis: Wenn die ausgewählten Felder den Typ "Zeichenkette" aufweisen, sollten Sie sie mit einem Zeichenkettenwert ersetzen. Die Verwendung des Standardwerts 0 oder eines anderen numerischen Werts als Ersatzwert für Zeichenkettenfelder führt zu einem Fehler.

Speichertypkonvertierung mithilfe des Füllerknotens

Mithilfe der Bedingung "Ersetzen" eines Füllerknotens können Sie problemlos den Feldspeichertyp für ein einzelnes Feld oder für mehrere Felder ändern. Beispiel: Mithilfe der Konvertierungsfunktion to_integer könnten Sie *income* von einer Zeichenkette in eine ganze Zahl konvertieren. Dazu wird folgender CLEM-Ausdruck verwendet: to_integer(income).

Abbildung 4-54
Verwendung eines Füllerknotens zur Konvertierung des Feldspeichertyps



Sie können die verfügbaren Funktionen anzeigen und mit Expression Builder automatisch einen CLEM-Ausdruck erstellen. Wählen Sie in der Dropdown-Liste "Funktionen" die Option Konvertierung aus, um eine Liste der Funktionen für die Konvertierung des Speichertyps anzuzeigen. Folgende Konvertierungsfunktionen stehen zur Verfügung:

- to_integer(ITEM)
- to_real(ITEM)
- to_number(ITEM)
- to_string(ITEM)
- to_time(ITEM)
- to_timestamp(ITEM)
- to_date(ITEM)
- to_datetime(ITEM)

Konvertieren von Datums- und Zeitwerten. Beachten Sie, dass die Konvertierungsfunktionen (und alle anderen Funktionen, für die ein spezieller Eingabetyp, wie beispielsweise ein Wert für Datum oder Uhrzeit, erforderlich ist) von den aktuell im Dialogfeld für die Stream-Optionen angegebenen Formaten abhängen. Wenn Sie beispielsweise ein Zeichenkettenfeld mit den Werten *Jan 2003*, *Feb 2003* usw. in einen Datumsspeicher konvertieren möchten, wählen Sie MON JJJJ als Standard-Datumsformat für den Stream aus.

Konvertierungsfunktionen sind auch im Ableitungsknoten zur temporären Konvertierung während einer Ableitungsberechnung verfügbar. Mit dem Ableitungsknoten können Sie auch andere Bearbeitungen vornehmen, wie beispielsweise die Umkodierung von Zeichenkettenfeldern mit kategorialen Werten. Für weitere Informationen siehe Thema Umkodieren von Werten mit dem Ableitungsknoten auf S. 178.

Anonymisierungsknoten

Mit dem Anonymisierungsknoten können Sie Feldnamen und/oder Feldwerte verschleiern, wenn Sie mit Daten arbeiten, die in ein Modell weiter unten im Knoten aufgenommen werden sollen. Auf diese Weise kann das generierte Modell frei verteilt werden (beispielsweise an den technischen Support), ohne dass die Gefahr besteht, dass unbefugte Benutzer vertrauliche Daten wie beispielsweise Personalakten oder Patientenakten anzeigen können.

Je nachdem, wo Sie den Anonymisierungsknoten im Stream platzieren, müssen Sie möglicherweise Änderungen an anderen Knoten vornehmen. Wenn Sie beispielsweise einen Anonymisierungsknoten oberhalb eines Auswahlknotens im Stream einfügen, müssen die Auswahlkriterien des Auswahlknotens geändert werden, wenn sie für Werte gelten sollen, die nun anonymisiert wurden.

Die für die Anonymisierung verwendete Methode beruht auf mehreren Faktoren. Bei Feldnamen und allen Feldwerten mit Ausnahme von stetigen Messniveaus werden die Daten durch eine Zeichenkette der folgenden Form ersetzt:

*prefix*_S*n*

Dabei ist *Präfix*_ entweder eine vom Benutzer angegebene Zeichenkette oder die Standardzeichenkette anon_ und *n* ist ein ganzzahliger Wert, der bei 0 beginnt und für jeden eindeutigen Wert erhöht wird (z. B. anon_S0, anon_S1 usw.).

Feldwerte mit dem Typ “Stetig” müssen transformiert werden, da sich numerische Bereiche mit ganzen oder reellen Zahlen befassen und nicht mit Zeichenketten. Daher können sie nur durch Transformation des Bereichs in einen anderen Bereich anonymisiert werden, wodurch die ursprünglichen Daten verschleiert werden. Die Transformation von Wert *x* im Bereich wird wie folgt durchgeführt:

A*(*x* + B)

Dabei gilt:

A ist ein Skalierungsfaktor, der größer als 0 sein muss.

B ist ein Verschiebungs-Offset, das zu den Werten addiert wird.

Beispiel

Bei einem Feld *ALTER*, bei dem der Skalierungsfaktor *A* auf 7 und das Verschiebungs-Offset *B* auf 3 gesetzt ist, werden die Werte für *ALTER* wie folgt transformiert:

7*(ALTER + 3)

Festlegen der Optionen für den Anonymisierungsknoten

Hier können Sie auswählen, bei welchen Feldern die Werte weiter unten im Stream verschleiert werden sollen.

Beachten Sie, dass die Datenfelder oberhalb des Anonymisierungsknotens instanziiert werden müssen, damit Anonymisierungsoperationen durchgeführt werden können. Sie können die Daten durch Klicken auf die Schaltfläche Werte lesen in einem Typknoten bzw. auf der Registerkarte "Typen" eines Quellenknotens instanziieren.

Abbildung 4-55
Festlegen der Anonymisierungsoptionen

Feld. Listet die Felder im aktuellen Daten-Set auf. Wenn bereits Feldnamen anonymisiert wurden, werden die anonymisierten Namen hier angezeigt.

Messung. Das Messniveau des Felds.

Werte anonymisieren. Wählen Sie ein oder mehrere Felder aus, klicken Sie auf diese Spalte und wählen Sie die Option Ja, um die Feldwerte mit dem Standardpräfix anon_ zu anonymisieren; wählen Sie Angeben, um ein Dialogfeld anzuzeigen, in dem Sie entweder Ihr eigenes Präfix eingeben oder – bei Feldwerten des Typs *Stetig* – angeben können, ob bei der Transformation der Feldwerte Zufallswerte oder vom Benutzer angegebene Werte verwendet werden sollen. Beachten Sie, dass *stetige* und nicht-*stetige* Feldtypen nicht in derselben Operation zusammen mit Daten eines anderen Typs angegeben werden können; Sie müssen diesen Vorgang separat für die einzelnen Feldtypen durchführen.

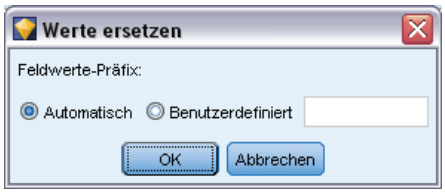
Aktuelle Felder anzeigen. Wählen Sie diese Option aus, um die Felder für Daten-Sets anzuzeigen, die aktiv mit dem Anonymisierungsknoten verbunden sind. Diese Option ist standardmäßig aktiviert.

Nicht verwendete Feldeinstellungen anzeigen. Wählen Sie diese Option aus, um die Felder für Daten-Sets anzuzeigen, die zu einem früheren Zeitpunkt (jetzt jedoch nicht mehr) mit dem Knoten verbunden waren. Diese Option wird vor allem beim Kopieren von Knoten aus einem Stream in einen anderen oder beim Speichern und erneuten Laden von Knoten eingesetzt.

Angabe der Vorgehensweise bei der Anonymisierung von Feldwerten

Im Dialogfeld "Werte ersetzen" können Sie auswählen, ob das Standardpräfix für anonymisierte Feldwerte oder ein benutzerdefiniertes Präfix verwendet werden soll. Wenn Sie in diesem Dialogfeld auf OK klicken, ändert sich die Einstellung von "Werte anonymisieren" auf der Registerkarte "Einstellungen" für die ausgewählten Felder in Ja.

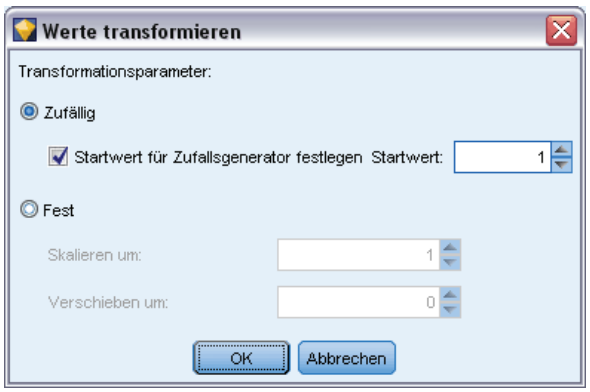
Abbildung 4-56
Dialogfeld "Werte ersetzen"



Feldwerte-Präfix. Das Standardpräfix für anonymisierte Feldwerte lautet anon_. Falls Sie ein anderes Präfix verwenden möchten, wählen Sie die Option Benutzerdefiniert und geben Sie das gewünschte Präfix ein.

Das Dialogfeld "Werte transformieren" wird nur für Felder des Typs "Stetig" angezeigt und ermöglicht Ihnen anzugeben, ob bei der Transformation der Feldwerte Zufallswerte oder vom Benutzer angegebene Werte verwendet werden sollen.

Abbildung 4-57
Dialogfeld "Werte transformieren"



Zufällig. Wählen Sie diese Option, um Zufallswerte für die Transformation zu verwenden. Startwert für Zufallsgenerator festlegen ist standardmäßig ausgewählt; geben Sie einen Wert im Feld Startwert an oder verwenden Sie den Standardwert.

Fest. Wählen Sie diese Option, um Ihre eigenen Werte für die Transformation anzugeben.

- **Skalieren um.** Der Wert, mit dem die Feldwerte in der Transformation multipliziert werden. Der Mindestwert ist 1; der Höchstwert ist normalerweise 10; er kann jedoch gesenkt werden, um einen Überlauf zu vermeiden.
- **Verschieben um.** Der Wert, der in der Transformation zu den Feldwerten addiert wird. Der Mindestwert ist 0; der Höchstwert ist normalerweise 1000; er kann jedoch gesenkt werden, um einen Überlauf zu vermeiden.

Anonymisieren von Feldwerten

Bei den auf der Registerkarte “Einstellungen” für die Anonymisierung ausgewählten Feldern werden die Werte in folgenden Fällen anonymisiert:

- Wenn Sie den Stream ausführen, der den Anonymisierungsknoten enthält
- Wenn Sie eine Vorschau der Werte anzeigen

Um eine Vorschau der Werte anzuzeigen, klicken Sie auf der Registerkarte “Anonymisierte Werte” auf die Schaltfläche Werte anonymisieren. Wählen Sie als Nächstes einen Feldnamen in der Dropdown-Liste aus.

Beim Messniveau “Stetig” werden folgende Elemente angezeigt:

- Mindest- und Höchstwert des ursprünglichen Bereichs
- Die zur Transformation der Werte verwendete Gleichung

Abbildung 4-58
Anonymisieren von Feldwerten

Bei einem anderen Messniveau als “Stetig” werden der ursprüngliche und der anonymisierte Wert für das betreffende Feld angezeigt.

Abbildung 4-59
Anonymisieren von Feldwerten

Wenn die Anzeige einen gelben Hintergrund aufweist, deutet dies darauf hin, dass sich entweder die Einstellung für das ausgewählte Feld seit der letzten Anonymisierung geändert hat oder dass Änderungen an den Daten oberhalb des Anonymisierungsknotens vorgenommen wurden, sodass die anonymisierten Werte möglicherweise nicht mehr korrekt sind. Das aktuelle Werte-Set wird angezeigt. Klicken Sie erneut auf die Schaltfläche Werte anonymisieren, um ein neues Werte-Set entsprechend der aktuellen Einstellung zu generieren.

Werte anonymisieren. Erstellt anonymisierte Werte für das ausgewählte Feld und zeigt diese in der Tabelle an. Bei Verwendung von Zufallsstartwerten für ein Feld vom Typ “Stetig” wird durch Klicken auf diese Schaltfläche jedes Mal ein anderes Werte-Set erstellt.

Werte löschen. Löscht die ursprünglichen und die anonymisierten Werte aus der Tabelle.

Umkodierungsknoten

Der Umkodierungsknoten ermöglicht die Transformation eines Sets kategorialer Werte in ein anderes. Die Umkodierung dient zur Reduzierung von Kategorien bzw. Neugruppierung von Daten für die Analyse. Beispielsweise können Sie die Werte für *Produkt* in drei Gruppen umkodieren, wie zum Beispiel *Küchenzubehör*, *Bad und Bettwäsche* sowie *Elektrogeräte*. Diese Operation wird häufig direkt aus einem Verteilungsknoten ausgeführt. Dazu werden die Werte gruppiert und ein Umkodierungsknoten wird erstellt. Für weitere Informationen siehe Thema Verwendung von Verteilungsknoten in Kapitel 5 auf S. 315.

Die Umkodierung kann für ein oder mehrere symbolische Felder durchgeführt werden. Außerdem können Sie festlegen, dass die neuen Werte für das bestehende Feld eingesetzt werden sollen, oder ein neues Feld generieren.

Vor der Verwendung eines Umkodierungsknotens sollten Sie überlegen, ob ein anderer Feldoperationsknoten für die betreffende Aufgabe geeigneter ist:

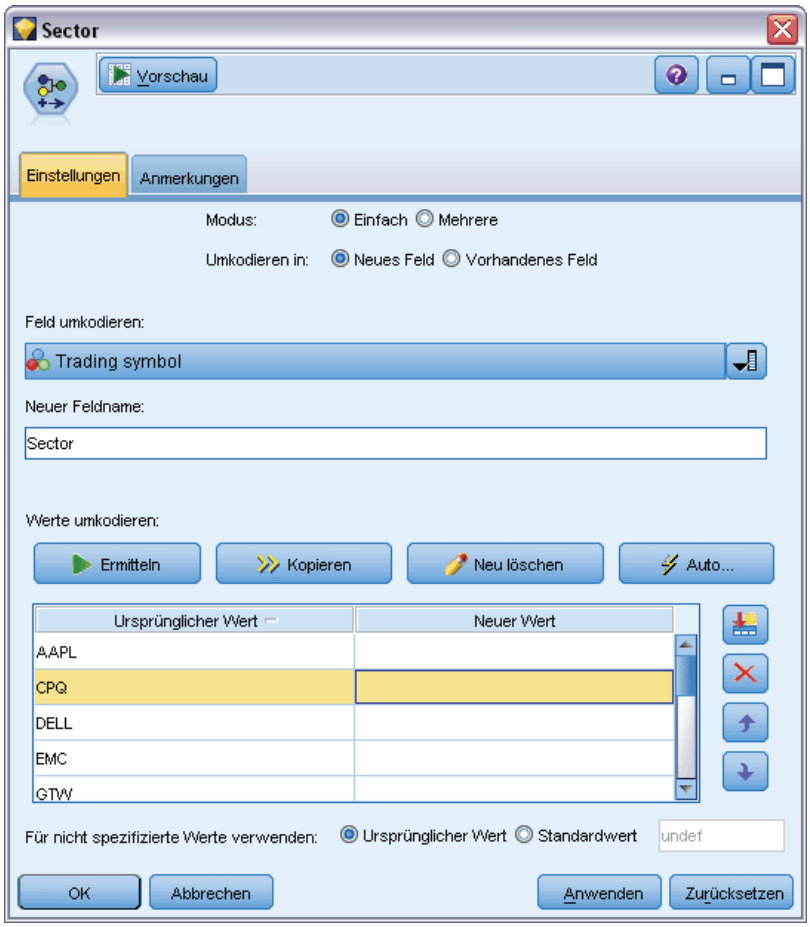
- Um numerische Bereiche automatisch in Sets (z. B Ränge oder Prozentsätze) umzuwandeln, sollten Sie einen Klassierknoten verwenden. Für weitere Informationen siehe Thema Klassierknoten auf S. 192.
- Wenn Sie numerische Bereiche manuell in Sets umwandeln möchten, sollten Sie einen Ableitungsknoten verwenden. Beispiel: Angenommen, Sie möchten Gehaltswerte in spezielle Gehaltsbereichskategorien zusammenfassen, dann sollten Sie jede Kategorie manuell mithilfe eines Ableitungsknotens definieren.
- Um eines oder mehrere Flag-Felder auf der Grundlage der Werte eines kategorialen Felds, beispielsweise *Hypothektyp*, zu erstellen, sollten Sie einen Dichotomknoten verwenden.
- Soll ein kategoriales Feld in ein Feld mit numerischem Speichertyp konvertiert werden, verwenden Sie einen Ableitungsknoten. So können Sie beispielsweise *Nein* und *Ja* in die Werte 0 und 1 konvertieren. Für weitere Informationen siehe Thema Umkodieren von Werten mit dem Ableitungsknoten auf S. 178.

Festlegen der Optionen für den Umkodierungsknoten

Die Verwendung des Umkodierungsknotens erfolgt in drei Schritten:

- Wählen Sie zunächst aus, ob Sie mehrere Felder umkodieren möchten oder nur ein einziges Feld.
- Wählen Sie als Nächstes aus, ob die Umkodierung in das bestehende Feld erfolgen oder ob ein neues Feld erstellt werden soll.
- Verwenden Sie schließlich die dynamischen Optionen im Dialogfeld “Umkodierungsknoten”, um die Sets wunschgemäß zuzuordnen.

Abbildung 4-60
Dialogfeld "Umkodierungsknoten"



Modus. Wählen Sie Einfach aus, um die Kategorien für ein einzelnes Feld umzukodieren. Wählen Sie Mehrere aus, um Optionen zu aktivieren, die die Transformation von mehreren Feldern gleichzeitig erlauben.

Umkodieren in. Wählen Sie Neues Feld aus, um das ursprüngliche nominale Feld beizubehalten und ein weiteres Feld abzuleiten, das die umkodierten Werte enthält. Wählen Sie die Option Vorhandenes Feld, um die Werte im ursprünglichen Feld mit den neuen Klassifikationen zu überschreiben. Dies ist im Grunde ein "Füll"-Vorgang.

Nach der Angabe des Modus und der Ersetzungsoptionen müssen Sie das Transformationsfeld auswählen und mithilfe der dynamischen Optionen in der unteren Hälfte des Dialogfelds die neuen Klassifikationswerte angeben. Diese Optionen variieren in Abhängigkeit vom oben ausgewählten Modus.

Umkodierungsfeld(er) Mit der Feldauswahl-Schaltfläche auf der rechten Seite können Sie eines (Modus "Einfach") oder mehrere (Modus "Mehrere") kategoriale Felder auswählen.

Neuer Feldname. Dient zur Angabe eines Namens für das neue nominale Feld mit den umkodierten Werten. Diese Option ist nur im Modus "Einfach" verfügbar, wenn oben Neues Feld ausgewählt wurde. Bei Auswahl von Vorhandenes Feld wird der ursprüngliche Feldname beibehalten. Im Modus "Mehrere" wird diese Option mit Steuerelementen zur Angabe einer Erweiterung für jedes neue Feld ersetzt. Für weitere Informationen siehe Thema Umkodieren mehrerer Felder auf S. 191.

Werte umkodieren. Diese Tabelle ermöglicht eine klare Zuordnung von alten Set-Werten zu den hier angegebenen.

- **Ursprünglicher Wert.** Diese Spalte listet bestehende Werte für die Auswahlfelder auf.
- **Neuer Wert.** In dieser Spalte können Sie neue Kategoriewerte eingeben oder einen aus der Dropdown-Liste auswählen. Wenn Sie automatisch einen Umkodierungsknoten mit Werten aus einem Verteilungsdiagramm generieren, sind diese Werte in der Dropdown-Liste enthalten. Dadurch können Sie schnell und einfach bestehende Werte einem bekannten Set von Werten zuordnen. Beispiel: Gesundheitsorganisationen gruppieren Diagnosen manchmal unterschiedlich je nach Netzwerk oder Ländereinstellung. Nach einer Fusion oder Übernahme müssen alle Beteiligten die neuen oder sogar die bereits vorhandenen Daten einheitlich klassifizieren. Anstatt jeden Zielwert aus einer langen Liste einzeln manuell einzugeben, können Sie die Master-Liste der Werte in IBM® SPSS® Modeler einlesen, ein Verteilungsdiagramm für das Feld *Diagnose* ausführen und einen Umkodierungsknoten für dieses Feld direkt aus dem Diagramm erstellen. Dadurch werden alle Zielwerte für die Diagnose in der Dropdown-Liste "Neue Werte" verfügbar.

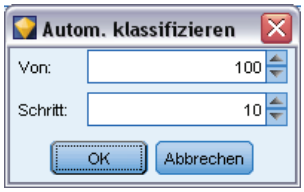
▶ Klicken Sie auf Ermitteln, um die ursprünglichen Werte für ein oder mehrere oben ausgewählte Felder zu lesen.

▶ Klicken Sie auf Kopieren, um für noch nicht zugeordnete Felder die ursprünglichen Werte in die Spalte *Neuer Wert* einzufügen. Die nicht zugeordneten ursprünglichen Werte werden in die Dropdown-Liste aufgenommen.

▶ Klicken Sie auf Neue löschen, um alle Spezifikationen in der Spalte *Neuer Wert* zu löschen. *Hinweis*: Mit dieser Option werden die Werte nicht aus der Dropdown-Liste gelöscht.

▶ Klicken Sie auf Automatisch, um automatisch aufeinander folgende ganze Zahlen für jeden der ursprünglichen Werte zu erstellen. Nur ganzzahlige Werte (keine reellen Werte wie 1,5; 2,5 usw.) können generiert werden.

Abbildung 4-61
Dialogfeld für die automatische Klassifizierung



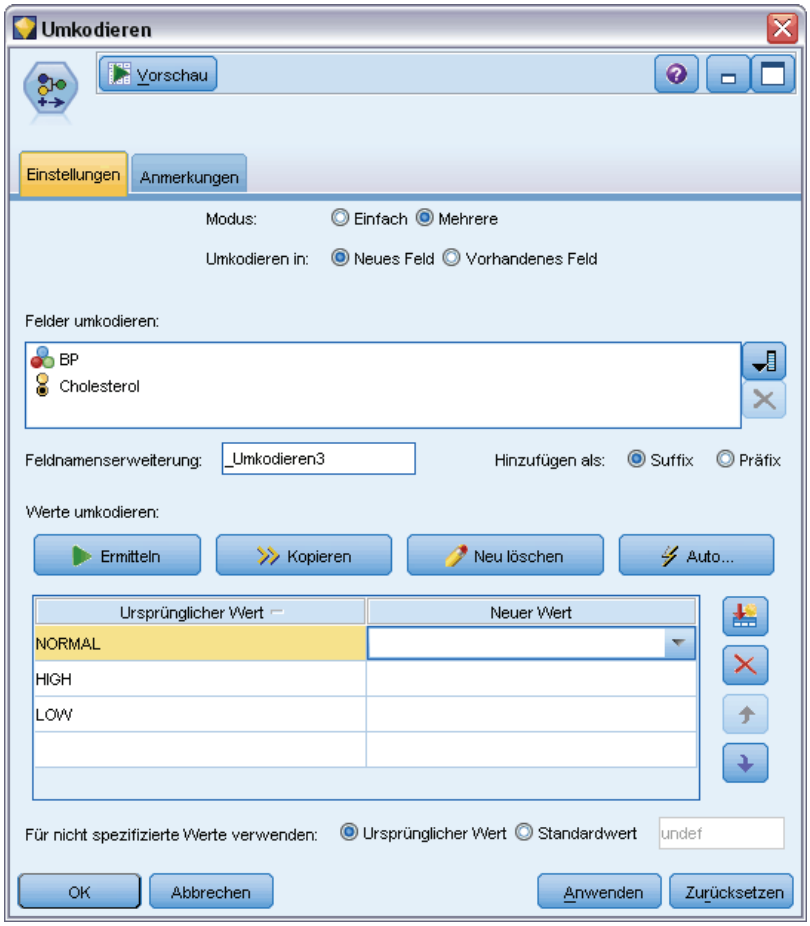
Sie können beispielsweise automatisch fortlaufende Produkt-IDs für Produktnamen erstellen oder Kursnummern für das Lehrangebot einer Universität. Diese Funktion entspricht der Transformation "Automatisch umkodieren" für Sets in IBM® SPSS® Statistics.

Für nicht spezifizierte Werte verwenden. Diese Option wird verwendet, um nicht spezifizierte Werte in das neue Feld einzutragen. Sie können entweder auswählen, dass der ursprüngliche Wert beibehalten werden soll, indem Sie Originalwert auswählen, oder einen Standardwert angeben.

Umkodieren mehrerer Felder

Um die Kategoriewerte für mehrere Felder gleichzeitig zuzuordnen, legen Sie als Modus Mehrere fest. Dadurch werden neue Einstellungen im Dialogfeld "Umkodieren" aktiviert. Diese werden im Folgenden beschrieben.

Abbildung 4-62
Dynamische Dialogfeldoptionen für die Umkodierung mehrerer Felder



Felder umkodieren. Mit der Feldauswahl-Schaltfläche auf der rechten Seite können Sie die Felder auswählen, die transformiert werden sollen. Mithilfe der Schaltfläche "Feldauswahl" können Sie alle Felder gleichzeitig auswählen oder Felder mit gleichem Typ wie "Nominal" oder "Flag".

Feldnamenerweiterung. Bei der gleichzeitigen Umkodierung mehrerer Felder ist es effizienter, anstatt einzelner Feldnamen eine gemeinsame Erweiterung anzugeben, um die alle neue Felder ergänzt werden. Geben Sie eine Erweiterung, wie beispielsweise `_recode`, an und wählen Sie

aus, ob diese Erweiterung an den Anfang oder an das Ende der ursprünglichen Feldnamen gestellt werden soll.

Speichertyp und Messniveau für umkodierte Felder

Der Umkodierungsknoten erstellt bei der Umkodierung immer ein nominales Feld. In einigen Fällen kann sich dadurch das Messniveau ändern, wenn der Umkodierungssmodus Vorhandenes Feld verwendet wird.

Der Speichertyp des neuen Felds (wie die Daten *gespeichert*, nicht wie sie *verwendet* werden) wird anhand der folgenden Optionen in der Registerkarte "Einstellungen" berechnet:

- Wenn bei nicht spezifizierten Werten die Verwendung eines Standardwerts festgelegt ist, wird der Speichertyp durch Untersuchung der neuen Werte und des Standardwerts sowie durch die Bestimmung des geeigneten Speichers ermittelt. Beispiel: Wenn alle Werte als ganze Zahlen analysiert werden können, weist das Feld den Speichertyp "Ganze Zahl" auf.
- Wenn bei nicht spezifizierten Werten die Verwendung der ursprünglichen Werte festgelegt ist, beruht der Speichertyp auf dem Speichertyp des ursprünglichen Felds. Wenn alle Werte als Speichertyp des ursprünglichen Felds analysiert werden können, wird dieser Speichertyp beibehalten; andernfalls wird der Speichertyp ermittelt, indem der geeignetste Speichertyp gesucht wird, der sowohl die alten als auch die neuen Werte umfasst. Beispiel: Bei der Umkodierung eines Sets mit ganzen Zahlen { 1, 2, 3, 4, 5 } mit der Umkodierung 4 => 0, 5 => 0 wird ein neues Set mit ganzen Zahlen { 1, 2, 3, 0 } generiert, wohingegen die Umkodierung 4 => "Over 3", 5 => "Over 3" das Zeichenketten-Set { "1", "2", "3", "Over 3" } ergibt.

Hinweis: Wenn der ursprüngliche Typ nicht instanziiert war, ist auch der neue Typ nicht instanziiert.

Klassierknoten

Mit dem Klassierknoten können Sie automatisch neue nominale Felder auf der Grundlage eines oder mehrerer bestehender stetiger Felder (numerischer Bereich) erstellen. Sie können beispielsweise ein stetiges Einkommensfeld in ein neues kategoriales Feld transformieren, das Einkommensgruppen gleicher Breite oder als Abweichungen vom Mittelwert enthält. Alternativ können Sie ein kategoriales "Supervisor"-Feld auswählen, damit die Stärke der ursprünglichen Assoziation zwischen den beiden Feldern erhalten bleibt.

Die Durchführung der Klassierung kann aus einer Reihe von Gründen nützlich sein. Hier einige Beispiele:

- **Algorithmusanforderungen.** Für bestimmte Algorithmen, beispielsweise "Naive Bayes" und "Logistische Regression", sind kategoriale Eingaben erforderlich.
- **Leistung.** Die Leistung von Algorithmen wie "Multinomiale logistische Regression" kann eventuell gesteigert werden, wenn die Anzahl der unterschiedlichen Werte der Eingabefelder reduziert wird. Sie könnten beispielsweise statt der ursprünglichen Werte den Median oder den Mittelwert für jede Klasse verwenden.
- **Datenschutz.** Vertrauliche persönliche Daten, wie beispielsweise Gehälter, können anstatt als tatsächliche Werte in Bereichen angegeben werden, um dem Datenschutz gerecht zu werden.

Es steht eine Reihe von Klassiermethoden zur Verfügung. Sobald Sie Klassen für das neue Feld erstellt haben, können Sie anhand der Trennwerte einen Ableitungsknoten generieren.

Vor der Verwendung eines Klassierknotens sollten Sie überlegen, ob ein anderes Verfahren für die betreffende Aufgabe geeigneter ist:

- Zur manuellen Angabe von Trennwerten für Kategorien, beispielsweise vordefinierte Gehaltsbereiche, verwenden Sie einen Ableitungsknoten. Für weitere Informationen siehe Thema Ableitungsknoten auf S. 167.
- Zur Erstellung neuer Kategorien für bestehende Sets verwenden Sie einen Umkodierungsknoten. Für weitere Informationen siehe Thema Umkodierungsknoten auf S. 187.

Umgang mit fehlenden Werten

Der Klassierknoten behandelt fehlende Werte folgendermaßen:

- **Vom Benutzer angegebene Leerstellen.** Fehlende Werte, die als Leerstellen angegeben sind, werden während der Transformierung aufgenommen. Wenn Sie beispielsweise –99 mithilfe des Typknotens als Leerwert gekennzeichnet haben, dann wird dieser Wert in den Klassiervorgang aufgenommen. Um Leerstellen beim Klassieren zu ignorieren, sollten Sie mithilfe eines Füllerknotens die Leerwerte durch den systemdefinierten Nullwert ersetzen.
- **Systemdefiniert fehlende Werte ($null$).** Nullwerte werden während der Klassiertransformation ignoriert und bleiben nach der Transformation weiterhin Nullwerte.

Auf der Registerkarte “Einstellungen” finden Sie Optionen für verfügbare Verfahren. Auf der Registerkarte “Ansicht” werden die Trennwerte angezeigt, die für die Daten ermittelt wurden, die den Knoten zuvor durchlaufen haben.

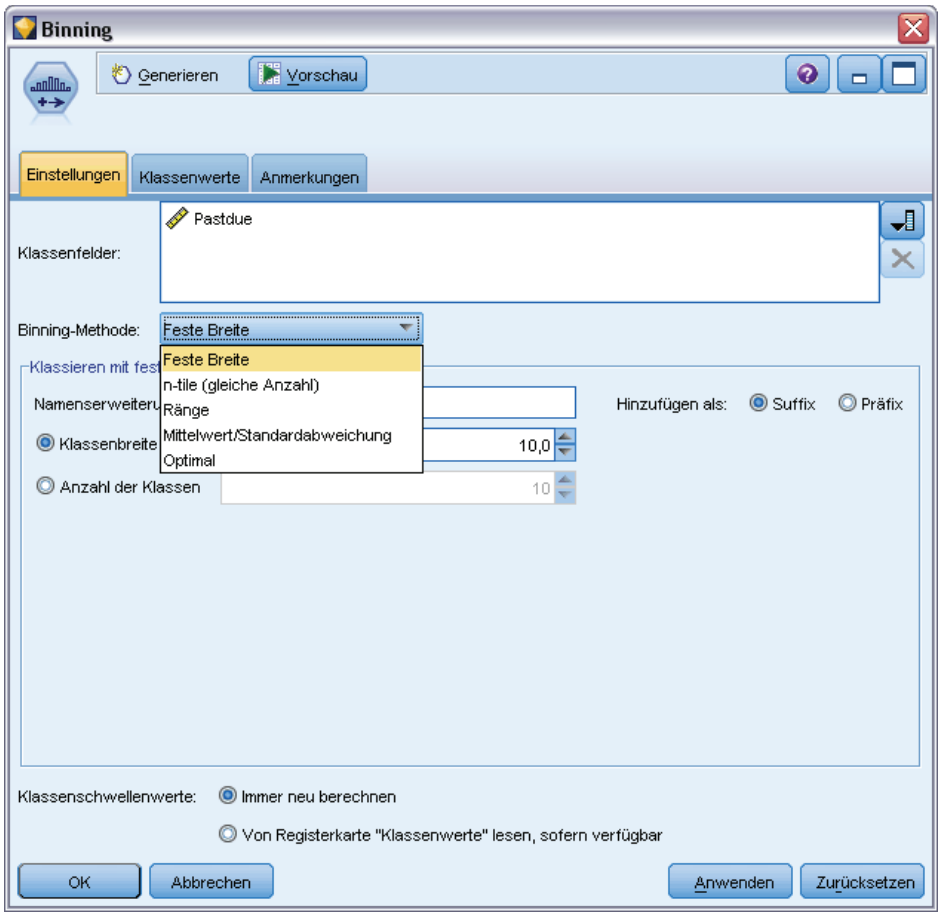
Festlegen der Optionen für den Klassierknoten

Mit dem Klassierknoten können Sie mit folgenden Verfahren automatisch Klassen (Kategorien) generieren:

- Klassieren mit fester Breite
- n-tile (gleiche Anzahl oder gleiche Summe)
- Mittelwert und Standardabweichung
- Ränge
- Optimiert in Bezug auf ein kategoriales “Supervisor”-Feld

Der untere Teil des Dialogfelds ändert sich dynamisch in Abhängigkeit von der ausgewählten Klassiermethode.

Abbildung 4-63
Dialogfeld "Klassierknoten", Registerkarte "Einstellungen"



Klassenfelder. Stetige Felder (numerischer Bereich) mit ausstehender Transformation werden hier angezeigt. Mit dem Klassierknoten können Sie mehrere Felder gleichzeitig klassieren. Zum Hinzufügen bzw. Entfernen von Feldern dienen die Schaltflächen auf der rechten Seite.

Klassiermethode. Dient zur Auswahl der Methode, die zur Ermittlung der Trennwerte für die neuen Feldklassen (Kategorien) verwendet werden. In den nachfolgenden Themenabschnitten werden die jeweils in den einzelnen Fällen verfügbaren Optionen behandelt.

Klassenschwellenwerte. Gibt an, wie die Klassenschwellenwerte berechnet werden.

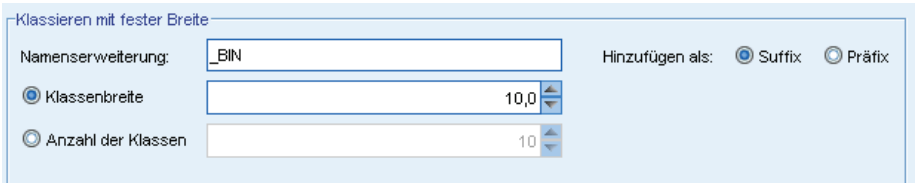
- **Immer neu berechnen.** Trennwerte und Klassenzuweisungen werden jedes Mal neu berechnet, wenn der Knoten ausgeführt wird.
- **Von Registerkarte "Klassenwerte" lesen, sofern verfügbar.** Trennwerte und Klassenzuweisungen werden nur bei Bedarf berechnet (beispielsweise, wenn neue Daten hinzugefügt wurden).

In den folgenden Themenabschnitten werden Optionen für die verfügbaren Klassiermethoden erörtert.

Klassen mit fester Breite

Wenn Sie als Klassiermethode Feste Breite auswählen, wird im Dialogfeld ein neues Options-Set angezeigt.

Abbildung 4-64
Dialogfeld "Klassierknoten" (Registerkarte "Einstellungen") mit Optionen für Klassen mit fester Breite



Namenserweiterung. Dient zur Angabe einer Erweiterung für die generierten Felder. *_BIN* ist die Standarderweiterung. Außerdem können Sie angeben, ob die Erweiterung am Anfang (Präfix) oder am Ende (Suffix) des Feldnamens eingefügt werden soll. Sie könnten beispielsweise ein neues Feld namens *Einkommen_BIN* erstellen.

Klassenbreite. Geben Sie einen Wert an (ganzzahlig oder reell), der zur Berechnung der "Breite" der Klasse verwendet werden soll. Sie können beispielsweise den Standardwert, 10 verwenden, um das Feld *Alter* zu klassieren. Da *Alter* einen Bereich von 18–65 aufweist, würden folgende Klassen generiert werden:

Tabelle 4-1
Klassen für Alter im Bereich 18–65

Klasse 1	Klasse 2	Klasse 3	Klasse 4	Klasse 5	Klasse 6
>=13 bis <23	>=23 bis <33	>=33 bis <43	>=43 bis <53	>=53 bis <63	>=63 bis <73

Der Start der Klassenintervalle wird aus dem niedrigsten gescannten Wert minus der Hälfte der (angegebenen) Klassenbreite berechnet. Beispiel: In den oben angegebenen Klassen wird der Wert 13 verwendet, um die Intervalle gemäß folgender Berechnung zu starten: 18 *[niedrigster Datenwert]* – 5 *[0,5 × (Klassenbreite von 10)]* = 13.

Anzahl der Klassen. Mit dieser Option können Sie eine ganze Zahl angeben, die zur Bestimmung der Anzahl der Klassen (Kategorien) mit fester Breite für die neuen Felder verwendet wird.

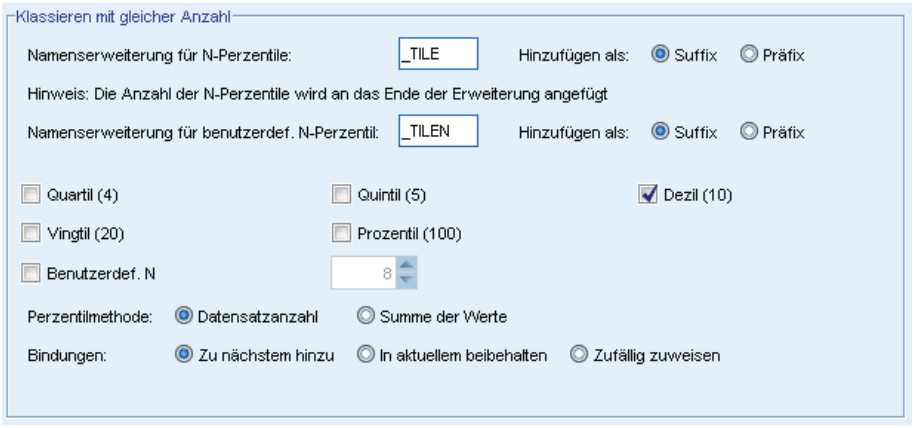
Nach der Ausführung des Klassierknotens in einem Stream können Sie die Klassenschwellenwerte anzeigen, die durch Klicken auf die Registerkarte Vorschau im Dialogfeld "Klassierknoten" generiert wurden. Für weitere Informationen siehe Thema Vorschau der generierten Klassen auf S. 202.

n-tile (gleiche Anzahl oder gleiche Summe)

Mit der Klassiermethode der n-tile erstellen Sie nominale Felder, mit denen die gescannten Datensätze so in Perzentil-Gruppen (oder Quartil-Gruppen, Dezil-Gruppen usw.) aufgeteilt werden können, dass jede Gruppe dieselbe Anzahl an Datensätzen aufweist oder dass die Summe der Werte in den einzelnen Gruppen gleich ist. Die Datensätze werden in aufsteigender Reihenfolge gemäß dem Wert des angegebenen Klassenfelds eingestuft. Datensätze mit dem niedrigsten Wert für die ausgewählte Klassenvariable erhalten somit den Rang 1, die nächste

Gruppe von Datensätzen den Rang 2 usw. Die Schwellenwerte für die einzelnen Klassen werden automatisch auf der Grundlage der Daten und der verwendeten n-til-Methode erzeugt.

Abbildung 4-65
Dialogfeld "Klassierknoten" (Registerkarte "Einstellungen") mit Optionen für Klassen mit gleicher Anzahl



Namenserweiterung für N-Perzentile. Geben Sie eine Erweiterung an, die für die mithilfe von Standard-N-Perzentilen generierten Felder verwendet wird. Die Standarderweiterung ist *_TILE* plus *N*; dabei steht *N* für die Nummer des Perzentils. Außerdem können Sie angeben, ob die Erweiterung am Anfang (Präfix) oder am Ende (Suffix) des Feldnamens eingefügt werden soll. Sie könnten beispielsweise ein neues Feld namens *Einkommen_BIN4* erstellen.

Namenserweiterung für benutzerdefinierte N-Perzentile. Dient zur Angabe einer Erweiterung für einen benutzerdefinierten n-til-Bereich. Die Standarderweiterung lautet *_TILEN*. *N* wird in diesem Fall *nicht* durch die benutzerdefinierte Zahl ersetzt.

Folgende N-Perzentile stehen zur Verfügung:

- **Quartil.** Generiert vier Klassen, die jeweils 25 % der Fälle enthalten.
- **Quintil.** Generiert fünf Klassen, die jeweils 20 % der Fälle enthalten.
- **Dezil.** Generiert zehn Klassen, die jeweils 10 % der Fälle enthalten.
- **Vingtil.** Generiert 20 Klassen, die jeweils 5 % der Fälle enthalten.
- **Perzentil.** Generiert 100 Klassen, die jeweils 1% der Fälle enthalten.
- **Benutzerdef. N** Wählen Sie diese Option aus, um die Anzahl der Klassen festzulegen. Der Wert 3 beispielsweise ergibt 3 in Bereiche eingeteilte Kategorien (2 Trennwerte), die jeweils 33,3 % der Fälle enthalten.

Falls weniger diskrete Werte in den Daten vorhanden sind als N-Perzentile angegeben wurden, werden nicht alle n-tile verwendet. In diesen Fällen spiegelt die neue Verteilung vermutlich die ursprüngliche Verteilung der Daten wider.

n-til-Methode. Legt fest, welche Methode für die Zuweisung der Datensätze zu den Klassen verwendet wird.

- **Datensatzanzahl.** Versucht, jeder Klasse eine gleich große Anzahl an Datensätzen zuzuweisen.
- **Summe.** Versucht die Datensätze so zu den Klassen zuzuweisen, dass die Summe der Werte in jeder Klasse gleich groß ist. Bei der Zielausrichtung von Absatzbemühungen sind Sie mit dieser Methode beispielsweise in der Lage, die Interessenten gemäß dem Wert je Datensatz zu Dezil-Gruppen zuzuweisen, wobei die Interessenten mit den höchsten Werten zur obersten Klasse gehören. Beispiel: Ein Pharmaunternehmen stuft die Ärzte gemäß der Anzahl ihrer Verschreibungen in Dezil-Gruppen ein. Jedes Dezil umfasst in etwa dieselbe Anzahl an Verschreibungen; die Anzahl der Personen, die diese Verschreibungen ausgestellt haben, ist jedoch nicht identisch. Die Personen mit den meisten Verschreibungen würden sich dabei in Dezil 10 wiederfinden. Hinweis: Bei dieser Vorgehensweise wird angenommen, dass alle Werte größer als null sind; ist dies nicht der Fall, können unerwartete Ergebnisse eintreten.

Bindungen. Eine Bindungsbedingung entsteht, wenn beide Seiten eines Trennwerts identisch sind. Wenn Sie beispielsweise Dezile zuweisen und mehr als 10 % der Datensätze denselben Wert im Klassenfeld aufweisen, können nicht alle Datensätze in derselben Klasse untergebracht werden, ohne den Schwellenwert entsprechend nach oben oder nach unten zu verschieben. Die Bindungen können wahlweise aufwärts in die nächste Klasse verschoben oder auch in der aktuellen Klase beibehalten werden; die Bindungen müssen jedoch in jedem Fall aufgelöst werden, sodass alle Datensätze mit identischen Werten in dieselbe Klasse fallen, auch wenn dadurch einige Klassen mehr Datensätze erhalten als erwartet. Auch die Schwellenwerte der nachfolgenden Klassen müssen angepasst werden, sodass die Werte für dieselbe Zahlengruppe unterschiedlich zugewiesen werden, je nach der verwendeten Methode zum Auflösen der Bindungen.

- **Zu nächstem hinzu.** Wählen Sie diese Option aus, um die Bindungswerte nach oben zur nächsten Klasse zu verschieben.
- **In aktuellem beibehalten.** Hiermit werden die Bindungswerte in der aktuellen (niedrigeren) Klasse belassen. Bei dieser Methode werden insgesamt ggf. weniger Klassen erstellt.
- **Zufällig zuweisen.** Wählen Sie diese Option aus, um die Bindungswerte nach dem Zufallsprinzip einer Klasse zuzuweisen. Dadurch wird versucht, die Anzahl der Datensätze in jeder Klasse gleich zu halten.

Beispiel: n-til-Einteilung nach Anzahl der Datensätze

Die nachstehende Tabelle zeigt, wie vereinfachte Feldwerte bei der n-til-Einteilung nach Anzahl der Datensätze als Quartile eingestuft werden. Die Ergebnisse sind dabei abhängig von der ausgewählten Bindungsoption.

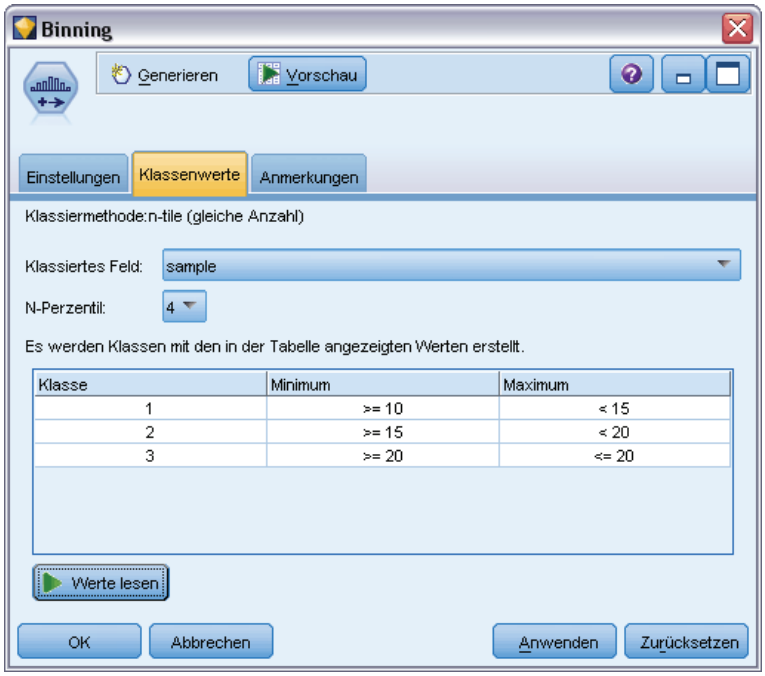
Werte	Zu nächstem hinzu	In aktuellem beibehalten
10	1	1
13	2	1
15	3	2
15	3	2
20	4	3

Die Anzahl der Elemente pro Klasse wird folgendermaßen berechnet:

Gesamtzahl der Werte/Anzahl der N-Perzentile

In dem vereinfachten Beispiel oben ist die erwünschte Anzahl der Elemente pro Klasse 1,25 (5 Werte/4 Quartile). Der Wert 13 (Wert Nummer 2) überspannt den gewünschten Schwellenwert für die Anzahl (1,25) und wird daher, je nach der ausgewählten Bindungsoption, unterschiedlich behandelt. Im Modus Zu nächstem hinzu wird er in Klasse 2 aufgenommen. Im Modus In aktuellem beibehalten wird er in Klasse 1 belassen, wodurch der Wertebereich für Klasse 4 so weit verschoben wird, dass er außerhalb des Bereichs der vorhandenen Datenwerte liegt. Daher werden nur drei Klassen erstellt und die Schwellenwerte für jede Klasse werden entsprechend angepasst.

Abbildung 4-66
Schwellenwerte für erzeugte Klassen

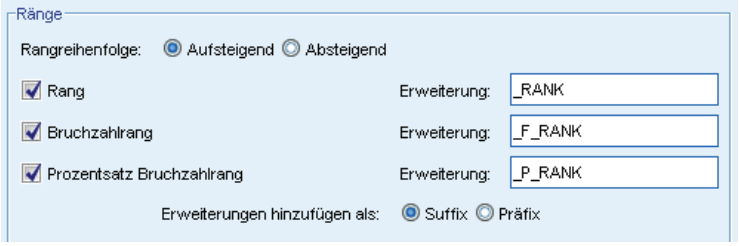


Hinweis: Die Geschwindigkeit beim Klassieren nach n-tilen kann ggf. durch Parallelverarbeitung gesteigert werden.

Rangfolge bilden

Wenn Sie als Klassiermethode Ränge auswählen, wird im Dialogfeld ein neues Options-Set angezeigt.

Abbildung 4-67
Dialogfeld "Klassierknoten" (Registerkarte "Einstellungen") mit Optionen für Ränge



Bei der Rangbildung werden neue Felder erstellt, die Ränge, Bruchzahlränge und Perzentilwerte für numerische Felder enthalten, je nach den unten angegebenen Optionen.

Rangreihenfolge. Wählen Sie Aufsteigend (der niedrigste Wert wird mit "1" gekennzeichnet) oder Absteigend (der höchste Wert wird mit "1" gekennzeichnet).

Rang. Mit dieser Option weisen Sie den Fällen in aufsteigender bzw. absteigender Reihenfolge (oben angegeben) Ränge zu. Der Bereich der Werte im neuen Feld ist 1–*N*. Dabei ist *N* die Anzahl der diskreten Werte im ursprünglichen Feld. Gebundenen Werten wird der Durchschnitt ihres Ranges zugewiesen.

Bruchzahlrang. Mit dieser Option weisen Sie Fällen Ränge zu, wobei der Wert des neuen Felds gleich dem Rang dividiert durch die Summe der Gewichtungen der nicht fehlenden Fälle ist. Bruchzahlränge fallen in den Bereich 0–1.

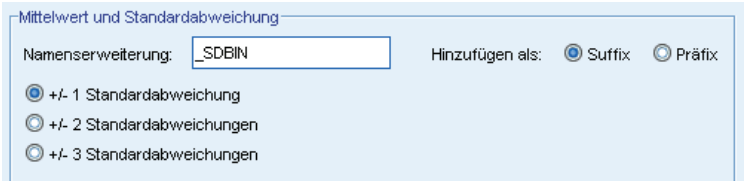
Prozentsatz Bruchzahlrang. Die einzelnen Ränge werden durch die Anzahl der Datensätze mit gültigen Werten dividiert und mit 100 multipliziert. Als Prozentsatz angegebene Bruchzahlränge fallen in den Bereich 1–100.

Erweiterung. Bei allen Rangoptionen können Sie benutzerdefinierte Erweiterungen erstellen und angeben, ob die Erweiterung am Anfang (Präfix) oder am Ende (Suffix) des Feldnamens eingefügt werden soll. Sie könnten beispielsweise ein neues Feld namens *Einkommen_P_RANK* erstellen.

Mittelwert/Standardabweichung

Wenn Sie als Klassiermethode Mittelwert/Standardabweichung auswählen, wird im Dialogfeld ein neues Options-Set angezeigt.

Abbildung 4-68
Dialogfeld "Klassierknoten" (Registerkarte "Einstellungen") mit Optionen für Mittelwert/Standardabweichung



Mit dieser Methode werden ein oder mehrere neue Felder mit in Bereiche eingeteilten Kategorien erstellt, die auf den Werten für Mittelwert und Standardabweichung der Verteilung für die angegebenen Felder beruhen. Wählen Sie die Anzahl der zu verwendenden Abweichungen aus.

Namenserweiterung. Dient zur Angabe einer Erweiterung für die generierten Felder. *_SDBIN* ist die Standarderweiterung. Außerdem können Sie angeben, ob die Erweiterung am Anfang (Präfix) oder am Ende (Suffix) des Feldnamens eingefügt werden soll. Sie könnten beispielsweise ein neues Feld namens *Einkommen_SDBIN* erstellen.

- **+/– 1 Standardabweichung.** Mit dieser Option werden drei Klassen erzeugt.
- **+/– 2 Standardabweichungen.** Mit dieser Option werden fünf Klassen erzeugt.
- **+/– 3 Standardabweichungen.** Mit dieser Option werden sieben Klassen erzeugt.

Die Auswahl von "+/– 1 Standardabweichung" beispielsweise führt zu den drei unten berechneten Klassen:

Klasse 1	Klasse 2	Klasse 3
x < (Mean – Std. Dev)	(Mean – Std. Dev) <= x <= (Mean + Std. Dev)	x > (Mean + Std. Dev)

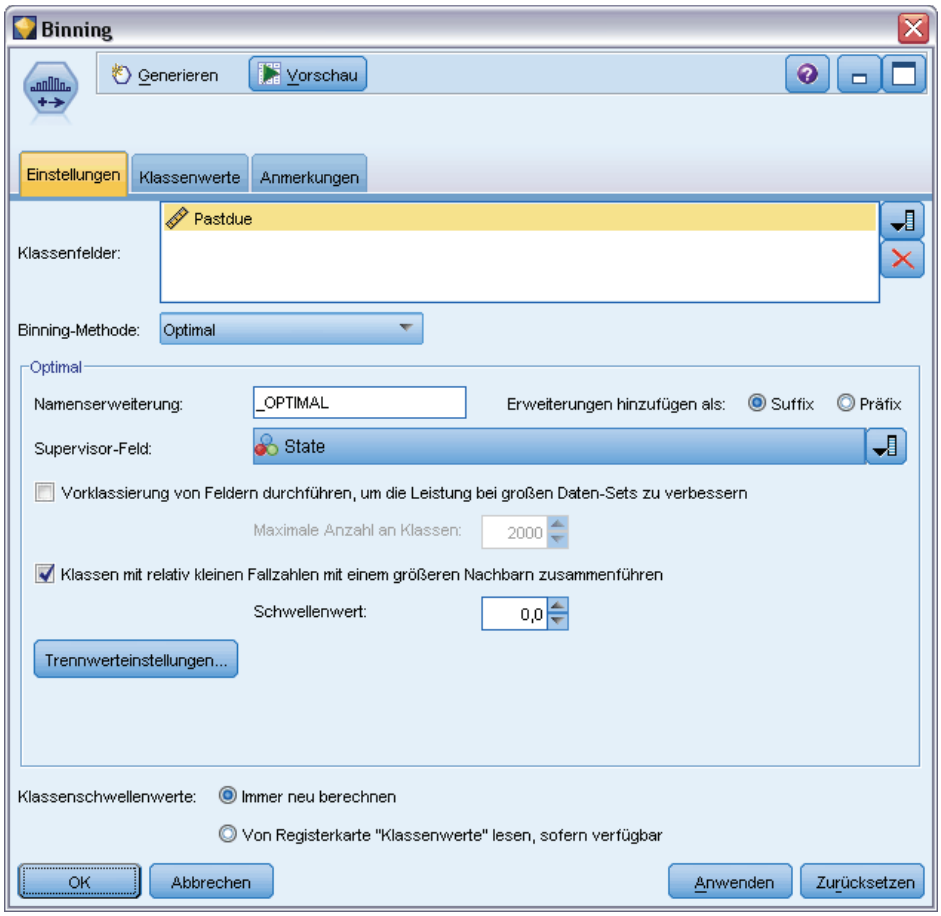
Bei einer Normalverteilung liegen 68 % der Fälle innerhalb einer Standardabweichung vom Mittelwert, 95 % innerhalb von zwei Standardabweichungen und 99 % innerhalb von drei Standardabweichungen. Das Erstellen von in Bereiche eingeteilten Kategorien auf der Grundlage der Standardabweichungen kann jedoch zu definierten Bereichen außerhalb des tatsächlichen Datenbereichs und sogar außerhalb des Bereichs der möglichen Datenwerte führen (z. B. ein negativer Gehaltsbereich).

Optimales Klassieren

Wenn das zu klassierende Feld eine starke Assoziation mit einem anderen kategorialen Feld aufweist, können Sie das kategoriale Feld als "Supervisor"-Feld auswählen, damit die Klassen so erstellt werden, dass die Stärke der ursprünglichen Assoziation zwischen den beiden Feldern beibehalten wird.

Beispiel: Angenommen, Sie haben mithilfe der Cluster-Analyse Statuswerte auf der Grundlage der Säumnisquoten für Eigenheimkredite gruppiert, mit den höchsten Quoten im ersten Cluster. In diesem Fall können Sie *Prozent nach Fälligkeit* und *Prozent der Zwangsvollstreckungen* als Klassenfelder und das vom Modell generierte Feld für die Cluster-Zugehörigkeit als Supervisor-Feld auswählen.

Abbildung 4-69
Optionen für optimales bzw. überwachtes Klassieren



Namenserweiterung. Geben Sie eine Erweiterung für die generierten Felder an und legen Sie fest, ob die Erweiterung am Anfang (Präfix) oder am Ende (Suffix) des Feldnamens eingefügt werden soll. Sie könnten beispielsweise ein neues Feld namens *überfällig_OPTIMAL* und ein weiteres namens *Zwangsvollstreckung_OPTIMAL* generieren.

Supervisor-Feld. Ein kategoriales Feld, das zur Erstellung der Klassen verwendet wird.

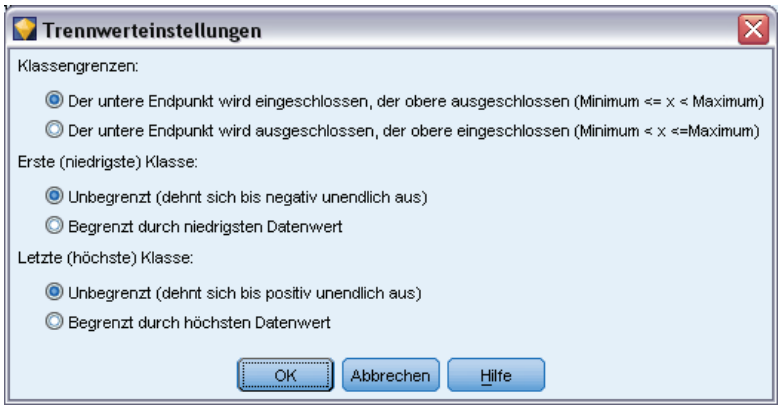
Vorklassierung von Feldern durchführen, um die Leistung bei großen Daten-Sets zu verbessern. Gibt an, ob eine Vorverarbeitung durchgeführt werden soll, um die optimale Klassierung zu rationalisieren. Bei dieser Gruppe werden Skalenwerte mithilfe einer einfachen nicht überwachten Klassiermethode in eine große Anzahl von Klassen gruppiert, die Werte innerhalb der einzelnen Klassen werden durch den Mittelwert repräsentiert und das Fallgewicht wird entsprechend angepasst, bevor mit dem überwachten Klassieren fortgefahren wird. In der Praxis bedeutet dies, dass bei diesem Verfahren zugunsten einer höheren Geschwindigkeit gewisse Einbußen bei der Präzision in Kauf genommen werden. Es empfiehlt sich für große Daten-Sets. Außerdem können Sie angeben, wie viele Klassen eine Variable nach der Vorverarbeitung maximal aufweisen soll, wenn diese Option verwendet wird.

Klassen mit relativ kleinen Fallzahlen mit einem größeren Nachbarn zusammenführen. Wenn diese Option aktiviert ist, wird eine Klasse zusammengeführt, falls das Verhältnis zwischen ihrer Größe (Anzahl der Fälle) und der Größe einer benachbarten Klasse kleiner ist als der angegebene Schwellenwert; beachten Sie, dass größere Schwellenwerte eine stärkere Zusammenführung mit sich bringen können.

Trennwerteinstellungen

Im Dialogfeld "Trennwerteinstellungen" können Sie erweiterte Optionen für den Algorithmus "Optimales Klassieren" angeben. Diese Optionen legen fest, wie der Algorithmus die Klassen unter Verwendung des Zielfelds berechnen soll.

Abbildung 4-70
Trennwerteinstellungen für optimales Klassieren



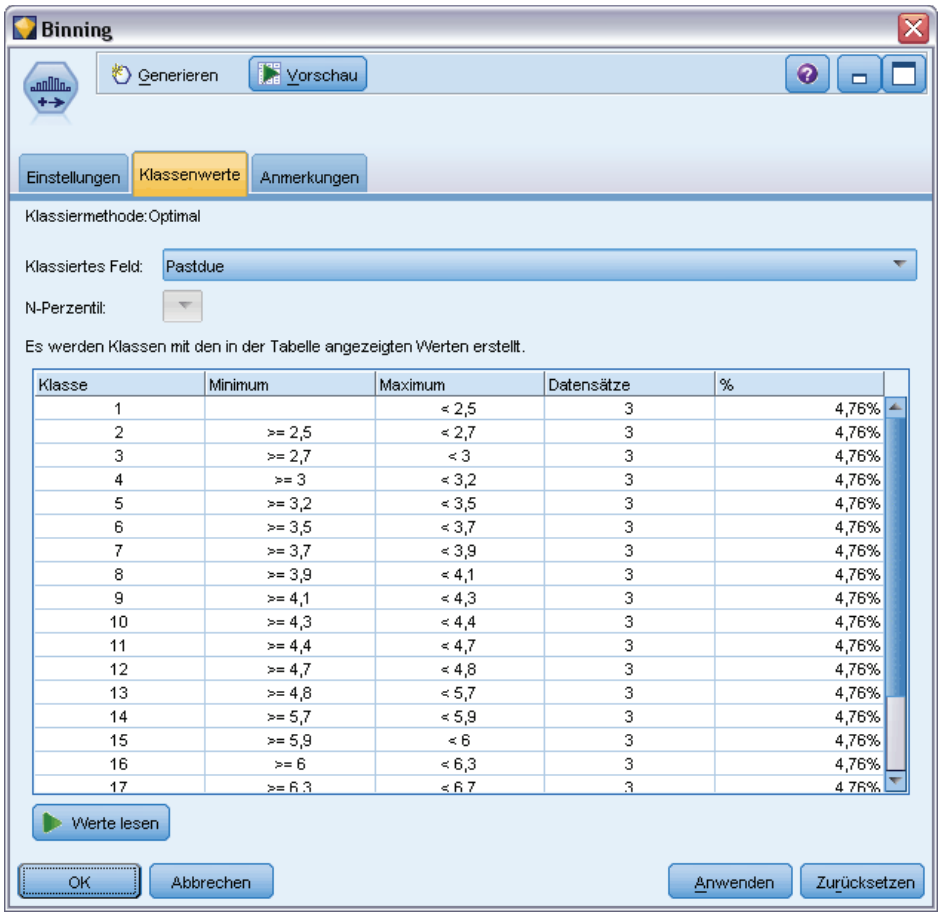
Klassengrenzen. Sie können angeben, ob der untere bzw. obere Endpunkt eingeschlossen (Minimum <= *x*) oder ausgeschlossen (Minimum < *x*) werden soll.

Erste und letzte Klasse. Für die erste und letzte Klasse können Sie angeben, ob die Klassen keine Begrenzung aufweisen (also gegen positiv bzw. negativ unendlich streben) sollen oder ob sie durch den niedrigsten bzw. höchsten Datenpunkt begrenzt werden sollen.

Vorschau der generierten Klassen

Auf der Registerkarte "Klassenwerte" im Klassierknoten können Sie die Schwellenwerte für die generierten Klassen anzeigen. Mithilfe des Menüs "Generieren" können Sie außerdem einen Ableitungsknoten generieren, mit dem Sie diese Schwellenwerte von einem Daten-Set auf ein anderes anwenden können.

Abbildung 4-71
Dialogfeld "Klassierknoten", Registerkarte "Klassenwerte"



Klassiertes Feld. Mithilfe der Dropdown-Liste können Sie ein Feld für die Anzeige auswählen. Für die angezeigten Feldnamen werden die ursprünglichen Feldnamen verwendet, um Verwirrung zu vermeiden.

N-Perzentil. Mithilfe der Dropdown-Liste können Sie ein N-Perzentil, beispielsweise 10 oder 100, für die Anzeige auswählen. Diese Option ist nur bei Klassen verfügbar, die mit der n-til-Methode (gleiche Anzahl oder gleiche Summe) generiert wurden.

Klassenschwellenwerte. Hier werden Schwellenwerte für die einzelnen generierten Klassen sowie die Anzahl der Datensätze angezeigt, die auf die einzelnen Klassen entfallen. Nur bei der Methode "Optimales Klassieren" wird die Anzahl der Datensätze in jeder Klasse als Prozentsatz der Gesamtzahl angezeigt. Beachten Sie, dass die Schwellenwerte beim Klassieren nach Rang nicht zum Einsatz kommen.

Werte lesen. Liest klassierte Werte aus dem Daten-Set. Beachten Sie, dass Schwellenwerte auch überschrieben werden, wenn neue Daten durch den Stream geleitet werden.

Erzeugen eines Ableitungsknotens

Im Menü “Generieren” können Sie einen Ableitungsknoten auf der Grundlage der aktuellen Schwellenwerte erstellen. Dies ist sinnvoll, um bewährte Klassenschwellenwerte aus einem Daten-Set auf ein anderes anzuwenden. Außerdem ist, sobald diese Aufteilungspunkte bekannt sind, eine Ableitung bei großen Daten-Sets effizienter (d. h. schneller) als ein Klassiervorgang.

Knoten “RFM-Analyse”

Mit dem Knoten “RFM-Analyse” (Recency-, Frequency-, Monetary-Analyse) können Sie quantitativ ermitteln, welche Kunden wahrscheinlich die besten sind, indem Sie untersuchen, wann sie zuletzt etwas von Ihnen erworben haben (Recency (Aktualität)), wie häufig sie eingekauft haben (Frequency (Häufigkeit)) und wie viel sie für alle Transaktionen zusammengenommen ausgegeben haben (Monetary (Geldwert)).

Der RDM-Analyse liegt zugrunde, dass Kunden, die einmal ein Produkt bzw. eine Dienstleistung erworben haben, dies mit größerer Wahrscheinlichkeit erneut tun. Die kategorisierten Kundendaten werden in eine Reihe von Klassen aufgeteilt, wobei die Klassierkriterien nach Bedarf angepasst werden können. In jeder Klasse wird den Kunden ein Score zugewiesen; diese Scores werden dann zu einem RFM-Gesamt-Score kombiniert. Der Score stellt die Zugehörigkeit des Kunden zu den für die einzelnen RFM-Parameter erstellten Klassen dar. Die klassierten Daten reichen möglicherweise für Ihre Bedürfnisse aus, indem sie beispielsweise die häufigsten Kunden mit den höchsten Werten ermitteln. Alternativ können sie zur weiteren Modellierung und Analyse einem Stream übergeben werden.

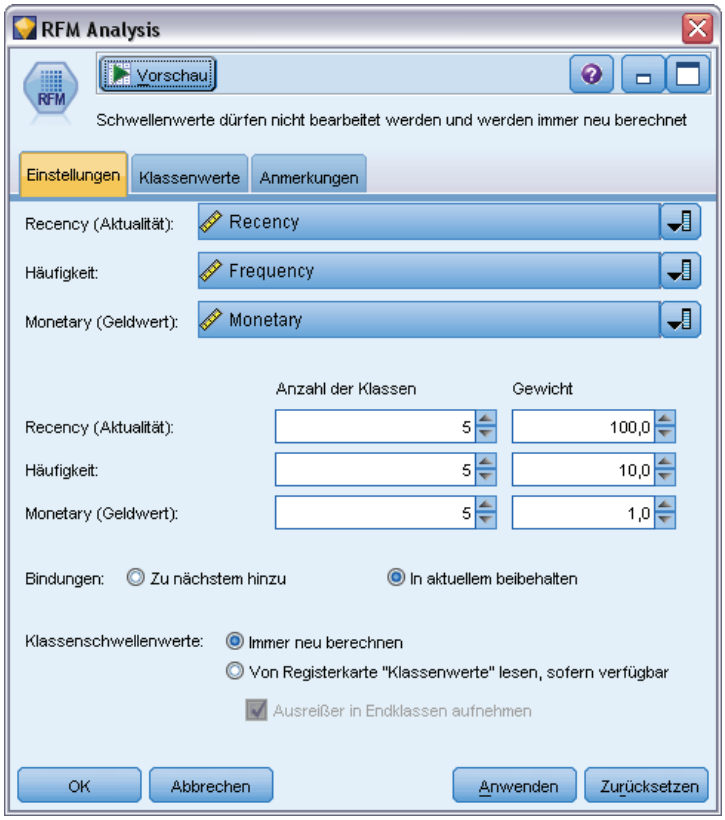
Beachten Sie: So nützlich die Möglichkeit zur Analyse und Rangeinteilung von RFM-Scores auch ist, müssen Sie sich bei der Verwendung doch bestimmter Faktoren bewusst sein. Es besteht die Versuchung, verstärkt auf die Kunden mit den höchsten Rangwertungen zuzugehen. Eine übermäßige Umwerbung dieser Kunden kann jedoch auch zu Verstimmungen und einen Rückgang in den Wiederholungsgeschäften führen. Außerdem sollte nicht vergessen werden, dass Kunden mit niedrigen Scores nicht vernachlässigt sollten, sondern dass es sinnvoller sein kann, sie zu pflegen, damit sie bessere Kunden werden. Umgekehrt deuten hohe Scores alleine, je nach Markt, noch nicht unbedingt auf gute Absatzchancen hin. So ist ein Kunde in Klasse 5 für “Recency”, der also vor sehr kurzer Zeit etwas erworben hat, nicht unbedingt der beste Zielkunde für Unternehmen, die teure, langlebige Produkte verkaufen, wie Autos oder Fernseher.

Hinweis: Je nachdem, wie Ihre Daten gespeichert sind, müssen Sie möglicherweise dem RFM-Analyseknoten einen RFM-Aggregatknoten vorschalten, um die Daten in ein brauchbares Format umzuwandeln. So müssen Eingabedaten beispielsweise im Kundenformat vorliegen, mit einer Zeile pro Kunden; wenn die Daten des Kunden in Transaktionsform vorliegen, können Sie durch Verwendung eines Knotens vom Typ “RFM-Aggregat” weiter oben im Stream die Felder für Aktualität, Häufigkeit und Geldwert ableiten. Für weitere Informationen siehe Thema RFM-Aggregatknoten in Kapitel 3 auf S. 85.

Die Knoten “RFM-Aggregat” und “RFM-Analyse” in IBM® SPSS® Modeler sind für die Verwendung einer unabhängigen Klassierung eingerichtet. Damit werden also Daten für jedes der Maße Aktualität, Häufigkeit und Geldwert in Ränge eingeteilt und klassiert, ohne Berücksichtigung ihrer Werte oder der beiden anderen Maße.

Knoten "RFM-Analyse" – Einstellungen

Abbildung 4-72
Festlegen der Optionen für die RFM-Analyse



Recency (Aktualität). Mithilfe der Feldauswahl-Schaltfläche (rechts neben dem Textfeld) können Sie das Aktualitätsfeld auswählen. Dabei kann es sich um ein Datum, einen Zeitstempel oder eine einfache Zahl handeln. Beachten Sie: Wenn ein Datum oder ein Zeitstempel das Datum der aktuellsten Transaktion angibt, wird der höchste Wert als der aktuellste betrachtet; wenn eine Nummer angegeben ist, steht sie für die Zeit, die seit der aktuellsten Transaktion verstrichen ist, und der niedrigste Wert wird als der aktuellste betrachtet.

Hinweis: Wenn dem Knoten "RFM-Analyse" im Stream der Knoten "RFM-Aggregat" vorangeht, sollten die vom Knoten "RFM-Aggregat" generierten Felder "Recency (Aktualität)", "Frequency (Häufigkeit)" und "Monetary (Geldwert)" im Knoten "RFM-Analyse" als Eingaben ausgewählt werden.

Häufigkeit. Wählen Sie mithilfe der Feldauswahl das zu verwendende Häufigkeitsfeld aus.

Monetary (Geldwert). Wählen Sie mithilfe der Feldauswahl das zu verwendende Feld für den Geldwert aus.

Anzahl der Klassen. Wählen Sie für jeden der drei Ausgabetypen aus, wie viele Klassen erstellt werden sollen. Der Standardwert ist 5.

Hinweis: Die Mindestzahl beträgt 2, die Höchstzahl 9 Klassen.

Gewicht. Standardmäßig erhalten bei der Berechnung der Scores die Aktualitätsdaten die größte Bedeutsamkeit; danach folgt die Häufigkeit und dann erst der Geldwert. Falls erforderlich können Sie die Gewichtung für eines oder mehrere dieser Elemente bearbeiten, um die Reihenfolge der Bedeutsamkeit zu ändern.

Der RFM-Score berechnet sich wie folgt: (Recency-Score x Recency-Gewicht) + (Frequency-Score x Frequency-Gewicht) + (Monetary-Score x Monetary-Gewicht).

Bindungen. Gibt an, wie identische (gebundene) Scores klassiert werden sollen. Folgende Optionen stehen zur Auswahl:

- **Zu nächstem hinzu.** Wählen Sie diese Option aus, um die Bindungswerte nach oben zur nächsten Klasse zu verschieben.
- **In aktuellem beibehalten.** Hiermit werden die Bindungswerte in der aktuellen (niedrigeren) Klasse belassen. Bei dieser Methode werden insgesamt ggf. weniger Klassen erstellt. (Dies ist der Standardwert.)

Klassenschwellenwerte. Dient zur Angabe, ob RFM-Scores und Klassenzuweisungen bei jeder Ausführung des Knotens neu berechnet werden sollen oder nur nach Bedarf (z. B. wenn neue Daten hinzugefügt wurden). Bei Auswahl von Von Registerkarte “Klassenwerte” lesen, sofern verfügbar können Sie die oberen und unteren Trennwerte für die verschiedenen Klassen auf der Registerkarte “Klassenwerte” ändern.

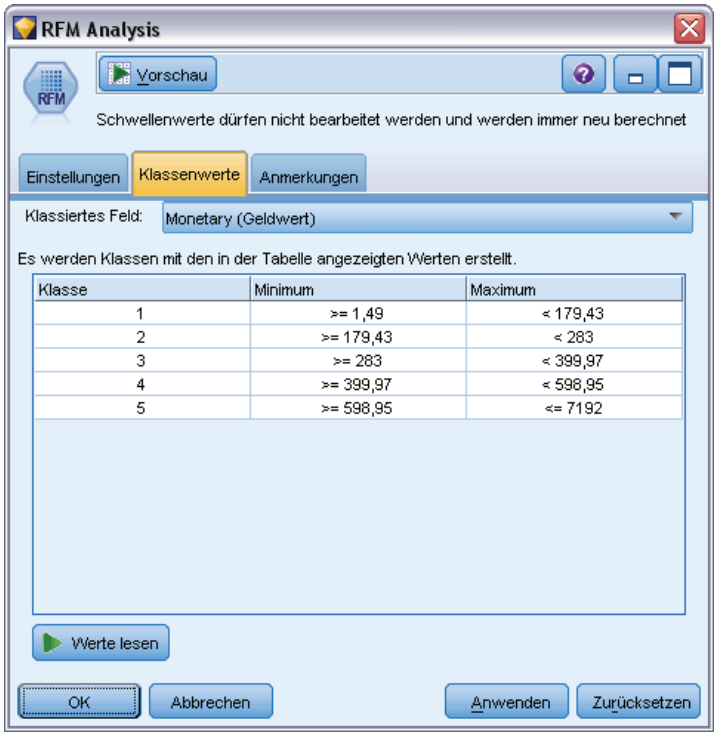
Bei Ausführung klassiert der Knoten RFM-Analyse die Felder mit den Rohwerten für Aktualität, Häufigkeit und Geldwert und fügt folgende neue Felder zum Daten-Set hinzu:

- Aktualitäts-Score. Ein Rang (Klassenwert) für “Recency (Aktualität)”
- Häufigkeits-Score. Ein Rang (Klassenwert) für “Frequency (Häufigkeit)”
- Geldwert-Score. Ein Rang (Klassenwert) für “Monetary (Geldwert)”
- RFM-Score. Die gewichtete Summe des Recency-, Frequency- und Monetary-Score.

Ausreißer in Endklassen aufnehmen. Bei Auswahl dieses Kontrollkästchens werden Datensätze, die unterhalb der untersten Klasse liegen, zur untersten Klasse hinzugefügt und Datensätze oberhalb der höchsten Klasse werden in die höchste Klasse aufgenommen; andernfalls erhalten sie einen Nullwert. Dieses Feld steht nur bei Auswahl von Von Registerkarte “Klassenwerte” lesen, sofern verfügbar zur Verfügung.

Knoten "RFM-Analyse" – Klassierung

Abbildung 4-73
Festlegen der Klassenwerte für die RFM-Analyse



Auf der Registerkarte "Klassenwerte" können Sie die Schwellenwerte für die generierten Klassen anzeigen und in bestimmten Fällen bearbeiten.

Hinweis: Die Werte auf dieser Registerkarte können nur bearbeitet werden, wenn auf der Registerkarte "Einstellungen" die Option Von Registerkarte "Klassenwerte" lesen, sofern verfügbar ausgewählt wurde.

Klassiertes Feld. Mithilfe der Dropdown-Liste können Sie ein Feld für die Aufteilung in Klassen auswählen. Verfügbar sind die auf der Registerkarte "Einstellungen" ausgewählten Werte.

Tabelle der Klassenwerte. Hier werden die Schwellenwerte für jeden generierten Bin angezeigt. Bei Auswahl von Von Registerkarte "Klassenwerte" lesen, sofern verfügbar auf der Registerkarte "Einstellungen" können Sie die oberen und unteren Trennwerte für die verschiedenen Klassen ändern, indem Sie auf die entsprechende Zelle doppelklicken.

Werte lesen. Liest klassierte Werte aus dem Daten-Set ein und füllt die Tabelle der Klassenwerte aus. Beachten Sie: Bei Auswahl von Immer neu berechnen auf der Registerkarte "Einstellungen" werden die Schwellenwerte der Klassen überschrieben, wenn neue Daten den Stream durchlaufen.

Partitionsknoten

Partitionsknoten werden zur Generierung eines Partitionsfelds verwendet, das Daten in getrennte Untergruppen bzw. Stichproben für die Trainings, Test- und Validierungsphase der Modellerstellung aufteilt. Indem Sie mit einer Stichprobe das Modell erstellen und es mit einer separaten Stichprobe testen, erhalten Sie einen guten Hinweis dafür, wie gut das Modell sich für größere Datenmengen generalisieren lässt, die den aktuellen Daten ähneln.

Der Partitionsknoten generiert ein nominales Feld, dessen Rolle auf Partition eingestellt ist. Wenn ein geeignetes Feld bereits in Ihren Daten vorhanden ist, kann dieses alternativ mithilfe eines Typknotens als Partition gekennzeichnet werden. In diesem Fall ist kein gesonderter Partitionsknoten erforderlich. Jedes instanziierte nominale Feld mit zwei oder drei Werten kann verwendet werden, nicht jedoch Flag-Felder. Für weitere Informationen siehe Thema Festlegen der Feldrolle auf S. 150.

In einem Stream können mehrere Partitionsfelder definiert werden. Wenn dies geschieht, muss allerdings bei jedem Modellierungsknoten, der Partitionierung verwendet, ein einzelnes Partitionsfeld ausgewählt werden. (Wenn nur eine einzige Partition vorhanden ist, wird diese immer automatisch verwendet, wenn die Partitionierung aktiviert ist.)

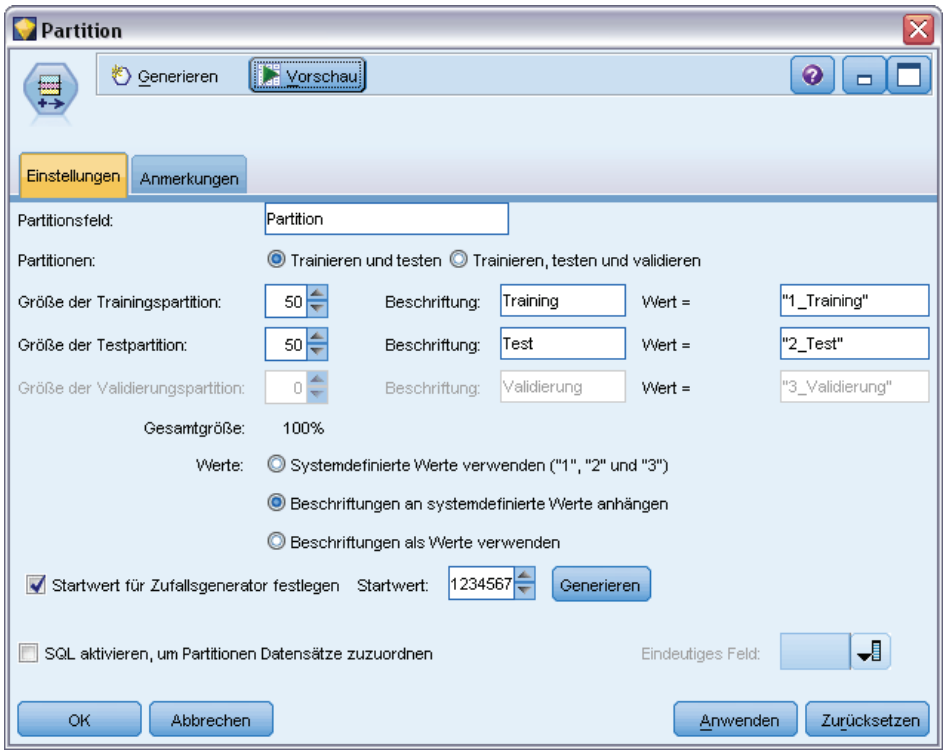
Aktivieren der Partitionierung. Um die Partition in einer Analyse zu verwenden, muss auf der Registerkarte “Modelloptionen” des entsprechenden Modellerstellungs- oder Analyseknotens die Partitionierung aktiviert sein. Wenn die Auswahl der Option aufgehoben ist, kann die Partitionierung deaktiviert werden, ohne das Feld zu entfernen.

Um ein Partitionsfeld auf der Grundlage eines anderen Kriteriums, wie beispielsweise Datumsbereich oder Standort, zu erstellen, können Sie auch einen Ableitungsknoten verwenden. Für weitere Informationen siehe Thema Ableitungsknoten auf S. 167.

Beispiel. Bei der Erstellung eines RFM-Streams zur Ermittlung aktueller Kunden, die positiv auf frühere Marketingkampagnen reagiert haben, verwendet die Marketingabteilung einer Vertriebsgesellschaft einen Partitionsknoten, um die Daten in Trainings- und Test-Partitionen zu unterteilen.

Partitionsknotenoptionen

Abbildung 4-74
Dialogfeld "Partitionsknoten", Registerkarte "Einstellungen"



Partitionsfeld. Gibt den Namen des vom Knoten erstellten Felds an.

Partitionen. Sie können die Daten in zwei Stichproben (Trainieren und testen) oder drei Stichproben (Trainieren, testen und validieren) partitionieren.

- **Trainieren und testen.** Partitioniert die Daten in zwei Stichproben, sodass Sie das Modell mit einer Stichprobe trainieren und mit der zweiten Stichprobe testen können.
- **Trainieren, testen und validieren.** Partitioniert die Daten in drei Stichproben, sodass Sie das Modell mit einer Stichprobe trainieren und mit der zweiten Stichprobe testen und verfeinern können und schließlich die Ergebnisse mit der dritten Stichprobe validieren können. Dadurch wird allerdings die Größe der einzelnen Partitionen entsprechend verringert. Außerdem ist dieses Verfahren wohl für sehr große Daten-Sets am besten geeignet.

Größe der Partition. Gibt die relative Größe der einzelnen Partitionen an. Wenn die Summe der Partitionsgrößen weniger als 100 % beträgt, werden die Datensätze, die nicht in einer Partition enthalten sind, verworfen. Beispiel: Ein Benutzer hat 10 Millionen Datensätze und eine Partitionsgröße von 5 % für das Training und von 10 % für das Testen angegeben. Nach der Ausführung des Knotens sollten ca. 500.000 Trainings- und ca. 1 Million Testdatensätze vorhanden sein. Die restlichen Datensätze müssten verworfen worden sein.

Werte. Gibt die Werte an, die für die einzelnen Partitionsstichproben in den Daten verwendet werden.

- **Systemdefinierte Werte verwenden ("1", "2" und "3")** Verwendet eine ganze Zahl für jede Partition. Beispiel: Alle Datensätze, die in der Training-Stichprobe enthalten sind, weisen den Wert 1 für das Partitionsfeld auf. Dadurch wird sichergestellt, dass die Daten zwischen verschiedenen Ländereinstellungen übertragbar sind und dass bei einer Reinstanziierung des Partitionsfelds an einer anderen Stelle (beispielsweise beim erneuten Einlesen der Daten aus einer Datenbank) die Sortierreihenfolge beibehalten wird (sodass 1 noch immer für die Trainingspartition steht). Die Werte bedürfen jedoch einiger Interpretation.
- **Beschriftungen an systemdefinierte Werte anhängen.** Kombiniert die ganze Zahl mit einer Beschriftung. Beispiel: Trainingspartitions-Datensätze, die den Wert *1_Training* aufweisen. Dadurch kann leicht erkannt werden, wozu die einzelnen Werte gehören, und gleichzeitig wird die Sortierreihenfolge beibehalten. Die Werte sind jedoch für eine bestimmte Ländereinstellung spezifisch.
- **Beschriftungen als Werte verwenden.** Verwendet die Beschriftung ohne ganze Zahl, beispielsweise *Training*. Dadurch können die Werte durch Bearbeitung der Beschriftungen angegeben werden. Dadurch werden die Daten jedoch von der Ländereinstellung abhängig und bei der erneuten Instanziierung einer Partitionsspalte werden die Werte in die natürliche Sortierreihenfolge gebracht, die nicht unbedingt mit ihrer "semantischen" Reihenfolge übereinstimmen muss.

Startwert für Zufallsgenerator festlegen. Bei der Stichprobenziehung oder Partitionierung von Datensätzen auf der Grundlage eines Zufallsprozentsatzes können Sie mit dieser Option dieselben Ergebnisse in einer anderen Sitzung replizieren. Wenn Sie den vom Zufallszahlengenerator verwendeten Startwert angeben, stellen Sie sicher, dass bei jeder Ausführung des Knotens dieselben Datensätze zugewiesen werden. Geben Sie den gewünschten Startwert ein oder klicken Sie auf die Schaltfläche Generieren, um automatisch einen Startwert zu generieren. Wenn diese Option nicht ausgewählt ist, wird bei jeder Ausführung des Knotens eine andere Stichprobe generiert.

Hinweis: Bei Verwendung der Option Startwert für Zufallsgenerator festlegen mit Datensätzen, die aus einer Datenbank eingelesen wurden, ist möglicherweise vor der Stichprobenziehung ein Sortierknoten erforderlich, um zu gewährleisten, dass bei jeder Ausführung des Knotens dasselbe Ergebnis erzielt wird. Dies liegt daran, dass der Startwert für den Zufallsgenerator von der Reihenfolge der Datensätze abhängt, die in relationalen Datenbanken nicht unbedingt gleich bleibt. Für weitere Informationen siehe Thema Sortierknoten in Kapitel 3 auf S. 88.

SQL aktivieren für die Zuweisung von Datensätzen zu Partitionen (nur für Tier-1-Datenbanken) Aktivieren Sie dieses Kontrollkästchen, um SQL-Pushback für die Zuweisung von Datensätzen zu Partitionen zu verwenden. Wählen Sie in der Dropdown-Liste Eindeutiges Feld ein Feld mit eindeutigen Werten (z. B. ein ID-Feld), um sicherzustellen, dass Datensätze zufällig aber auf wiederholbare Weise zugewiesen werden.

Datenbankstufen werden in der Beschreibung des Quellenknotens "Datenbank" erläutert. Für weitere Informationen siehe Thema Datenbankquellenknoten in Kapitel 2 auf S. 15.

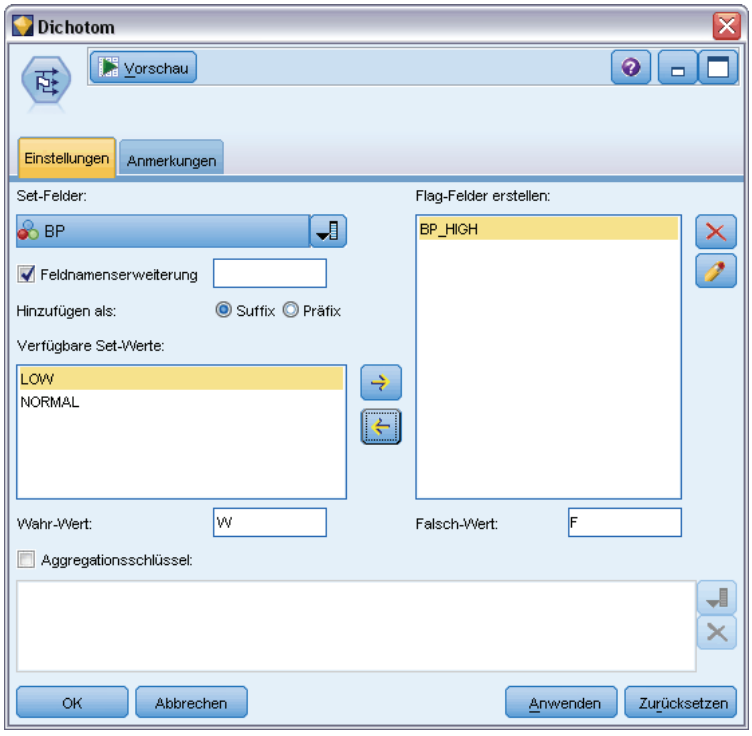
Generieren von Auswahlknoten

Mithilfe des Menüs “Generieren” im Partitionsknoten können Sie automatisch einen Auswahlknoten für jede Partition erstellen. Beispielsweise können Sie alle Datensätze in der Trainingspartition auswählen, um unter Verwendung nur dieser Partition eine weitere Evaluation bzw. weitere Analysen zu erstellen.

Dichotomknoten

Der Dichotomknoten wird zur Ableitung von Flag-Feldern auf der Grundlage der Kategoriewerte verwendet, die für ein oder mehrere nominale Felder definiert sind. Ihr Daten-Set könnte beispielsweise das nominale Feld *Blutdruck* mit den Werten *Hoch*, *Normal* und *Niedrig* enthalten. Zur Erleichterung der Datenbearbeitung können Sie ein Flag-Feld für hohen Blutdruck erstellen, das angibt, ob der Patient unter hohem Blutdruck leidet oder nicht.

Abbildung 4-75
Erstellen eines Flag-Felds für hohen Blutdruck



Festlegen der Optionen für den Dichotomknoten

Set-Felder. Listet alle Datenfelder mit einem Messniveau von *Nominal* (Set) auf. Wählen Sie eines aus der Liste aus, um die Werte im Set anzuzeigen. Sie können eine Auswahl aus diesen Werten treffen, um ein Flag-Feld zu erstellen. Dabei müssen die Daten mithilfe eines aufwärts liegenden Quellen- oder Typknotens vollständig instanziiert werden, bevor die verfügbaren

nominalen Felder (und die zugehörigen Werte) sichtbar werden. Für weitere Informationen siehe Thema Typknoten auf S. 136.

Feldnamenerweiterung. Bei Auswahl dieser Option werden Steuerelemente zur Angabe einer Erweiterung aktiviert, die dem neuen Flag-Feld als Suffix oder Präfix hinzugefügt werden kann. Standardmäßig werden neue Feldnamen automatisch erstellt, indem der ursprüngliche Feldname mit dem Feldwert zu einer Beschriftung kombiniert wird, wie beispielsweise *Feldname_Feldwert*.

Verfügbare Set-Werte. Die Werte im oben ausgewählten Set werden hier angezeigt. Wählen Sie einen oder mehrere Werte aus, für die Flags generiert werden sollen. Wenn die Werte in einem Feld namens *Blutdruck* beispielsweise *Hoch*, *Mittel* und *Niedrig* lauten, dann können Sie *Hoch* auswählen und zu der Liste auf der rechten Seite hinzufügen. Dadurch wird ein Feld mit einem Flag für Datensätze erstellt, die einen Wert enthalten, der auf einen hohen Blutdruck hinweist.

Flag-Felder erstellen. Die neu erstellten Flag-Felder werden hier aufgeführt. Mit den Steuerelementen für die Feldnamenerweiterung können Sie Optionen zur Benennung des neuen Felds angeben.

Wahr-Wert. Dient zur Angabe des Wahr-Werts, den der Knoten zum Festlegen eines Flags verwendet. Standardmäßig lautet dieser Wert `T`.

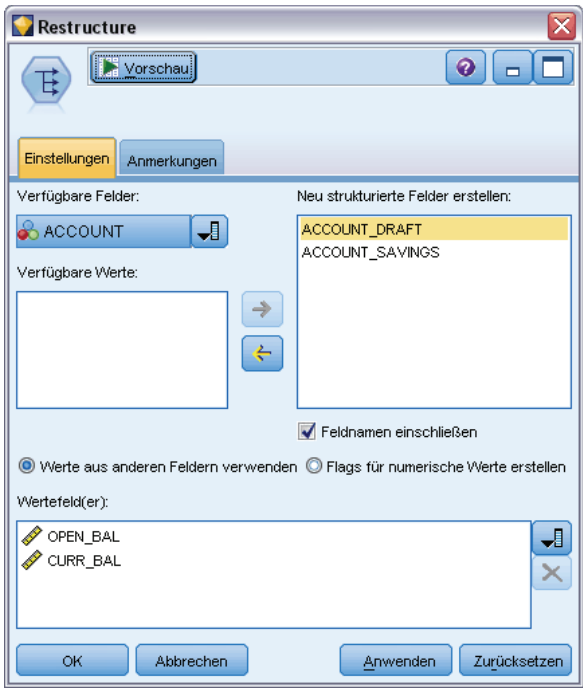
Falsch-Wert. Dient zur Angabe des Falsch-Werts, den der Knoten zum Festlegen eines Flags verwendet. Standardmäßig lautet dieser Wert `F`.

Aggregationsschlüssel. Mit dieser Option werden Gruppen anhand der unten angegebenen Schlüsselfelder zu Gruppen zusammengefasst. Bei Auswahl von Aggregationsschlüssel werden alle Flag-Felder in einer Gruppe aktiviert, wenn *ein beliebiger* Datensatz auf "wahr" gesetzt wurde. Mit der Feldauswahl-Schaltfläche können Sie angeben, welche Schlüsselfelder zur Aggregierung von Datensätzen verwendet werden.

Neustrukturierungsknoten

Mit dem Neustrukturierungsknoten erzeugen Sie mehrere Felder auf der Grundlage der Werte eines nominalen oder Flag-Felds. Diese neu erzeugten Felder können Werte aus einem anderen Feld enthalten oder auch ein numerisches Flag (0 oder 1). Dieser Knoten weist ähnliche Funktionen auf wie der Dichotomknoten. Er bietet jedoch größere Flexibilität. Hiermit können Sie Felder mit einem beliebigen Typ (auch numerische Flags) anhand der Werte aus einem anderen Feld anlegen. Anschließend können Sie die Aggregation oder andere Bearbeitungsschritte für andere abwärts liegende Knoten durchführen. (Mit dem Dichotomknoten können Sie Felder in einem einzigen Schritt aggregieren, was beim Erstellen von Flag-Feldern nützlich ist.)

Abbildung 4-76
Erzeugen neu strukturierter Felder für "Account"



Das folgende Daten-Set enthält beispielsweise ein nominales Feld *Account* mit den Werten *Savings* und *Draft*. Für jedes Konto werden der Anfangssaldo und der aktuelle Saldo festgehalten; einige Kunden besitzen mehrere Konten von jedem Typ. Angenommen, Sie möchten erfahren, ob ein Kunde ein Konto mit einem bestimmten Typ besitzt und, wenn ja, wie hoch der Saldo in jedem Kontentyp ist. Mit dem Neustrukturierungsknoten erzeugen Sie je ein Feld für die Werte für *Account* und Sie wählen den Wert *Current_Balance* aus. In jedes neue Feld wird der aktuelle Saldo für den jeweiligen Datensatz eingetragen.

Tabelle 4-2
Beispieldaten vor der Neustrukturierung

CustID	Account	Open_Bal	Current_Bal
12701	Text	1000	1005.32
12702	Sparkonto	100	144.51
12703	Sparkonto	300	321.20
12703	Sparkonto	150	204.51
12703	Text	1200	586.32

Tabelle 4-3
Beispieldaten nach der Neustrukturierung

CustID	Account	Open_Bal	Current_Bal	Account_Draft_ Current_Bal	Account_Savings _Current_Bal
12701	Text	1000	1005.32	1005.32	$null$
12702	Sparkonto	100	144.51	$null$	144.51
12703	Sparkonto	300	321.20	$null$	321.20

CustID	Account	Open_Bal	Current_Bal	Account_Draft_ Current_Bal	Account_Savings _Current_Bal
12703	Sparkonto	150	204.51	$null$	204.51
12703	Text	1200	586.32	586.32	$null$

Verwenden des Neustrukturierungsknotens mit dem Aggregatknoten

In vielen Fällen soll der Neustrukturierungsknoten mit einem Aggregatknoten gekoppelt werden. Im obigen Beispiel besitzt ein Kunde (mit der ID 12703) drei Konten. Mit einem Aggregatknoten können Sie den Gesamtsaldo für jeden Kontentyp berechnen. Das Schlüsselfeld ist *CustID* und die Aggregatfelder sind die soeben neu strukturierten Felder *Account_Draft_Current_Bal* und *Account_Savings_Current_Bal*. Die nachstehende Tabelle zeigt die Ergebnisse.

Tabelle 4-4
Beispieldaten nach der Neustrukturierung und Aggregation

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

Festlegen von Optionen für den Umstrukturierungsknoten

Verfügbare Felder. Listet alle Datenfelder mit einem Messniveau von *Nominal* (Set) oder *Flag* auf. Wählen Sie ein Feld in der Liste aus, sodass die Werte im Set oder im Flag angezeigt werden, und wählen Sie die gewünschten Werte für die Erstellung der neu strukturierten Felder aus. Dabei müssen die Daten mithilfe eines aufwärts liegenden Quellen- oder Typknotens vollständig instanziiert werden, bevor die verfügbaren Felder (und die zugehörigen Werte) sichtbar werden. Für weitere Informationen siehe Thema Typknoten auf S. 136.

Verfügbare Werte. Die Werte im oben ausgewählten Set werden hier angezeigt. Wählen Sie einen oder mehrere Werte aus, für die neu strukturierte Felder generiert werden sollen. Wenn die Werte im Feld *Blutdruck* beispielsweise *Hoch*, *Mittel* und *Niedrig* lauten, dann können Sie *Hoch* auswählen und zu der Liste auf der rechten Seite hinzufügen. Dadurch wird ein Feld mit einem bestimmten Wert (siehe unten) für Datensätze erstellt, die den Wert *Hoch* enthalten.

Neu strukturierte Felder erstellen. Hier werden die soeben erstellten, neu strukturierten Felder aufgeführt. Standardmäßig werden neue Feldnamen automatisch erstellt, indem der ursprüngliche Feldname mit dem Feldwert zu einer Beschriftung kombiniert wird, wie beispielsweise *Feldname_Feldwert*.

Feldnamen einschließen. Deaktivieren Sie diese Option, wenn der ursprüngliche Feldname den neuen Feldnamen nicht als Präfix vorangestellt werden soll.

Werte aus anderen Feldern verwenden. Geben Sie mindestens ein Feld an, dessen Werte in die neu strukturierten Felder eingetragen werden sollen. Mit der Feldauswahl können Sie das oder die gewünschten Felder bestimmen. Für jedes angegebene Feld wird ein neues Feld erstellt. Der Name des Feldes, aus dem die Werte stammen, wird an den Namen des umstrukturierten

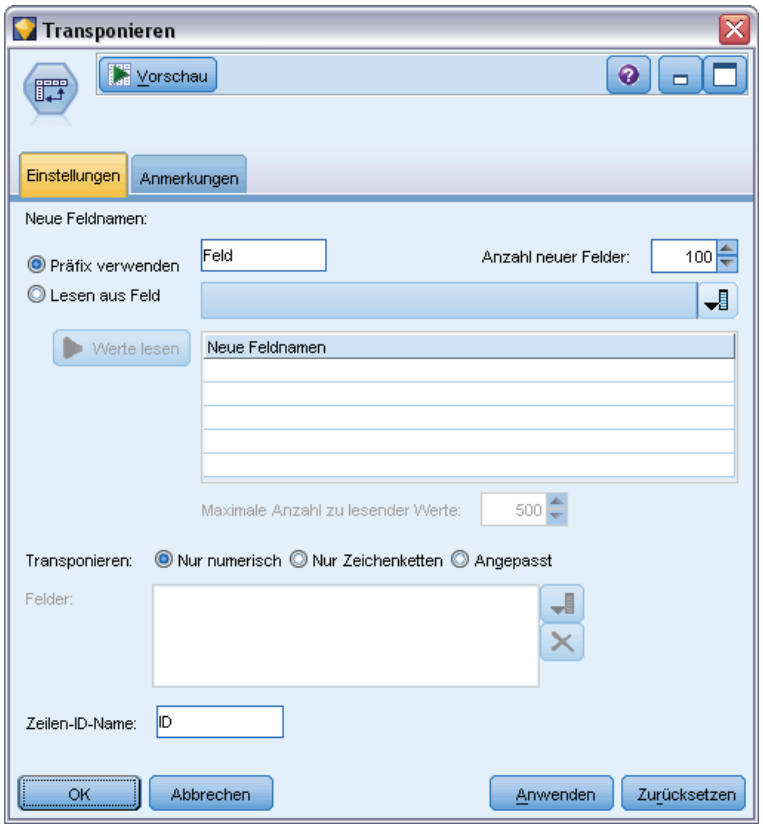
Feldes angehängt, z. B. *Blutdruck_Hohes_Alter* oder *Blutdruck_Niedriges_Alter*. Jedes neue Feld übernimmt den Typ des Felds, aus dem der ursprüngliche Wert stammt.

Flags für numerische Werte erstellen. Wenn Sie diese Option auswählen, wird kein Wert aus einem anderen Feld bernommen, sondern die neuen Felder werden mit numerischen Wert-Flags (0 für Falsch, 1 für Wahr) gefüllt.

Transponierknoten

Standardmäßig bestehen die Spalten aus Feldern und die Zeilen aus Datensätzen oder Beobachtungen. Falls notwendig, können Sie mithilfe eines Transponierknotens die Daten in Zeilen und Spalten vertauschen, sodass aus Feldern Datensätze und aus Datensätzen Felder werden. Wenn Sie beispielsweise Zeitreihendaten verwenden, in der die Zeitreihen jeweils eine Zeile darstellen (also keine Spalte), können Sie die Daten vor der Analyse transponieren.

Abbildung 4-77
Registerkarte "Einstellungen" beim Transponierknoten



Festlegen von Optionen für Transponierknoten

Neue Feldnamen

Neue Felder können automatisch auf der Grundlage eines angegebenen Präfixes generiert oder aus einem bestehenden Feld in den Daten eingelesen werden.

Präfix verwenden. Mit dieser Option werden die neuen Feldnamen automatisch auf der Grundlage des angegebenen Präfixes erstellt (*Field1*, *Field2* usw.). Sie können das Präfix ganz nach Bedarf anpassen. Bei dieser Option muss die Anzahl der zu erstellenden Felder angegeben werden, unabhängig von der Anzahl der Zeilen in den ursprünglichen Daten. Wenn Sie unter Anzahl neuer Felder beispielsweise den Wert 100 festlegen, werden alle Daten ab der 101. Zeile verworfen. Enthalten die Originaldaten weniger als 100 Zeilen, bleiben einige Felder leer. (Sie können die Anzahl der Felder ganz nach Bedarf anpassen. Mit dieser Einstellung soll vermieden werden, dass z. B. eine Million Datensätze in eine Million Felder transponiert werden, was zu unüberschaubaren Ergebnissen führen würde.)

Beispiel: Angenommen, Ihnen liegen Daten mit Zeitreihen in den Zeilen und einem separaten Feld (Spalte) für jeden Monat vor. Sie können diese Daten transponieren, sodass jede Zeitreihe in einem separaten Feld (mit einer Zeile für jeden Monat) vorliegt.

Abbildung 4-78
Ursprüngliche Daten mit Zeitreihen in Zeilen

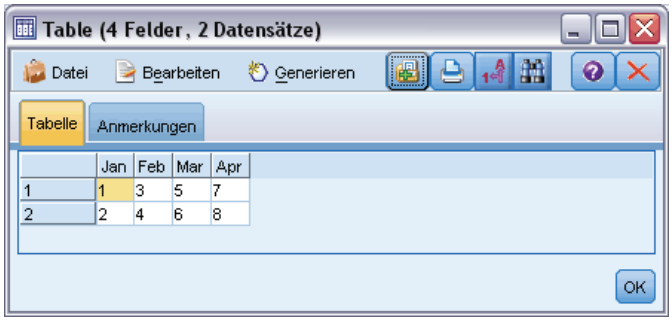
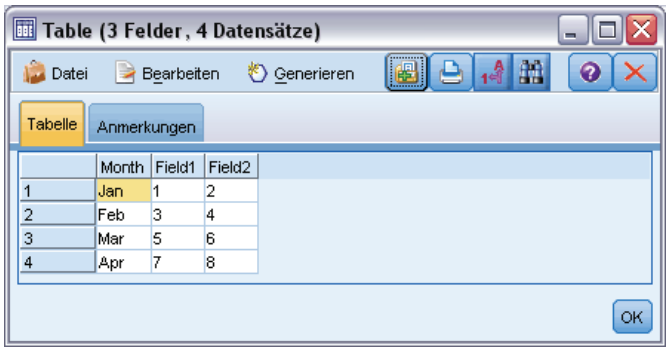


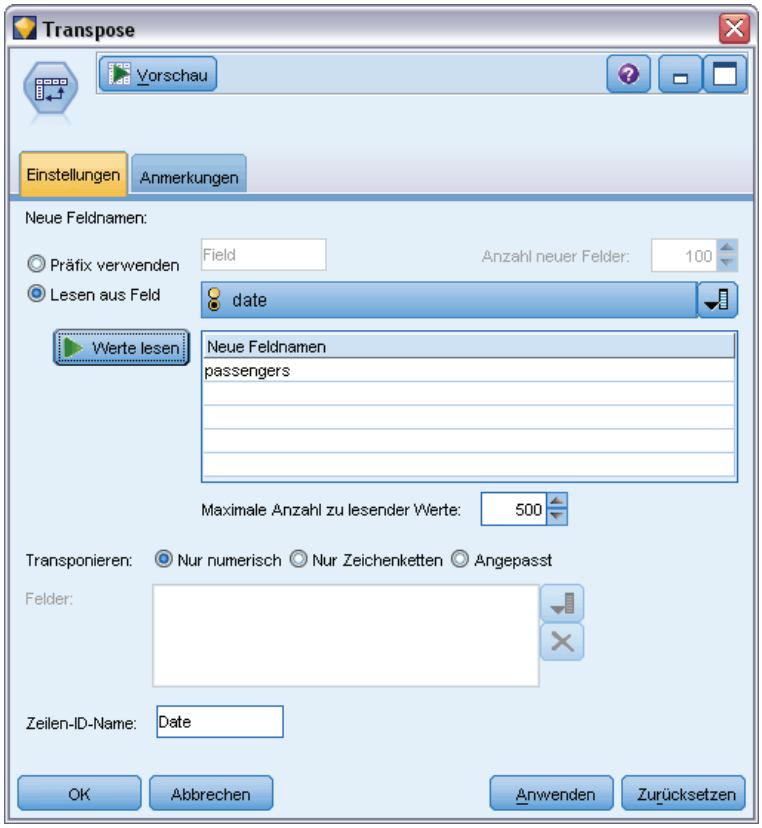
Abbildung 4-79
Transponierte Daten mit Zeitreihen in Spalten



Hinweis: Um die abgebildeten Ergebnisse zu erstellen, wurde die Option “Anzahl neuer Felder” von 100 in 2 geändert und der Zeilen-ID-Name wurde von ID in Monat geändert (siehe unten).

Lesen aus Feld. Liest die Feldnamen aus einem vorhandenen Feld. Bei dieser Option wird die Anzahl der neuen Felder durch die Daten bestimmt (bis zum angegebenen Höchstwert). Jeder Wert im ausgewählten Feld wird zu einem neuen Feld in den Ausgabedaten. Das ausgewählte Feld kann einen beliebigen Speichertyp besitzen (ganze Zahl, Zeichenkette, Datum usw.); um doppelte Feldnamen zu vermeiden, müssen die Werte im ausgewählten Feld jedoch eindeutig sein. (Die Anzahl der Werte soll also mit der Anzahl der Zeilen übereinstimmen.) Falls doppelte Feldnamen auftreten, wird eine Warnmeldung angezeigt.

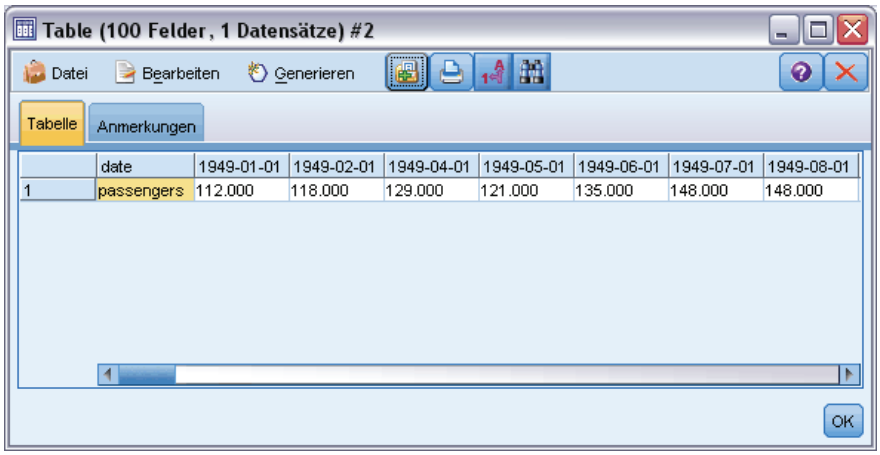
Abbildung 4-80
Einlesen der Feldnamen aus einem vorhandenen Feld.



- **Werte lesen.** Falls das ausgewählte Feld nicht instanziiert wurde, wählen Sie diese Option, damit die Liste der neuen Feldnamen gefüllt wird. Wurde das Feld bereits instanziiert, ist dieser Schritt nicht notwendig.
- **Maximale Anzahl zu lesender Werte.** Beim Einlesen von Feldnamen aus den Daten wird eine Obergrenze angegeben, um das Erstellen einer übermäßig großen Anzahl von Feldern zu verhindern. (Wie bereits beschrieben, würde das Transponieren von einer Million Datensätzen in eine Million Felder zu unüberschaubaren Ergebnissen führen.)

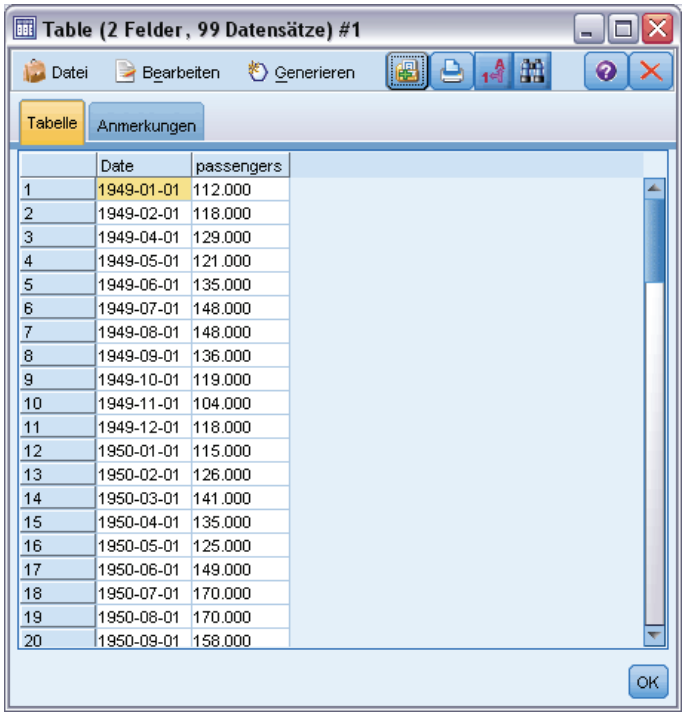
Wenn beispielsweise in der ersten Spalte der Daten der Name für die einzelnen Zeitreihen angegeben wird, können diese Werte in den transponierten Daten als Feldnamen verwendet werden.

Abbildung 4-81
Ursprüngliche Daten mit Zeitreihen in einer einzelnen Zeile



	date	1949-01-01	1949-02-01	1949-04-01	1949-05-01	1949-06-01	1949-07-01	1949-08-01
1	passengers	112.000	118.000	129.000	121.000	135.000	148.000	148.000

Abbildung 4-82
Transponierte Daten mit Zeitreihen in Spalten



	Date	passengers
1	1949-01-01	112.000
2	1949-02-01	118.000
3	1949-04-01	129.000
4	1949-05-01	121.000
5	1949-06-01	135.000
6	1949-07-01	148.000
7	1949-08-01	148.000
8	1949-09-01	136.000
9	1949-10-01	119.000
10	1949-11-01	104.000
11	1949-12-01	118.000
12	1950-01-01	115.000
13	1950-02-01	126.000
14	1950-03-01	141.000
15	1950-04-01	135.000
16	1950-05-01	125.000
17	1950-06-01	149.000
18	1950-07-01	170.000
19	1950-08-01	170.000
20	1950-09-01	158.000

Transponieren. Standardmäßig werden nur stetige Felder (numerischer Bereich) transponiert (ganze Zahl oder reelle Zahl als Speichertyp). Optional können Sie stattdessen eine Teilmenge numerischer Felder auswählen oder auch Zeichenkettenfelder transponieren. Alle transponierten Felder müssen allerdings denselben Speichertyp aufweisen (entweder numerisch oder Zeichenkette, nicht jedoch beide Typen), weil durch das Mischen der Eingabefelder gemischte Werte in den Ausgabespalten entstünden, was die Regel verletzt, dass alle Werte eines Feldes denselben Speichertyp besitzen müssen. Andere Speichertypen (Datum, Zeit, Zeitstempel) können nicht transponiert werden.

- **Nur numerisch.** Transponiert alle numerischen Felder (ganze Zahl oder reelle Zahl als Speichertyp). Die Anzahl der Zeilen in der Ausgabe entspricht der Anzahl der numerischen Felder in den ursprünglichen Daten.
- **Nur Zeichenketten.** Transponiert alle Zeichenkettenfelder.
- **Benutzerdefiniert.** Ermöglicht die Auswahl einer Teilmenge von numerischen Feldern. Die Anzahl der Zeilen in der Ausgabe entspricht der Anzahl der ausgewählten Felder. *Hinweis*: Diese Option ist nur für numerische Felder verfügbar.

Zeilen-ID-Name. Gibt den Namen des vom Knoten erstellten Zeilen-ID-Felds an. Die Werte für dieses Feld ergeben sich aus den Namen der Felder in den ursprünglichen Daten.

Tipp: Wenn beim Transponieren von Zeitreihendaten von Zeilen in Spalten die ursprünglichen Daten eine Zeile (z. B. Datum, Monat, Jahr) enthalten, die die Zeitperiode für die einzelnen Messungen angibt, müssen Sie sicherstellen, dass diese Beschriftungen als Feldnamen in IBM® SPSS® Modeler eingelesen werden (wie in den oben stehenden Beispielen beschrieben, bei denen Monat bzw. Datum als Feldnamen in den ursprünglichen Daten angezeigt werden) und nicht etwa in die erste Datenzeile aufgenommen werden. Dadurch wird eine Vermischung von Beschriftungen und Werten in den einzelnen Spalten vermieden (die dazu führen würde, dass Zahlen als Zeichenketten gelesen würden, da innerhalb einer Spalte nicht verschiedene Speichertypen vorkommen dürfen).

Zeitintervallknoten

Mit dem Zeitintervallknoten können Sie Intervalle angeben und Beschriftungen für Zeitreihendaten generieren, die in einer Zeitreihenmodellierung oder einem Zeitplotknoten zu Schätz- oder Vorhersagezwecken verwendet werden sollen. Die unterstützten Zeitintervalle reichen dabei von Sekunden bis hin zu Jahren. Wenn Sie beispielsweise eine Zeitreihe mit täglichen Messungen ausführen, die am 3. Januar 2005 begann, können Sie die Datensätze ab diesem Datum beschriften. Die zweite Zeile stünde dann für den 4. Januar usw. Darüber hinaus können Sie die Periodizität bestimmen, z. B. fünf Tage pro Woche oder acht Stunden pro Tag.

Des Weiteren können Sie den Bereich der Datensätze angeben, der für die Schätzung verwendet werden soll. Sie können auswählen, ob die frühesten Datensätze in der Zeitreihe ausgeschlossen werden sollen und ob Holdouts angegeben werden sollen. Dadurch können Sie das Modell testen, indem Sie die aktuellsten Datensätze in den Zeitreihendaten zurückhalten, um ihre bekannten Werte mit den geschätzten Werten für die betreffenden Zeitperioden zu vergleichen.

Außerdem können Sie angeben, für wie viele Zeitperioden in die Zukunft die Prognose erstellt werden soll, und Sie können zukünftige Werte angeben, die für die Vorhersage durch weiter unten im Stream liegende Zeitreihenmodellierungsknoten verwendet werden sollen.

Im Zeitintervallknoten wird ein *TimeLabel*-Feld in einem geeigneten Format für das angegebene Intervall und die Periode generiert, außerdem ein *TimeIndex*-Feld, mit dem jedem Datensatz eine eindeutige ganze Zahl zugewiesen wird. Auch einige zusätzliche Felder werden ggf. erzeugt, abhängig vom ausgewählten Intervall bzw. der Periodizität (z. B. die Minute oder Sekunde, in die eine Messung fällt).

Sie können die Werte auffüllen oder aggregieren, um so sicherzustellen, dass die Messungen in gleichmäßigen Zeitabständen erfolgen. Bei den Methoden zur Modellierung von Zeitreihendaten ist ein einheitliches Intervall zwischen den Messungen erforderlich; fehlende Werte werden durch

leere Zeilen dargestellt. Falls Ihre Daten diese Bedingung nicht bereits erfüllen, können Sie sie mithilfe dieses Knotens entsprechend transponieren.

Kommentare

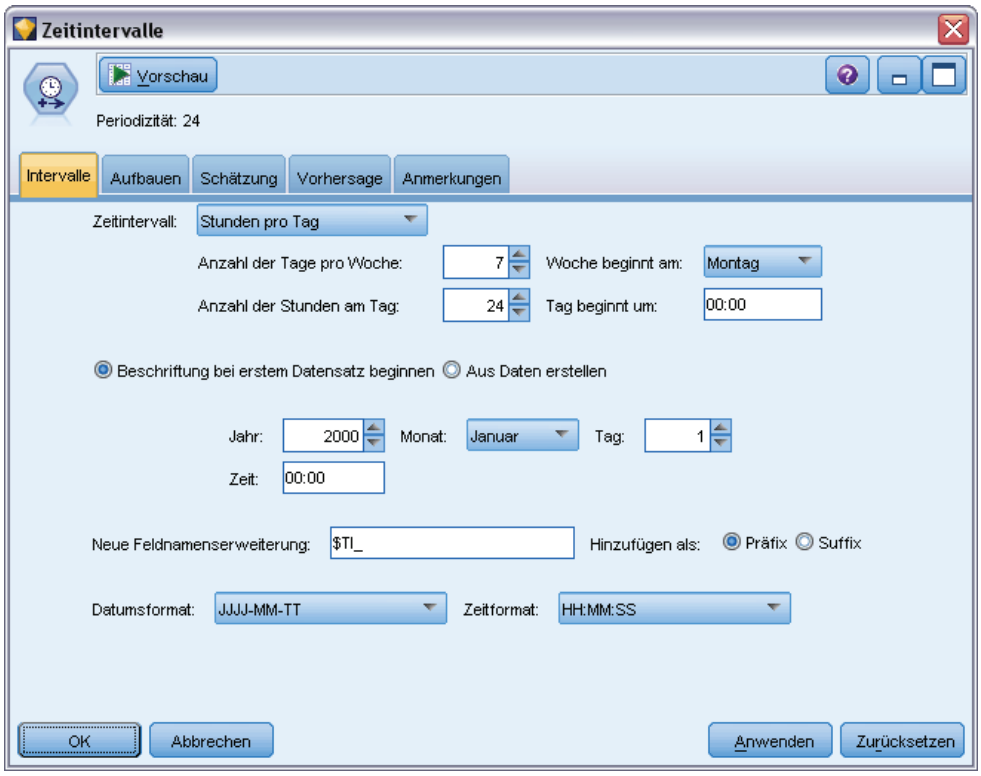
- Die periodischen Intervalle stimmen unter Umständen nicht mit der Echtzeit überein. Bei einer Reihe, die auf einer normalen Fünf-Tage-Arbeitswoche beruht, würde die Lücke zwischen Freitag und Montag als ein einziger Tag behandelt.
- Im Zeitintervallknoten wird vorausgesetzt, dass sich jede Zeitreihe in einem Feld oder einer Spalte befindet und dabei jeweils eine Zeile für jede Messung vorliegt. Falls notwendig, können Sie die Daten so transponieren, dass diese Anforderung erfüllt ist. Für weitere Informationen siehe Thema Transponierknoten auf S. 215.
- Bei Zeitreihen mit ungleichmäßigen Abständen können Sie ein Feld definieren, aus dem das Datum oder die Uhrzeit für die jeweilige Messung hervorgeht. Beachten Sie, dass hierfür ein Datums-, Zeit- oder Zeitstempelfeld im entsprechenden Format zur Verwendung als Eingabe erforderlich ist. Falls erforderlich, können Sie ein bestehendes Feld (beispielsweise ein Beschriftungsfeld vom Typ "Zeichenkette") mithilfe eines Füllerknotens in dieses Format konvertieren. Für weitere Informationen siehe Thema Speichertypkonvertierung mithilfe des Füllerknotens auf S. 181.
- Beim Betrachten der Details für die erzeugten Beschriftungs- und Indexfelder hilft es häufig, wenn Sie die Anzeige der Wertelabels aktivieren. Wenn Sie beispielsweise eine Tabelle mit Werten betrachten, die für monatliche Daten erzeugt wurde, können Sie mit dem Symbol für die Wertelabels in der Symbolleiste die Beschriftungen *Januar, Februar, März* usw. einblenden statt *1, 2, 3* usw.

Abbildung 4-83
Symbol für Wertelabels

Festlegen von Zeitintervallen

Auf der Registerkarte "Intervalle" bestimmen Sie das Intervall und die Periodizität zum Aufbauen oder Beschriften der Zeitreihe. Die jeweiligen Einstellungen sind abhängig vom ausgewählten Intervall. Bei der Option Stunden pro Tag können Sie beispielsweise die Anzahl der Tage in der Woche festlegen, außerdem die Anzahl der Stunden pro Tag sowie die Stunde, zu der der Tag beginnt. Für weitere Informationen siehe Thema Unterstützte Intervalle auf S. 228.

Abbildung 4-84
Zeitintervalleinstellungen für eine Zeitreihe im Stundenabstand



Beschriften oder Aufbauen der Zeitreihe

Sie können die Datensätze nacheinander beschriften oder auch die Zeitreihe auf der Grundlage eines angegebenen Datums-, Zeitstempel- oder Zeitfelds erstellen.

- **Beschriftung bei erstem Datensatz beginnen.** Legen Sie das Anfangsdatum und/oder die Anfangsuhrzeit fest, mit der die aufeinander folgenden Datensätze beschriftet werden sollen. Bei einer Beschriftung pro Tag geben Sie beispielsweise das Datum und die Stunde an, zu der die Zeitreihe beginnt. Für jede nachfolgende Stunde wird dann ein neuer Datensatz angelegt. Bei dieser Methode werden lediglich die Beschriftungen hinzugefügt; ansonsten bleiben die ursprünglichen Daten unverändert. Es wird stattdessen angenommen, dass die Datensätze bereits gleiche Abstände zueinander aufweisen, dass also gleiche Intervalle zwischen den Messungen vorliegen. Fehlende Messwerte müssen durch leere Zeilen in den Daten ersetzt werden.
- **Aus Daten erstellen.** Bei Zeitreihen mit ungleichmäßigen Abständen können Sie ein Feld definieren, aus dem das Datum oder die Uhrzeit für die Messungen hervorgeht. Beachten Sie, dass hierfür ein Datums-, Zeit- oder Zeitstempelfeld im entsprechenden Format zur Verwendung als Eingabe erforderlich ist. Wenn Ihnen beispielsweise ein Zeichenkettenfeld mit Werten wie *Jan 2000*, *Feb 2000* usw. vorliegt, können Sie dieses mithilfe eines Füllerknotens in ein Datumsfeld konvertieren. Für weitere Informationen siehe Thema Speichertypkonvertierung mithilfe des Füllerknotens auf S. 181. Mit der Option Aus Daten erstellen werden die Daten ebenfalls gemäß dem angegebenen Intervall transformiert,

indem Datensätze je nach Bedarf aufgefüllt oder aggregiert werden, beispielsweise durch Zusammenfassen von Wochen zu Monaten oder durch Ersetzen fehlender Datensätze durch Leerwerte oder extrapolierte Werte. Auf der Registerkarte “Aufbauen” legen Sie die Funktionen fest, mit denen die Datensätze aufgefüllt oder aggregiert werden. Für weitere Informationen siehe Thema Aufbauoptionen für Zeitintervalle auf S. 222.

Neue Feldnamenerweiterung. Hiermit legen Sie ein Präfix oder ein Suffix fest, das für alle durch den Knoten erzeugten Felder übernommen wird. Wenn Sie beispielsweise das Standardpräfix *$TI_* verwenden, erhalten die durch den Knoten erzeugten Felder die Bezeichnung *$TI_TimeIndex*, *$TI_TimeLabel* usw.

Datumsformat. Bestimmt das Format für das durch den Knoten erstellte *TimeLabel*-Feld gemäß dem aktuellen Intervall. Die Verfügbarkeit dieser Optionen ist abhängig von der aktuellen Auswahl.

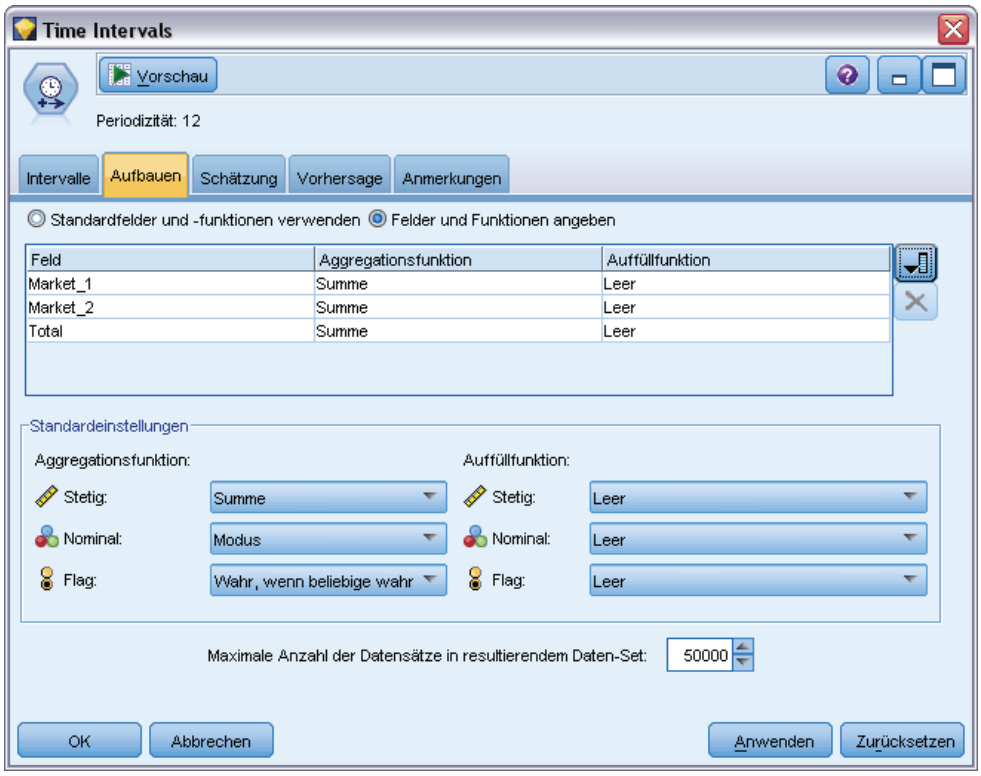
Zeitformat. Bestimmt das Format für das durch den Knoten erstellte *TimeLabel*-Feld gemäß dem aktuellen Intervall. Die Verfügbarkeit dieser Optionen ist abhängig von der aktuellen Auswahl.

Aufbauoptionen für Zeitintervalle

Auf der Registerkarte “Aufbauen” im Zeitintervallknoten legen Sie Optionen fest, mit denen die Felder gemäß dem angegebenen Intervall aggregiert oder aufgefüllt werden. Diese Einstellungen gelten nur dann, wenn auf der Registerkarte “Intervalle” die Option Aus Daten erstellen aktiviert ist. Liegt beispielsweise eine Mischung aus Wochen- und Monatsdaten vor, können Sie die Wochenwerte so aggregieren (zusammenfassen), dass ein gleichmäßiges monatliches Intervall entsteht. Alternativ können Sie ein Wochenintervall festlegen und die Zeitreihe auffüllen, indem Sie Leerwerte für die fehlenden Wochen einfügen oder fehlende Werte mithilfe einer bestimmten Auffüllfunktion extrapolieren.

Beim Auffüllen und Aggregieren der Daten werden die vorhandenen Datums- oder Zeitstempelfelder effektiv durch die erzeugten *TimeLabel*- und *TimeIndex*-Felder überschrieben und aus der Ausgabe entfernt. Auch Felder ohne Typ werden herausgenommen. Felder, mit denen ein Zeitraum gemessen wird (z. B. ein Feld, das nicht den Zeitpunkt festhält, zu dem ein Service-Call begann, sondern die Dauer dieses Gesprächs), werden beibehalten, sofern sie intern als Zeitfelder statt als Zeitstempelfelder gespeichert sind. Für weitere Informationen siehe Thema Festlegen von Feldspeichertyp und Formatierung in Kapitel 2 auf S. 32. Die anderen Felder werden gemäß den Optionen auf der Registerkarte “Aufbauen” aggregiert.

Abbildung 4-85
Zeitintervallknoten: Registerkarte "Aufbauen"



- **Standardfelder und -funktionen verwenden.** Gibt an, dass alle Felder je nach Bedarf aggregiert oder aufgefüllt werden sollen, ausgenommen Datums- und Zeitstempelfelder sowie Felder ohne Typ, wie oben beschrieben. Die Standardfunktion wird gemäß dem Messniveau angewendet. Stetige Felder werden beispielsweise anhand des Mittelwerts aggregiert, nominale Felder dagegen anhand des Modus. Im unteren Teil des Dialogfelds können Sie die Standardeinstellungen für die Messniveaus ändern.
- **Felder und Funktionen angeben.** Mit dieser Option können Sie die aufzufüllenden oder zu aggregierenden Felder auswählen und auch die jeweilige Funktion festlegen. Alle nicht ausgewählten Felder werden aus der Ausgabe herausgenommen. Mit den Symbolen rechts können Sie Felder in die Tabelle aufnehmen oder daraus entfernen. Wenn Sie auf eine Zelle in einer bestimmten Spalte klicken, können Sie die Aggregations- oder Auffüllfunktion für dieses Feld ändern und somit die Standardeinstellung außer Kraft setzen. Felder ohne Typ werden aus der Liste ausgeschlossen und können nicht zur Tabelle hinzugefügt werden.

Standard. Bestimmt die Aggregations- und Auffüllfunktionen, die standardmäßig für die verschiedenen Feldtypen verwendet werden. Diese Standardeinstellungen werden angewendet, wenn Sie die Option Standards verwenden aktivieren, und gelten auch als anfängliche Standardeinstellungen für alle Felder, die neu in die Tabelle aufgenommen werden. (Wenn Sie die Standardeinstellungen ändern, wirkt sich dies nicht auf die Einstellungen in der Tabelle aus, sondern nur auf die nachfolgend hinzugefügten Felder.)

Aggregationsfunktionen. Die folgenden Aggregationsfunktionen stehen zur Verfügung:

- **Stetig.** Verfügbare Funktionen für stetige Felder: Mittelwert, Summe, Modus, Min und Max.

- **Nominal.** Verfügbare Optionen: Modus, Erste und Letzte. "Erste" bezieht sich auf den ersten Wert ungleich null in der (nach Datum sortierten) Aggregationsgruppe, "Letzte" entsprechend auf den letzten Wert ungleich null in dieser Gruppe.
- **Flag.** Verfügbare Optionen: Wahr, wenn beliebige wahr, Modus, Erste und Letzte.

Auffüllfunktion. Die folgenden Auffüllfunktionen stehen zur Verfügung:

- **Stetig.** Verfügbare Optionen: Leer und Mittelwert der zuletzt verwendeten Punkte; hierbei wird der Mittelwert der drei jüngsten Werte ungleich null vor der zu erstellenden Zeitperiode gebildet. Falls weniger als drei Werte vorliegen, ist der neue Wert leer. Zu den letzten Werten zählen nur "echte" Werte; zuvor erstellte aufgefüllte Werte werden bei der Suche nach Werten ungleich null nicht berücksichtigt.
- **Nominal.** Leer und Zuletzt verwendeter Wert. "Zuletzt verwendet" bezieht sich auf den jüngsten Wert ungleich null vor der zu erstellenden Zeitperiode. Auch hier werden nur echte Werte berücksichtigt.
- **Flag.** Verfügbare Optionen: Leer, Wahr und Falsch.

Maximale Anzahl der Datensätze in resultierendem Daten-Set. Bestimmt die maximal zulässige Anzahl an erstellten Datensätzen, die ansonsten sehr groß würde, insbesondere dann, wenn Sekunden (absichtlich oder versehentlich) als Zeitintervall eingestellt wurden. Bei einer Zeitreihe mit nur zwei Werten (1. Jan. 2000 und 1. Jan. 2001) würden entsprechend 31.536.000 Datensätze erzeugt, wenn die Daten auf Sekunden aufgefüllt würden (60 Sekunden x 60 Minuten x 24 Stunden x 365 Tage). Sobald der angegebene Höchstwert erreicht ist, wird die Verarbeitung unterbrochen und eine Warnmeldung wird angezeigt.

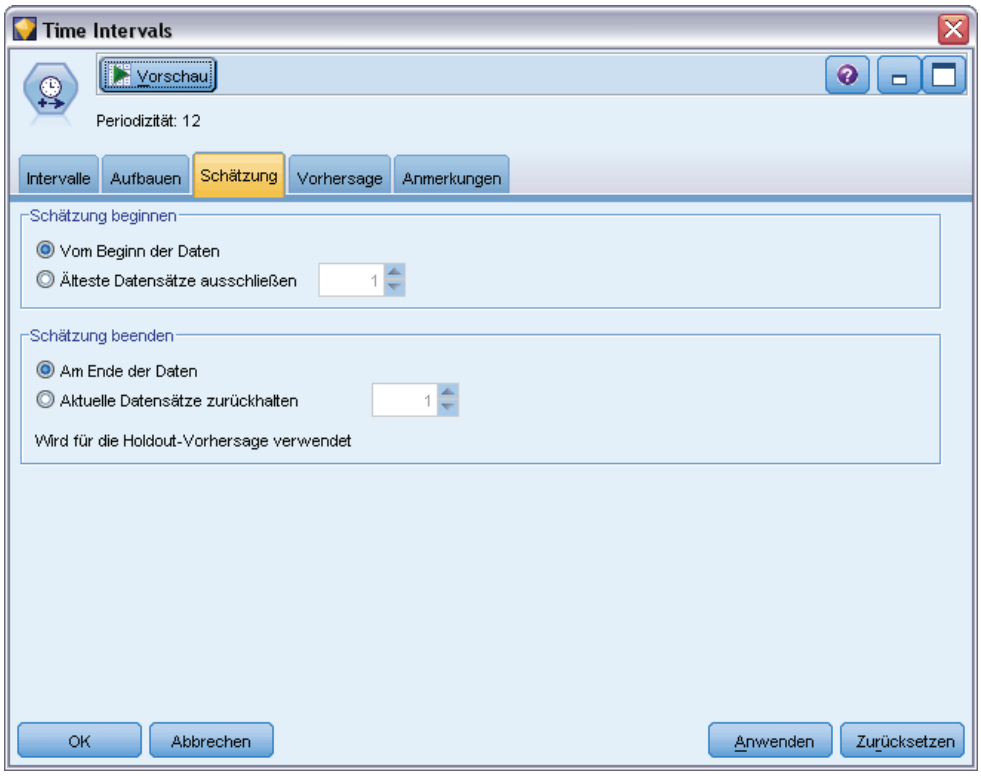
Anzahlfeld

Beim Aggregieren oder Auffüllen von Werten wird ein neues *Anzahlfeld* erzeugt, aus dem die Anzahl der Datensätze hervorgeht, die beim Ermitteln des neuen Datensatzes berücksichtigt werden. Wenn Sie beispielsweise vier Wochenwerte zu einem einzigen Monat aggregieren, ist die Anzahl gleich 4. Bei einem aufgefüllten Datensatz ist die Anzahl gleich 0. Der Name des Felds setzt sich aus der Bezeichnung *Anzahl* und dem auf der Registerkarte "Intervall" angegebenen Präfix oder Suffix zusammen.

Schätzperiode

Auf der Registerkarte "Schätzung" des Zeitintervallknotens können Sie den Bereich der in der Modellschätzung verwendeten Datensätze sowie etwaige Holdouts angeben. Diese Einstellungen können bei Bedarf in weiter unten im Stream liegenden Modellierungsknoten überschrieben werden, es ist jedoch zumeist praktischer, sie hier anzugeben, als für jeden Knoten einzeln.

Abbildung 4-86
Zeitintervallknoten – Registerkarte "Schätzung"



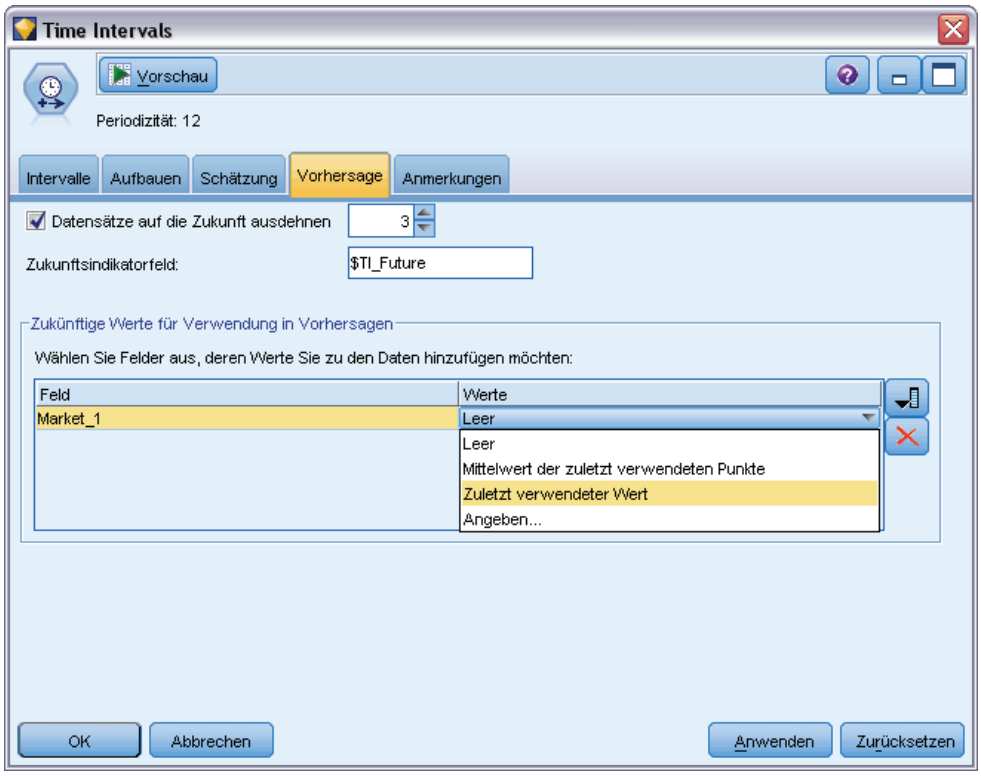
Schätzung beginnen. Sie können die Schätzperiode am Anfang der Daten beginnen oder ältere Werte ausschließen, die bei der Vorhersage nur von begrenztem Nutzen sind. Je nach den jeweils vorliegenden Daten kann eine Verkürzung der Schätzperiode die Leistung beschleunigen (und den für die Datenvorbereitung erforderlichen Zeitaufwand verkürzen), ohne dass signifikante Einbußen bei der Prognosegenauigkeit auftreten.

Schätzung beenden. Sie können das Modell mit allen Datensätzen bis zum Ende der Daten schätzen oder die aktuellsten Datensätze als Holdout zurückhalten, um damit das Modell zu evaluieren. In letzterem Fall "sagen" Sie im Grunde Werte "vorher", die bereits bekannt sind. Auf diese Weise können Sie die beobachteten und die vorhergesagten Werte vergleichen, um die Effektivität des Modells abzuschätzen.

Vorhersagen

Auf der Registerkarte "Vorhersage" des Zeitintervallknotens können Sie die Anzahl der Datensätze angeben, die Sie vorhersagen möchten, sowie die zukünftigen Werte, die bei der Vorhersage durch die weiter unten im Stream liegenden Zeitreihenmodellierungsknoten verwendet werden sollen. Diese Einstellungen können bei Bedarf in weiter unten im Stream liegenden Modellierungsknoten überschrieben werden, es ist jedoch zumeist praktischer, sie hier anzugeben, als für jeden Knoten einzeln.

Abbildung 4-87
Zeitintervallknoten – Registerkarte "Vorhersage"



Datensätze auf die Zukunft ausdehnen. Gibt an, wie viele Datensätze über die Schätzperiode hinaus vorhergesagt werden sollen. Beachten Sie, dass es von der Anzahl der auf der Registerkarte "Schätzung" angegebenen Holdouts abhängt, ob diese Datensätze tatsächlich "vorhergesagt" werden.

Zukunftsindikatorfeld. Beschriftung des generierten Felds, das angibt, ob ein Datensatz Prognosedaten enthält. Der Standardwert für die Beschriftung lautet *$TI_Zukunft*.

Zukünftige Werte für Verwendung in Vorhersagen. Für jeden Datensatz, den Sie vorhersagen möchten (ausgenommen Holdouts) müssen Sie bei Verwendung von Prädiktorfeldern (mit der Rolle auf *Eingabe* eingestellt) für jeden Prädiktor geschätzte Werte für die Vorhersageperiode angeben. Sie können die Werte entweder manuell angeben oder aus einer Liste auswählen.

- **Feld.** Klicken Sie auf die Feldauswahlschaltfläche und wählen Sie alle Felder aus, die als Prädiktor verwendet werden können. Beachten Sie, dass die hier ausgewählten Felder nicht unbedingt bei der Modellierung verwendet werden. Damit ein Feld tatsächlich als Prädiktor verwendet wird, muss es in einem weiter unten im Stream liegenden Modellierungsknoten ausgewählt werden. Dieses Dialogfeld bietet einfach eine praktische Möglichkeit zur Angabe zukünftiger Werte, damit diese gemeinsam von mehreren weiter unten liegenden Modellierungsknoten verwendet werden können und nicht separat in jedem Knoten angegeben werden müssen. Beachten Sie, dass die Liste der verfügbaren Felder durch die auf der Registerkarte "Aufbauen" getroffene Auswahl eingeschränkt sein kann. Wenn beispielsweise auf der Registerkarte "Aufbauen" die Option Felder und Funktionen angeben ausgewählt

wurde, werden alle Felder, die nicht aggregiert oder aufgefüllt wurden, aus dem Stream verworfen und können nicht bei der Modellierung verwendet werden.

Hinweis: Wenn für ein Feld, das nicht mehr im Stream verfügbar ist (weil es verworfen wurde oder weil auf der Registerkarte "Aufbauen" eine neue Auswahl getroffen wurde), zukünftige Werte angegeben werden, wird das Feld auf der Registerkarte "Vorhersage" in roter Farbe angezeigt.

- **Werte.** Sie können bei jedem Feld aus einer Liste von Funktionen wählen oder auf Angeben klicken, um entweder Werte manuell einzugeben oder eine Auswahl aus einer Liste vordefinierter Werte zu treffen. Wenn die Prädiktorfelder sich auf Elemente beziehen, über die Sie die Kontrolle haben oder die anderweitig im Voraus bekannt sind, sollten Sie die Werte manuell eingeben. Wenn Sie beispielsweise die Einnahmen des nächsten Monats für ein Hotel auf der Grundlage der Anzahl der Zimmerreservierungen vorhersagen, können Sie die Anzahl der Reservierungen angeben, die Ihnen tatsächlich für diesen Zeitraum vorliegen. Wenn ein Prädiktorfeld sich dagegen auf Daten bezieht, über die Sie keine Kontrolle haben, wie beispielsweise der Preis einer Aktie, können Sie eine Funktion verwenden, wie beispielsweise den aktuellsten Wert oder den Mittelwert der aktuellsten Punkte.

  Die verfügbaren Funktionen hängen vom Messniveau des Felds ab.

Messniveau	Funktionen
Stetiges oder nominales Feld	Blank Mittelwert der zuletzt verwendeten Punkte Zuletzt verwendeter Wert Specify
Flag-Felder	Blank Zuletzt verwendeter Wert True False Specify

Mittelwert der zuletzt verwendeten Punkte – berechnet den zukünftigen Wert aus dem Mittelwert der letzten drei Datenpunkte.

Zuletzt verwendeter Wert – legt den Wert des aktuellsten Datenpunkts als zukünftigen Wert fest.

Wahr/Falsch – setzt den zukünftigen Wert eines Flag-Felds je nach Angabe auf "Wahr" bzw. "Falsch".

Angeben – Öffnet ein Dialogfeld, in dem Sie zukünftige Werte manuell eingeben oder aus einer vordefinierten Liste auswählen können.

Abbildung 4-88
Angabe der zukünftigen Werte für Prädiktoren

Zukünftige Werte

Hier können Sie zukünftige Werte angeben, die für die Vorhersage durch weiter unten im Stream liegende Zeitreihenmodellierungsknoten verwendet werden sollen. Diese Einstellungen können bei Bedarf in weiter unten im Stream liegenden Modellierungsknoten überschrieben werden, es ist jedoch zumeist praktischer, sie hier anzugeben, als für jeden Knoten einzeln.

Sie können die Werte manuell eingeben oder auf die Auswahlschaltfläche rechts neben dem Dialogfeld klicken, um eine Auswahl aus einer Liste von Werten zu treffen, die für das aktuelle Feld definiert wurden.

Die Anzahl der zukünftigen Werte, die Sie angeben können, entspricht der Anzahl der Datensätze, um die Sie die Zeitreihe auf die Zukunft ausdehnen.

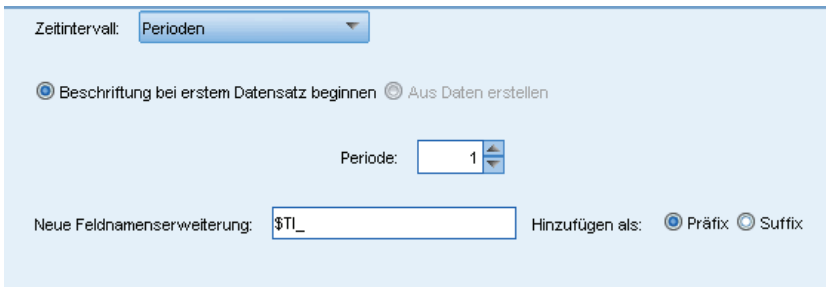
Unterstützte Intervalle

Der Zeitintervallknoten unterstützt die vollständige Palette an Intervallen von Sekunden bis hin zu Jahren, sowie zyklische (z. B. saisonale) und nicht zyklische Perioden. Das Intervall wird auf der Registerkarte “Intervalle” im Feld “Zeitintervall” angegeben.

Periods

Mit der Option Perioden beschriften Sie eine vorhandene, nicht zyklische Zeitreihe, die nicht unter die anderen angegebenen Intervalle fällt. Die Zeitreihe muss bereits die richtige Reihenfolge aufweisen und ein einheitliches Intervall zwischen den einzelnen Messungen besitzen. Bei Auswahl dieses Intervalls ist die Option Aus Daten erstellen nicht verfügbar.

Abbildung 4-89
Zeitintervalleinstellungen für nicht zyklische Perioden



Beispiel

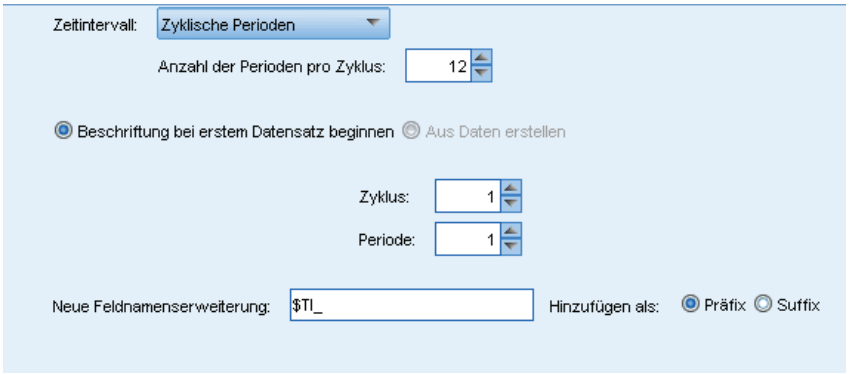
Die Datensätze werden inkrementell gemäß dem angegebenen Anfangswert beschriftet (*Periode 1*, *Periode 2* usw.). Neue Felder werden wie folgt erstellt:

$TI_TimeIndex (ganze Zahl)	$TI_TimeLabel (Zeichenkette)	$TI_Period (ganze Zahl)
1	Periode 1	1
2	Periode 2	2
3	Periode 3	3
4	Periode 4	4
5	Periode 5	5

Zyklische Perioden

Mit der Option Zyklische Perioden beschriften Sie eine vorhandene Zeitreihe mit einem wiederholenden Zyklus, der nicht unter die Standardintervalle fällt. Verwenden Sie diese Option beispielsweise, wenn Ihr Geschäftsjahr nur 10 Monate umfasst. Die Zeitreihe muss bereits die richtige Reihenfolge aufweisen und ein einheitliches Intervall zwischen den einzelnen Messungen besitzen. (Bei Auswahl dieses Intervalls ist die Option Aus Daten erstellen nicht verfügbar.)

Abbildung 4-90
Zeitintervalleinstellungen für zyklische Perioden



Beispiel

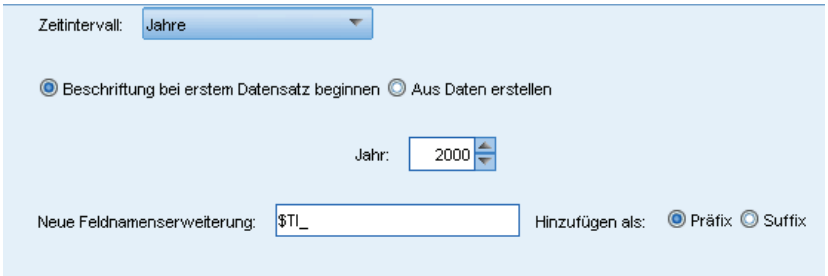
Die Datensätze werden inkrementell gemäß dem angegebenen Anfangszyklus und der Periode beschriftet (*Zyklus 1*, *Periode 1*, *Zyklus 1*, *Periode 2* usw.). Wenn beispielsweise 3 Perioden pro Zyklus festgelegt sind, werden neue Felder wie folgt erstellt:

$TI_TimeIndex (ganze Zahl)	$TI_TimeLabel (Zeichenkette)	$TI_Cycle (ganze Zahl)	$TI_Period (ganze Zahl)
1	Zyklus 1, Periode 1	1	1
2	Zyklus 1, Periode 2	1	2
3	Zyklus 1, Periode 3	1	3
4	Zyklus 2, Periode 1	2	1
5	Zyklus 2, Periode 2	2	2

Years

Bei den Jahren können Sie das Anfangsjahr festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option Aus Daten erstellen ein Zeitstempel- oder Datumsfeld festlegen, aus dem das Jahr für die einzelnen Datensätze hervorgeht.

Abbildung 4-91
Zeitintervalleinstellungen für eine Zeitreihe im Jahresabstand



Beispiel

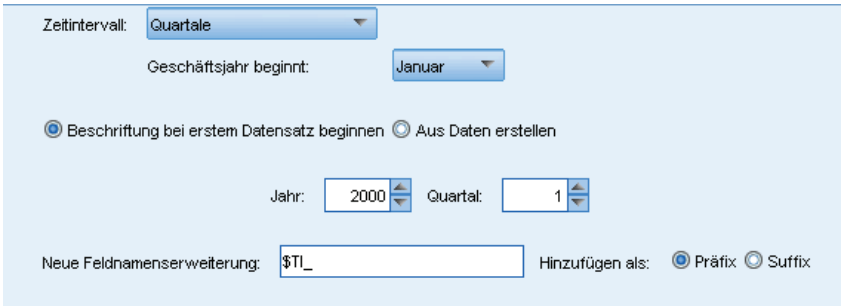
Neue Felder werden wie folgt erstellt:

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeichenkette)	$TI-Year (ganze Zahl)
1	2000	2000
2	2001	2001
3	2002	2002
4	2003	2003
5	2004	2004

Quarters

Bei einer Zeitreihe im Vierteljahresabstand können Sie den Monat festlegen, in dem das Geschäftsjahr beginnt. Darüber hinaus können Sie das Anfangsquartal und -jahr festlegen (z. B. Q1 2000), sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option Aus Daten erstellen ein Zeitstempel- oder Datumsfeld festlegen, aus dem das Quartal und das Jahr für die einzelnen Datensätze hervorgeht.

Abbildung 4-92
Zeitintervalleinstellungen für eine Zeitreihe im Vierteljahresabstand



Beispiel

Bei einem Geschäftsjahr, das im Januar beginnt, werden neue Felder wie folgt erstellt und gefüllt:

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeichenkette)	$TI-Year (ganze Zahl)	$TI-Quarter (ganze Zahl mit Beschriftungen)
1	Q1 2000	2000	1 (Q1)
2	Q2 2000	2000	2 (Q2)
3	Q3 2000	2000	3 (Q3)
4	Q4 2000	2000	4 (Q4)
5	Q1 2001	2001	1 (Q1)

Beginnt das Geschäftsjahr in einem anderen Monat (z. B. im Juli), werden neue Felder wie unten beschrieben erstellt. Sollen die Beschriftungen für die Monate in den einzelnen Quartalen angezeigt werden, aktivieren Sie die Anzeige der Beschriftungen durch Klicken auf das entsprechende Symbol in der Symbolleiste.

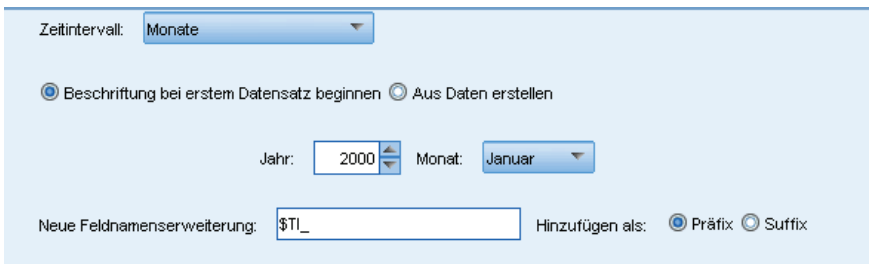
Abbildung 4-93
Symbol für Wertelabels

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeichenkette)	$TI-Year (ganze Zahl)	$TI-Quarter (ganze Zahl mit Beschriftungen)
1	Q1 2000/2001	1	1 (Q1 Jul-Sep)
2	Q2 2000/2001	1	2 (Q2 Okt-Dez)
3	Q3 2000/2001	1	3 (Q3 Jan-Mär)
4	Q4 2000/2001	1	4 (Q4 Apr-Jun)
5	Q1 2001/2002	2	1 (Q1 Jul-Sep)

Months

Sie können das Anfangsjahr und den Anfangsmonat festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option Aus Daten erstellen ein Zeitstempel- oder Datumsfeld festlegen, aus dem der Monat für die einzelnen Datensätze hervorgeht.

Abbildung 4-94
Zeitintervalleinstellungen für eine Zeitreihe im Monatsabstand



Beispiel

Neue Felder werden wie folgt erstellt:

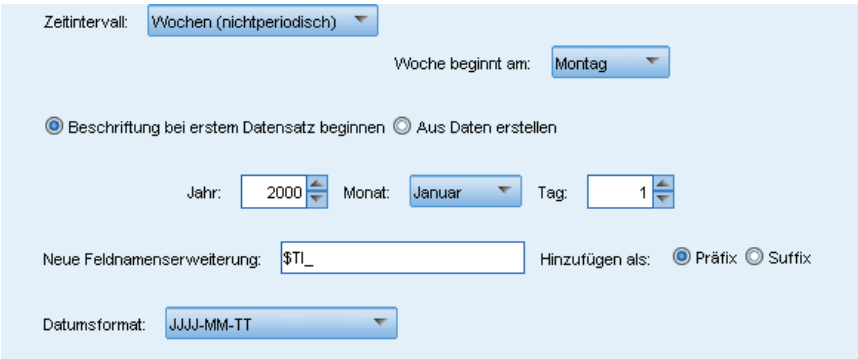
$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Datum)	$TI-Year (ganze Zahl)	$TI-Months (ganze Zahl mit Beschriftungen)
1	Jan 2000	2000	1 (Januar)
2	Feb 2000	2000	2 (Februar)
3	Mär 2000	2000	3 (März)
4	Apr 2000	2000	4 (April)
5	Mai 2000	2000	5 (Mai)

Wochen (nichtperiodisch)

Bei einer Zeitreihe im Wochenabstand können Sie den Tag der Woche angeben, an dem der Zyklus beginnt.

Beachten Sie, dass Wochen nur nichtperiodisch sein können, da verschiedene Monate, Quartale und Jahre nicht unbedingt jeweils dieselbe Anzahl an Wochen aufweisen. Daten mit Zeitstempel können bei nichtperiodischen Modellen jedoch leicht auf Wochenebene aggregiert oder aufgefüllt werden.

Abbildung 4-95
Zeitintervalleinstellungen für eine Zeitreihe im Wochenabstand



Beispiel

Neue Felder werden wie folgt erstellt:

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Datum)	$TI-Week (ganze Zahl)
1	1999-12-27	1
2	2000-01-03	2
3	2000-01-10	3
4	2000-01-17	4
5	2000-01-24	5

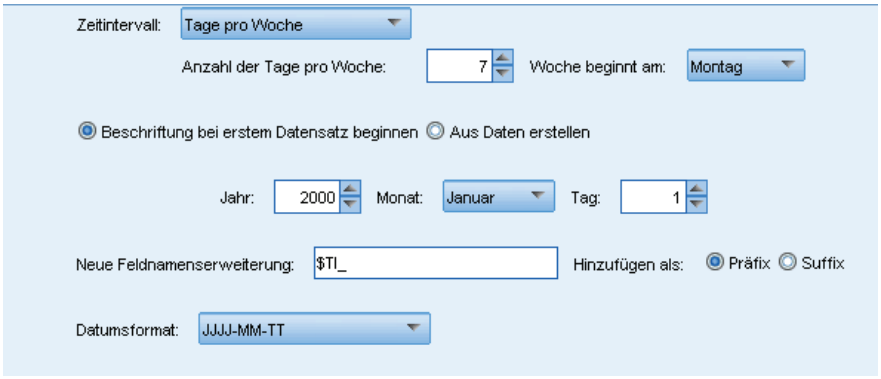
Im Feld *$TI-TimeLabel* für eine Woche wird der erste Tag der betreffenden Woche angezeigt. In der vorherigen Tabelle begann der Benutzer mit der Beschriftung beim 1. Januar 2000. Die Woche beginnt jedoch am Montag und der 1. Januar 2000 ist ein Samstag. Somit beginnt die Woche, in der der 1. Januar liegt, am 27. Dezember 1999 und wird als Beschriftung des ersten Punkts verwendet.

Das Datumsformat bestimmt, welche Zeichenketten für das Feld *$TI-TimeLabel* erstellt werden.

Tage pro Woche

Bei täglichen Messungen, die in einen Wochenzyklus fallen, können Sie die Anzahl der Tage pro Woche angeben und auch den Tag festlegen, an dem die Woche beginnt. Sie können das Anfangsdatum festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option Aus Daten erstellen ein Zeitstempel- oder Datumsfeld festlegen, aus dem das Datum für die einzelnen Datensätze hervorgeht.

Abbildung 4-96
Zeitintervalleinstellungen für eine Zeitreihe im Tagesabstand



Beispiel

Neue Felder werden wie folgt erstellt:

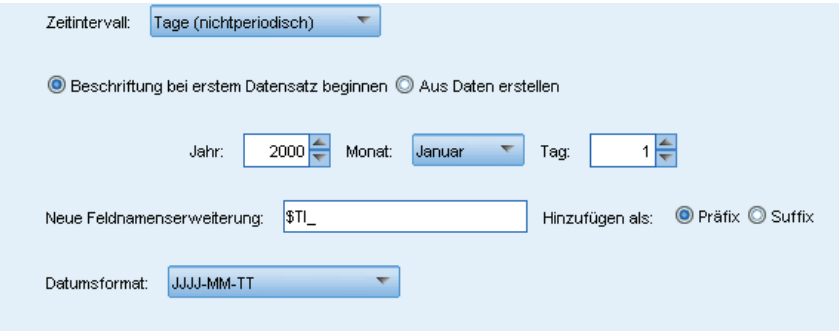
$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Datum)	$TI-Week (ganze Zahl)	$TI-Day (ganze Zahl mit Beschriftungen)
1	Jan 5 2005	1	3 (Mittwoch)
2	Jan 6 2005	1	4 (Donnerstag)
3	Jan 7 2005	1	5 (Freitag)
4	Jan 10 2005	2	1 (Montag)
5	Jan 11 2005	2	2 (Dienstag)

Hinweis: Die Woche beginnt stets mit Tag 1 in der ersten Zeitperiode und weist keinen Kalenderzyklus auf. Auf Woche 52 folgt daher Woche 53, dann Woche 54 usw. Die Woche entspricht nicht der Woche im Jahr, sondern bezeichnet lediglich die Anzahl der Wocheninkremente in der Zeitreihe.

Tage (nichtperiodisch)

Wählen Sie die Option “Tage (nichtperiodisch)” bei täglichen Messungen, die nicht in einen normalen Wochenzyklus fallen. Sie können das Anfangsdatum festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option Aus Daten erstellen ein Zeitstempel- oder Datumsfeld festlegen, aus dem das Datum für die einzelnen Datensätze hervorgeht.

Abbildung 4-97
Zeitintervalleinstellungen für eine Zeitreihe im Tagesabstand (nicht periodisch)



Beispiel

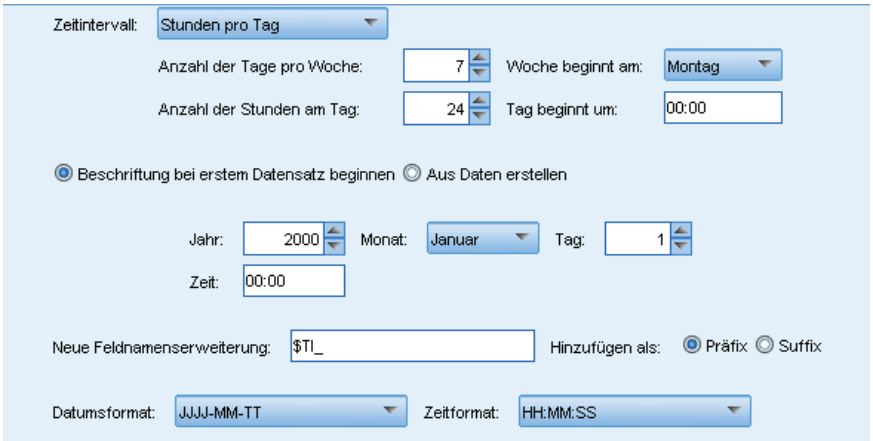
Neue Felder werden wie folgt erstellt:

$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Datum)
1	Jan 5 2005
2	Jan 6 2005
3	Jan 7 2005
4	Jan 8 2005
5	Jan 9 2005

Stunden pro Tag

Bei stündlichen Messungen, die in einen normalen Tageszyklus fallen, können Sie die Anzahl der Tage pro Woche festlegen, außerdem die Anzahl der Stunden pro Tag (z. B. ein Acht-Stunden-Arbeitstag), den Tag, an dem die Woche beginnt, sowie die Stunde, an dem jeder Tag beginnt. Die Stunden können minutengenau im 24-Stunden-System angegeben werden (z. B. 14:05).

Abbildung 4-98
Zeitintervalleinstellungen für eine Zeitreihe im Stundenabstand



Sie können das Anfangsdatum und die Anfangszeit festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option Aus Daten erstellen ein Zeitstempelfeld festlegen, aus dem das Datum und die Uhrzeit für die einzelnen Datensätze hervorgeht.

Beispiel

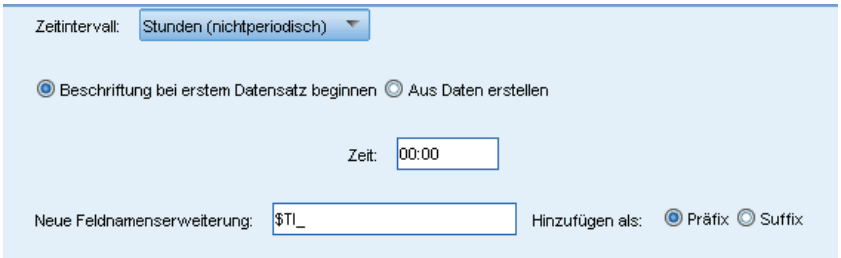
Neue Felder werden wie folgt erstellt:

$TI-TimeIndex (ganze Zahl)	**$TI-TimeLabel (Zeitstempel)**	**$TI-Day (ganze Zahl mit Beschriftungen)**	**$TI-Hour (ganze Zahl mit Beschriftungen)**
1	Jan 5 2005 8:00	3 (Mittwoch)	8 (8:00)
2	Jan 5 2005 9:00	3 (Mittwoch)	9 (9:00)
3	Jan 5 2005 10:00	3 (Mittwoch)	10 (10:00)
4	Jan 5 2005 11:00	3 (Mittwoch)	11 (11:00)
5	Jan 5 2005 12:00	3 (Mittwoch)	12 (12:00)

Stunden (nichtperiodisch)

Wählen Sie diese Option bei stündlichen Messungen, die nicht in einen normalen Tageszyklus fallen. Sie können die Anfangszeit festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option Aus Daten erstellen ein Zeitstempel- oder Zeitfeld festlegen, aus dem die Uhrzeit für die einzelnen Datensätze hervorgeht.

Abbildung 4-99
Zeitintervalleinstellungen für Jahresdaten



Die Stundenangaben beruhen auf dem 24-Stunden-System (13:00); auf die 24. Stunde folgt dabei die 25. Stunde, die Uhrzeit wird also nicht auf 1:00 zurückgesetzt.

Beispiel

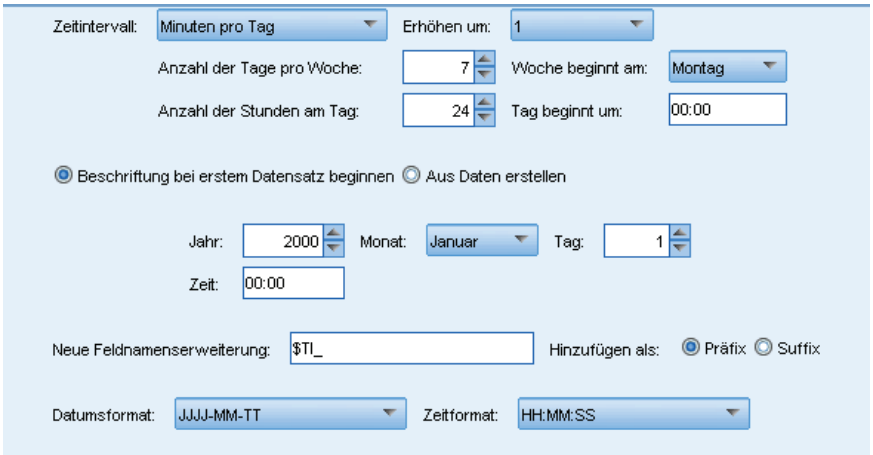
Neue Felder werden wie folgt erstellt:

$TI-TimeIndex (ganze Zahl)	**$TI-TimeLabel (Zeichenkette)**	**$TI-Hour (ganze Zahl mit Beschriftungen)**
1	8:00	8 (8:00)
2	9:00	9 (9:00)
3	10:00	10 (10:00)
4	11:00	11 (11:00)
5	12:00	12 (12:00)

Minuten pro Tag

Bei minutengenauen Messungen, die in einen Tageszyklus fallen, können Sie beispielsweise die Anzahl der Tage in der Woche festlegen, außerdem die Anzahl der Stunden pro Tag sowie die Uhrzeit, zu der der Tag beginnt. Die Uhrzeiten können mithilfe von Doppelpunkten minuten- und sekundengenau im 24-Stunden-System angegeben werden (z. B. 14:05:17). Darüber hinaus können Sie den Zeitraum in Minuten pro Inkrement angeben (minütlich, alle zwei Minuten usw.); das Inkrement muss dabei ein Wert sein, durch den 60 ohne Rest dividiert werden kann.

Abbildung 4-100
Zeitintervalleinstellungen für "Minuten pro Tag"



Sie können das Anfangsdatum und die Anfangszeit festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option Aus Daten erstellen ein Zeitstempelfeld festlegen, aus dem das Datum und die Uhrzeit für die einzelnen Datensätze hervorgeht.

Beispiel

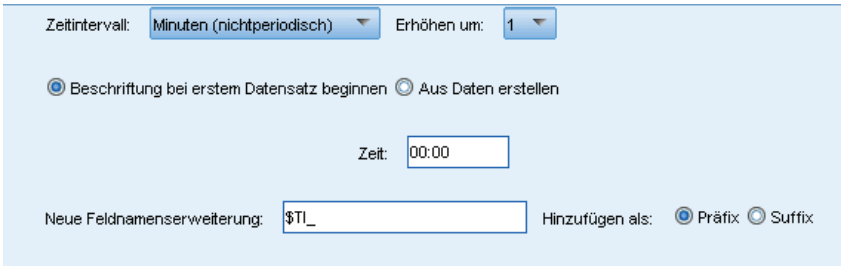
Neue Felder werden wie folgt erstellt:

$TI-TimeIndex (ganze Zahl)	**$TI-TimeLabel (Zeitstempel)**	**$TI-Minute**
1	2005-01-05 08:00:00	0
2	2005-01-05 08:01:00	1
3	2005-01-05 08:02:00	2
4	2005-01-05 08:03:00	3
5	2005-01-05 08:04:00	4

Minuten (nichtperiodisch)

Wählen Sie diese Option bei minütlichen Messungen, die nicht in einen normalen Tageszyklus fallen. Sie können den Zeitraum in Minuten pro Inkrement angeben (minütlich, alle zwei Minuten usw.); das Inkrement muss dabei ein Wert sein, durch den 60 ohne Rest dividiert werden kann.

Abbildung 4-101
Zeitintervalleinstellungen für "Minuten (nichtperiodisch)"



Sie können die Anfangszeit festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option Aus Daten erstellen ein Zeitstempel- oder Zeitfeld festlegen, aus dem die Uhrzeit für die einzelnen Datensätze hervorgeht.

Beispiel

Neue Felder werden wie folgt erstellt:

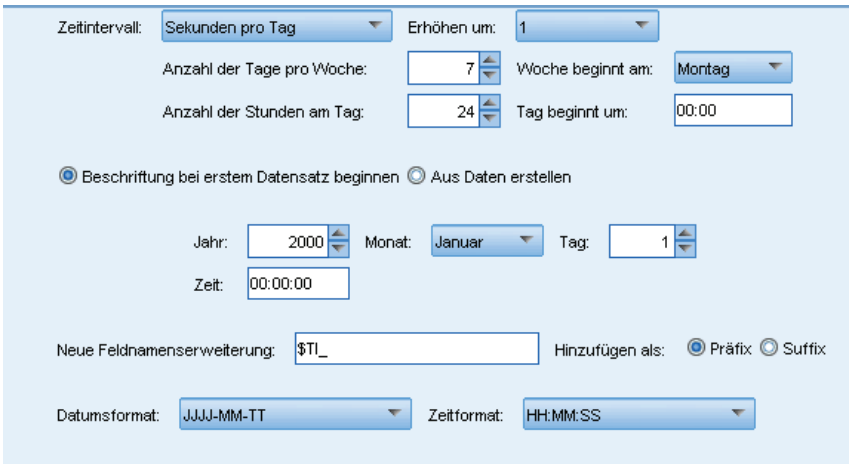
$TI-TimeIndex (ganze Zahl)	$TI-TimeLabel (Zeichenkette)	$TI-Minute
1	8:00	0
2	8:01	1
3	8:02	2
4	8:03	3
5	8:04	4

- In der *TimeLabel*-Zeichenkette sind die Stunden- und Minutenangaben durch einen Doppelpunkt getrennt. Auf die 24. Stunde folgt dabei die 25. Stunde; die Uhrzeit wird also nicht auf 1:00 zurückgesetzt.
- Die Minuten werden gemäß dem im Dialogfeld angegebenen Wert hochgezählt. Beim Inkrement 2 erhält das *TimeLabel*-Feld beispielsweise die Werte 8:00, 8:02 usw.; die Minutenangaben lauten entsprechend 0, 2 usw.

Sekunden pro Tag

Bei Sekundenintervallen, die in einen Tageszyklus fallen, können Sie beispielsweise die Anzahl der Tage in der Woche festlegen, außerdem die Anzahl der Stunden pro Tag sowie die Uhrzeit, zu der der Tag beginnt. Die Uhrzeiten können mithilfe von Doppelpunkten minuten- und sekundengenau im 24-Stunden-System angegeben werden (z. B. 14:05:17). Darüber hinaus können Sie den Zeitraum in Sekunden pro Inkrement angeben (sekündlich, alle zwei Sekunden usw.); das Inkrement muss dabei ein Wert sein, durch den 60 ohne Rest dividiert werden kann.

Abbildung 4-102
Zeitintervalleinstellungen für "Sekunden pro Tag"



Sie können das Anfangsdatum und die Anfangszeit festlegen, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder mit der Option Aus Daten erstellen ein Zeitstempelfeld festlegen, aus dem das Datum und die Uhrzeit für die einzelnen Datensätze hervorgeht.

Beispiel

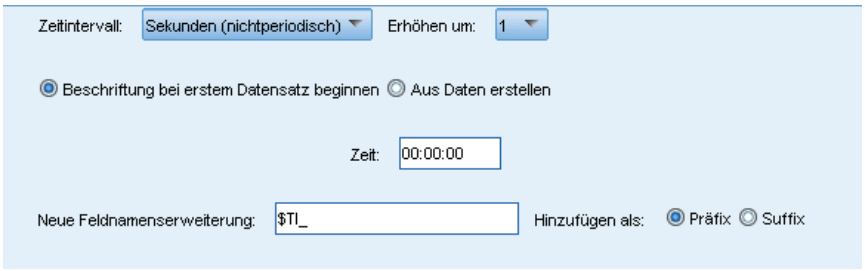
Neue Felder werden wie folgt erstellt:

$TI-TimeIndex (ganze Zahl)	**$TI-TimeLabel (Zeitstempel)**	**$TI-Minute**	**$TI-Second**
1	2005-01-05 08:00:00	0	0
2	2005-01-05 08:00:01	0	1
3	2005-01-05 08:00:02	0	2
4	2005-01-05 08:00:03	0	3
5	2005-01-05 08:00:04	0	4

Sekunden (nichtperiodisch)

Wählen Sie diese Option bei sekündlichen Messungen, die nicht in einen normalen Tageszyklus fallen. Sie können den Zeitraum in Sekunden pro Inkrement angeben (sekündlich, alle zwei Sekunden usw.); das Inkrement muss dabei ein Wert sein, durch den 60 ohne Rest dividiert werden kann.

Abbildung 4-103
Zeitintervalleinstellungen für "Sekunden (nichtperiodisch)"



Legen Sie die Anfangszeit fest, sodass die nachfolgenden Datensätze entsprechend beschriftet werden, oder wählen Sie mit der Option Aus Daten erstellen ein Zeitstempel- oder Zeitfeld, aus dem die Uhrzeit für die einzelnen Datensätze hervorgeht.

Beispiel

Neue Felder werden wie folgt erstellt:

$TI-TimeIndex (ganze Zahl)	**$TI-TimeLabel (Zeichenkette)**	**$TI-Minute**	**$TI-Second**
1	8:00:00	0	0
2	8:00:01	0	1
3	8:00:02	0	2
4	8:00:03	0	3
5	8:00:04	0	4

- In der *TimeLabel*-Zeichenkette sind die Stunden- und Minutenangaben sowie die Minuten- und Sekundenangaben jeweils durch einen Doppelpunkt getrennt. Auf die 24. Stunde folgt dabei die 25. Stunde; die Uhrzeit wird also nicht auf 1:00 zurückgesetzt.
- Die Sekunden werden gemäß dem Wert hochgezählt, der als Inkrement festgelegt ist. Beim Inkrement 2 erhält das *TimeLabel*-Feld beispielsweise die Werte 8:00:00, 8:00:02 usw.; die Sekundenangaben lauten entsprechend 0, 2 usw.

Verlaufsknoten

Verlaufsknoten werden am häufigsten für sequenzielle Daten, beispielsweise Zeitreihendaten, verwendet. Sie dienen zum Erstellen neuer Felder mit Daten aus Feldern in vorangegangenen Datensätzen. Bei Verwendung eines Verlaufsknotens sind meist Daten sinnvoll, die anhand eines bestimmten Felds vorsortiert sind. Dies lässt sich mit einem Sortierknoten erreichen.

Festlegen der Optionen für den Verlaufsknoten

Abbildung 4-104
Verlaufsknoten – Dialogfeld

Ausgewählte Felder. Mithilfe der Feldauswahl-Schaltfläche rechts neben dem Textfeld können Sie die Felder auswählen, für die der Verlauf erstellt werden soll. Alle ausgewählten Felder dienen zur Erstellung neuer Felder für alle Datensätze im Daten-Set.

Offset. Dient zur Angabe des jüngsten Datensatzes vor dem aktuellsten Datensatz, aus dem Verlaufsfeldwerte extrahiert werden sollen. Wenn für "Offset" beispielsweise der Wert "3" festgelegt ist, werden, wenn die einzelnen Datensätze diesen Knoten durchlaufen, die Feldwerte für den dritten vorangegangenen Datensatz in den aktuellen Datensatz aufgenommen. Verwenden Sie die Einstellungen für die Spannweite, um anzugeben, wie weit zurückliegende Datensätze für die Extraktion verwendet werden sollen. Mithilfe der Pfeile können Sie den Offset-Wert einstellen.

Spannweite. Geben Sie an, aus wie vielen früheren Datensätzen Werte extrahiert werden sollen. Beispiel: Wenn für "Offset" der Wert "3" und für "Spannweite" der Wert "5" angegeben wurde, werden jedem Datensatz, der den Knoten durchläuft, fünf Felder für jedes in der Liste "Ausgewählte Felder" angegebene Feld hinzugefügt. Wenn der Knoten also beispielsweise Datensatz 10 verarbeitet, werden für Datensatz 7 bis Datensatz 3 Felder hinzugefügt. Mithilfe der Pfeile können Sie den Wert für die Spannweite einstellen.

Wenn Verlauf nicht verfügbar. Wählen Sie eine der folgenden Optionen für die Behandlung von Datensätzen aus, die keine Verlaufswerte aufweisen. Dies bezieht sich normalerweise auf die ersten Datensätze oben im Daten-Set, für die es keine vorangegangenen Datensätze gibt, die als Verlauf dienen könnten.

- **Datensätze verwerfen.** Wählen Sie diese Option aus, um Datensätze zu verwerfen, wenn für das ausgewählte Feld kein Verlaufswert verfügbar ist.
- **Verlauf undefiniert lassen.** Wählen Sie diese Option aus, um Datensätze beizubehalten, wenn für das ausgewählte Feld kein Verlaufswert verfügbar ist. Das Verlaufsfeld wird mit einem nicht definierten Wert ausgefüllt, der als $null$ angezeigt wird.
- **Werte füllen mit.** Geben Sie einen Wert oder eine Zeichenkette an, die für Datensätze verwendet werden soll, wenn kein Verlaufswert verfügbar ist. Der Standard-Ersatzwert ist *undef*, die systemdefinierte Null. Nullwerte werden mit der Zeichenkette $null$ angezeigt.

Beachten Sie bei der Auswahl eines Ersatzwerts folgende Regeln, damit eine ordnungsgemäße Ausführung erfolgen kann:

- Die ausgewählten Felder sollten denselben Speichertyp aufweisen.
- Wenn alle ausgewählten Felder einen numerischen Speichertyp aufweisen, muss der Ersatzwert als ganze Zahl analysiert werden.
- Wenn alle ausgewählten Felder einen reellen Speichertyp aufweisen, muss der Ersatzwert als reelle Zahl analysiert werden.
- Wenn alle ausgewählten Felder einen symbolischen Speichertyp aufweisen, muss der Ersatzwert als Zeichenkette analysiert werden.
- Wenn alle ausgewählten Felder den Speichertyp "Datum/Uhrzeit" aufweisen, muss der Ersatzwert als Datums-/Uhrzeit-Feld analysiert werden.

Wenn eine der oben angegebenen Bedingungen nicht erfüllt ist, wird bei der Ausführung des Verlaufsknotens eine Fehlermeldung ausgegeben.

Knoten "Felder ordnen"

Mit dem Knoten "Felder ordnen" können Sie die natürliche Reihenfolge definieren, die bei der Anzeige der weiter unten im Stream liegenden Felder verwendet wird. Diese Reihenfolge betrifft die Anzeige von Feldern an unterschiedlichen Stellen, beispielsweise in Tabellen, Listen und in der Feldauswahl. Dieser Vorgang dient beispielsweise dazu, um bei der Arbeit mit umfangreichen Daten-Sets die relevanten Felder deutlicher hervorzuheben.

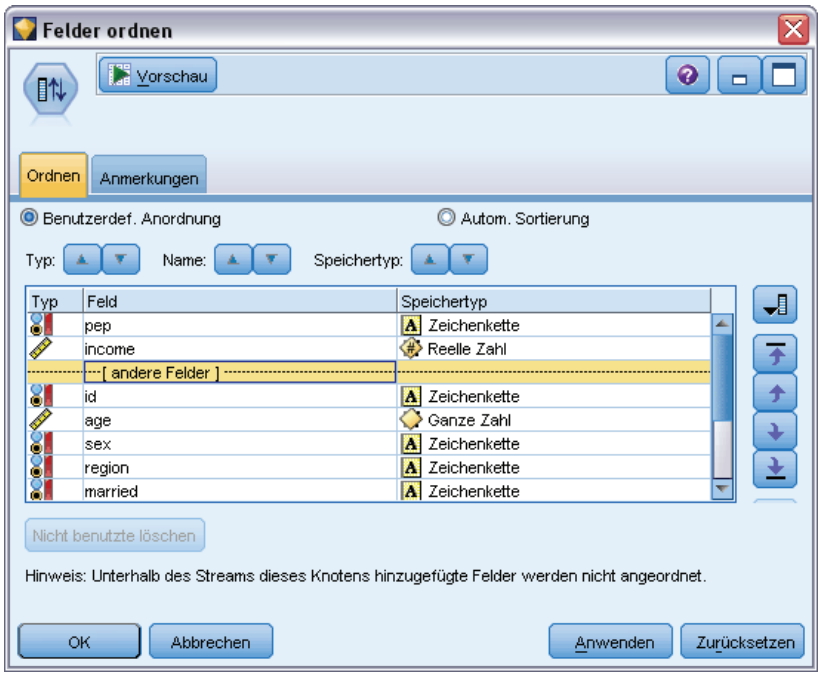
Festlegen der Optionen für "Felder ordnen"

Es gibt zwei Methoden für das Ordnen von Feldern: benutzerdefinierte Anordnung und automatische Sortierung.

Benutzerdefinierte Anordnung

Wählen Sie die Option Benutzerdef. Anordnung, um eine Tabelle mit Feldnamen und -typen zu definieren, in der Sie alle Felder anzeigen und mithilfe der Pfeilschaltflächen eine benutzerdefinierte Anordnung erstellen können.

Abbildung 4-105
Anordnung zur priorisierten Anzeige der relevanten Felder



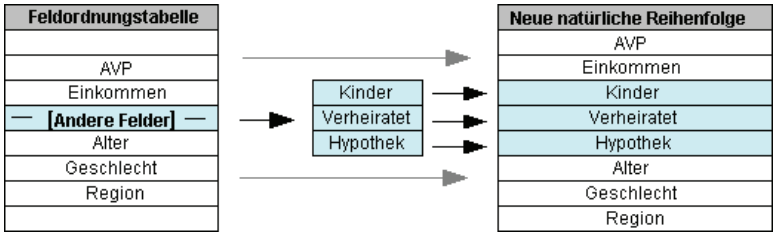
So können Sie Felder neu anordnen:

- Wählen Sie ein Feld in der Tabelle aus. Mittels Strg-Klicken können Sie mehrere Felder auswählen.
- Mithilfe der einfachen Pfeilschaltflächen können Sie die Felder eine Zeile nach oben bzw. unten verschieben.
- Mithilfe der Schaltflächen mit Pfeil und Balken können Sie die Felder ganz nach unten oder oben in der Liste verschieben.
- Die Reihenfolge der hier nicht angegebenen Felder können Sie angeben, indem Sie die mit [andere Felder] bezeichnete Trennzeile nach oben bzw. unten verschieben.

Andere Felder. Die Trennzeile [andere Felder] dient dazu, die Tabelle in zwei Hälften aufzuteilen.

- Die oberhalb der Trennzeile angezeigten Felder werden (wie in der Tabelle zu sehen) oberhalb aller natürlichen Reihenfolgen angeordnet, die zur Anzeige der Felder unterhalb dieses Knotens verwendet werden.
- Die unterhalb der Trennzeile angezeigten Felder werden (wie in der Tabelle zu sehen) unterhalb aller natürlichen Reihenfolgen angeordnet, die zur Anzeige der Felder unterhalb dieses Knotens verwendet werden.

Abbildung 4-106
Diagramm zur Darstellung, wie "andere Felder" in die neue Feldanordnung integriert werden.



- Alle anderen Felder, die nicht in der Tabelle von "Felder ordnen" angezeigt werden, werden zwischen den genannten "oberhalb" und "unterhalb" liegenden Feldern (durch die Lage der Trennzeile gekennzeichnet) angezeigt.

Weitere benutzerdefinierte Sortieroptionen:

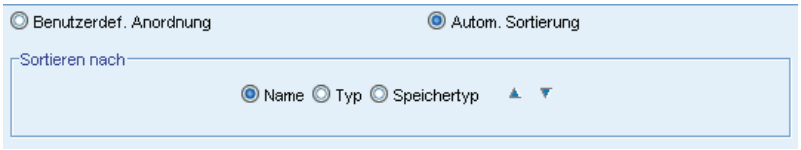
- Sie können Felder in aufsteigender oder absteigender Reihenfolge sortieren, indem Sie auf die Pfeile oberhalb der einzelnen Spaltentitel klicken (Typ, Name und Speichertyp). Beim Sortieren nach Spalte werden die hier nicht angegebenen Felder (durch die Zeile [andere Felder] angezeigt) in ihrer natürlichen Reihenfolge sortiert.
- Klicken Sie auf Nicht benutzte löschen, um alle nicht verwendeten Felder aus dem Knoten "Felder ordnen" zu löschen. Nicht verwendete Felder werden in der Tabelle mit roter Schriftfarbe angezeigt. Diese zeigt an, dass das Feld in vorangegangenen Operationen gelöscht wurde.
- Sie können die Anordnung für alle neuen Felder angeben (angezeigt mit einem Blitzsymbol, das auf ein neues oder nicht spezifiziertes Feld hinweist). Wenn Sie auf OK oder Anwenden klicken, wird das Symbol entfernt.

Hinweis: Wenn weiter oben im Stream Felder hinzugefügt werden, nachdem eine benutzerdefinierte Anordnung durchgeführt wurde, werden die neuen Felder unten in der benutzerdefinierten Liste hinzugefügt.

Automatische Sortierung

Wählen Sie Autom. Sort. aus, um einen Parameter für die Sortierung anzugeben. Die Dialogfeldoptionen ändern sich dynamisch, um Optionen für die automatische Sortierung zu bieten.

Abbildung 4-107
Ordnen aller Felder mithilfe der Optionen für die automatische Sortierung



Sortieren nach. Dient zur Auswahl einer der drei Methoden, die für die Sortierung der in den Knoten “Felder ordnen” eingelesenen Felder zur Verfügung stehen. Die Pfeilschaltflächen zeigen an, ob auf- oder absteigende Sortierreihenfolge verwendet wird. Wählen Sie eine Option aus, um eine Änderung vorzunehmen.

- Name
- Typ
- Speicher

Felder, die nach der Durchführung der automatischen Sortierung oberhalb des Knotens “Felder ordnen” hinzugefügt werden, werden automatisch in die richtige Position (je nach dem ausgewählten Sortiertyp) gebracht.

Kapitel 5

Diagrammknoten

Häufig verwendete Funktionen von Diagrammknoten

Die in IBM® SPSS® Modeler eingebrachten Daten werden in verschiedenen Phasen des Data Mining-Prozesses mithilfe von Diagrammen und Grafiken untersucht. Verbinden Sie beispielsweise einen Plot- oder einen Verteilungsknoten mit einer Datenquelle und informieren Sie sich über die Datentypen und die Verteilungen. Anschließend können Sie Datensätze und Felder bearbeiten und die Daten so für Downstream-Modellierungsoperationen vorbereiten. Darüber hinaus werden Diagramme häufig zum Prüfen der Verteilung und der Beziehungen zwischen neu abgeleiteten Feldern eingesetzt.

Die Palette "Diagramme" enthält die folgenden Knoten:

Der Diagrammtafelknoten bietet viele verschiedene Diagrammtypen in einem einzigen Knoten. Bei Verwendung dieses Knotens können Sie die Datenfelder auswählen, die Sie untersuchen möchten, und anschließend eines der für die ausgewählten Daten verfügbaren Diagramme auswählen. Der Knoten filtert automatisch alle Diagrammtypen heraus, die nicht für die Feldauswahl geeignet sind. Für weitere Informationen siehe Thema Diagrammtafelknoten auf S. 253.

Der Plotknoten zeigt die Beziehung zwischen numerischen Feldern an. Sie können einen Plot mithilfe von Punkten (Streudiagramm) oder mit Linien erstellen. Für weitere Informationen siehe Thema Plotknoten auf S. 303.

Der Verteilungsknoten zeigt das Auftreten symbolischer (kategorialer) Werte wie beispielsweise Hypothekenart oder Geschlecht. Verteilungsknoten eignen sich insbesondere zum Aufzeigen von Ungleichgewichten in den Daten, die mithilfe eines Balancierungsknotens vor dem Erstellen eines Modells ausgeglichen werden können. Für weitere Informationen siehe Thema Verteilungsknoten auf S. 312.

Der Histogrammknoten zeigt das Auftreten bestimmter Werte in numerischen Feldern. Damit werden häufig die Daten vor der weiteren Bearbeitung und der Modellerstellung untersucht. Ähnlich wie der Verteilungsknoten kann der Histogrammknoten oft Ungleichgewichte in den Daten aufdecken. Für weitere Informationen siehe Thema Histogramm – Registerkarte "Plot" auf S. 318.

Der Sammlungsknoten zeigt die Verteilung der Werte für ein numerisches Feld im Verhältnis zu den Werten eines anderen an. (Er erstellt histogrammähnliche Diagramme.) Er eignet sich besonders für die Darstellung einer Variablen oder eines Felds, dessen Werte sich mit der Zeit verändern. Mithilfe eines 3-D-Diagramms können Sie außerdem eine symbolische Achse anlegen, auf der die Verteilungen nach Kategorie aufgetragen sind. Für weitere Informationen siehe Thema Sammlung – Registerkarte "Plot" auf S. 322.

Ein Multidiagramm erstellt ein Plot, bei dem mehrere *Y*-Felder über einem einzelnen *X*-Feld dargestellt werden. Die *Y*-Felder werden als farbige Linien geplottet, die jeweils einem Plotknoten mit dem Stil Linie und dem X-Modus Sortieren entsprechen. Multidiagramme sind hilfreich, wenn die Fluktuation mehrerer Variablen im Laufe der Zeit untersucht werden soll. Für weitere Informationen siehe Thema Multidiagrammknoten auf S. 327.

Der Netzdiagrammknoten zeigt die Stärke der Beziehung zwischen den Werten aus mindestens zwei symbolischen (kategorialen) Feldern. Im Diagramm wird die Verbindungsstärke durch unterschiedlich breite Linien angezeigt. Mit Netzdiagrammknoten können Sie beispielsweise die Beziehung zwischen dem Kauf einer Gruppe von Artikeln auf einer e-Commerce-Website untersuchen. Für weitere Informationen siehe Thema Netzdiagrammknoten auf S. 331.

Der Zeitdiagrammknoten zeigt ein oder mehrere Sets mit Zeitreihendaten an. Normalerweise wird zuerst mithilfe eines Zeitintervallknotens ein *TimeLabel*-Feld erstellt, das dann zur Beschriftung der *x*-Achse verwendet wird. Für weitere Informationen siehe Thema Zeitdiagrammknoten auf S. 342.

Der Evaluationsknoten erleichtert die Evaluation und den Vergleich von Vorhersagemodellen. Das Evaluationsdiagramm zeigt, wie gut Modelle bestimmte Ergebnisse vorhersagen. Die Datensätze werden auf der Grundlage des vorhergesagten Werts und des Konfidenzwerts für die Prognose sortiert. Die Datensätze werden in gleich große Gruppen (**Quantile**) aufgeteilt. Anschließend wird der Wert des Geschäftskriteriums für jedes Quantil geplottet, vom höchsten Wert bis zum niedrigsten Wert. Mehrere Modelle werden als separate Linien im Plot dargestellt. Für weitere Informationen siehe Thema Evaluationsknoten auf S. 347.

Nachdem Sie einen Diagrammknoten zu einem Stream hinzugefügt haben, können Sie durch Doppelklicken auf den Knoten ein Dialogfeld zur Angabe von Optionen öffnen. Die meisten Diagramme enthalten eine Reihe spezieller Optionen, die auf einer oder mehreren Registerkarten gruppiert sind. Des Weiteren stehen einige Registerkartenoptionen zur Verfügung, die allen Diagrammen gemeinsam sind. In den nachstehenden Abschnitten finden Sie weitere Informationen zu diesen gemeinsamen Optionen.

Nachdem Sie die Optionen für einen Diagrammknoten konfiguriert haben, können Sie ihn im Dialogfeld oder als Teil eines Streams ausführen. Im generierten Diagrammfenster können Sie Ableitungsknoten (Set und Flag) und Auswahlknoten auf der Grundlage einer Datenauswahl bzw. eines Datenbereichs generieren, wodurch eine Untergruppe der Daten erstellt wird. Diese leistungsstarke Funktion kann beispielsweise zur Ermittlung und zum Ausschluss von Ausreißern verwendet werden.

Formatierungen, Überlagerungen, Fenster und Animation

Überlagerungen und Formatierungen

Formatierung (und Überlagerungen) verleihen einer Visualisierung Dimensionalität. Die Wirkung einer Formatierung (Gruppierung, Cluster oder Stapeln) hängt vom Visualisierungstyp, dem Feld-/Variablentyp sowie dem Grafikelementtyp und der Statistik ab. So kann beispielsweise ein kategoriales Feld für die Farbe verwendet werden, um Punkte in einem Streudiagramm zu gruppieren oder die Stapel in einem gestapelten Balkendiagramm zu erstellen. Außerdem kann

ein stetiger numerischer Bereich für die Farbe verwendet werden, um die Werte des Bereichs für jeden Punkt in einem Streudiagramm anzugeben.

Sie sollten die verschiedenen Formatierungsmöglichkeiten und Überlagerungen ausprobieren, um diejenige zu finden, die Ihren Bedürfnissen am besten entspricht. Die folgenden Beschreibungen sollen Ihnen die Auswahl erleichtern.

Hinweis: Nicht alle Formatierungen und Überlagerungen stehen für alle Virsualisierungstypen zur Verfügung.

- **Farbe.** Wenn die Farbe durch ein kategoriales Feld festgelegt wird, wird die Visualisierung anhand der einzelnen Kategorien aufgespalten: Jede Kategorie erhält eine andere Farbe. Wenn die Farbe durch einen stetigen numerischen Bereich angegeben wird, variiert die Farbe je nach dem Wert des Bereichsfelds. Wenn das Grafikelement (z. B. ein Balken oder eine Box) für mehrere Datensätze/Fälle steht und ein Bereichsfeld für die Farbe verwendet wird, variiert die Farbe je nach dem *Mittelwert* des Bereichsfelds.
- **Form.** Die Form wird durch ein kategoriales Feld festgelegt, wodurch die Visualisierung in Elemente mit verschiedenen Formen aufgespalten wird: Jede Kategorie erhält eine eigene Form.
- **Transparenz.** Wenn die Transparenz durch ein kategoriales Feld festgelegt wird, wird die Visualisierung anhand der einzelnen Kategorien aufgespalten: Jede Kategorie erhält eine andere Transparenzebene. Wenn die Transparenz durch einen stetigen numerischen Bereich angegeben wird, variiert die Transparenz je nach dem Wert des Bereichsfelds. Wenn das Grafikelement (z. B. ein Balken oder eine Box) für mehrere Datensätze/Fälle steht und ein Bereichsfeld für die Transparenz verwendet wird, variiert die Farbe je nach dem *Mittelwert* des Bereichsfelds. Beim höchsten Wert sind die Grafikelemente vollständig transparent. Beim kleinsten Wert sind sie vollständig undurchsichtig.
- **Datenbeschriftung.** Datenbeschriftungen werden durch ein Feld beliebigen Typs festgelegt. Dabei werden anhand des Werts des Felds Beschriftungen erstellt, mit denen die Grafikelemente versehen werden.
- **Größe.** Wenn die Größe durch ein kategoriales Feld festgelegt wird, wird die Visualisierung anhand der einzelnen Kategorien aufgespalten: Jede Kategorie erhält eine andere Größe. Wenn die Größe durch einen stetigen numerischen Bereich angegeben wird, variiert die Größe je nach dem Wert des Bereichsfelds. Wenn das Grafikelement (z. B. ein Balken oder eine Box) für mehrere Datensätze/Fälle steht und ein Bereichsfeld für die Größe verwendet wird, variiert die Größe je nach dem *Mittelwert* des Bereichsfelds.

Abbildung 5-1
Diagramm mit Farbüberlagerungsformatierung

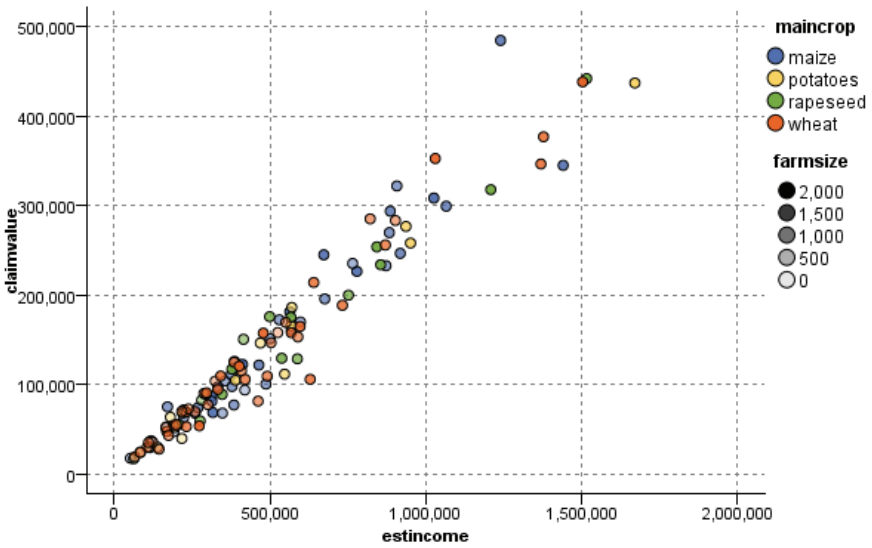
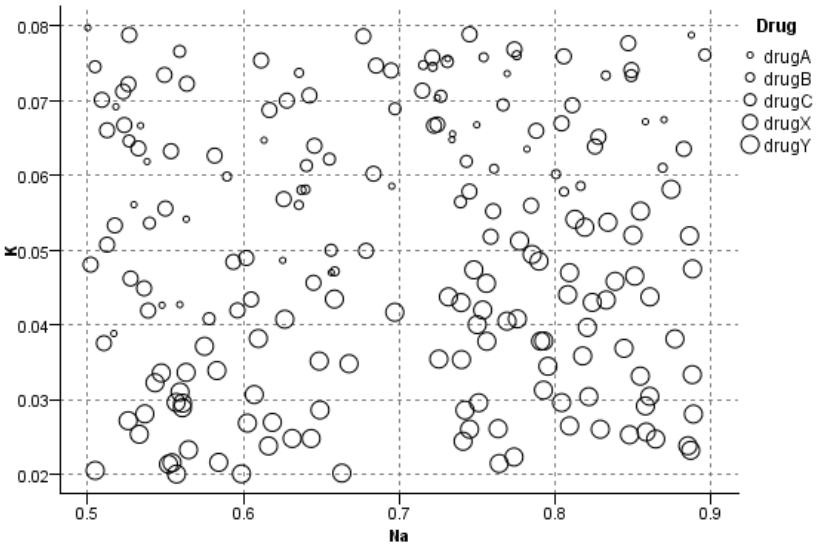


Abbildung 5-2
Diagramm mit Größenüberlagerungsformatierung

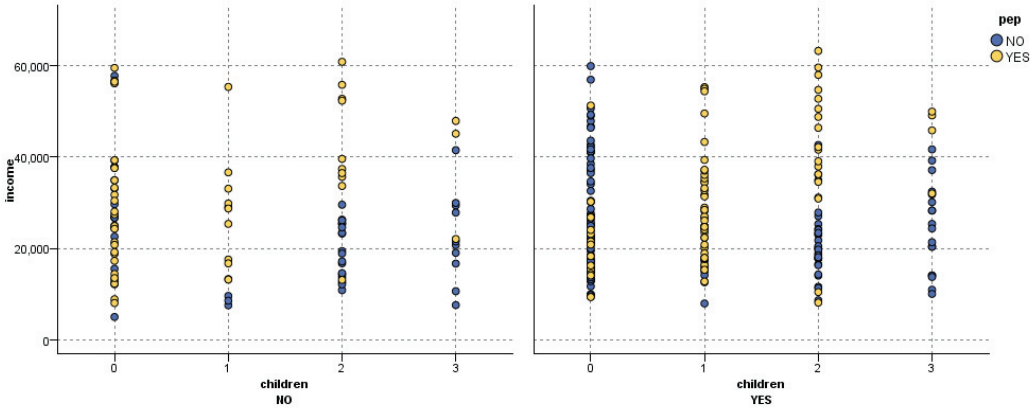


Fenstereinteilung und Animation

Einteilung in Felder. Bei der Fenstereinteilung (Facettierung) wird eine Tabelle mit Diagrammen erstellt. Für jede Kategorie in den Fenstereinteilungsfeldern wird ein Diagramm erstellt, aber alle diese Fenster werden gleichzeitig angezeigt. Die Fenstereinteilung ist nützlich, um zu überprüfen,

ob die Visualisierung den Bedingungen der Fenstereinteilungsfelder unterworfen ist. Sie können beispielsweise ein Histogramm nach dem Geschlecht in Fenster einteilen, um zu ermitteln, ob die Häufigkeitsverteilungen bei Männern und Frauen gleich sind. Damit können Sie also ermitteln, ob das Gehalt geschlechtsabhängig ist. Wählen Sie ein kategoriales Feld für die Fenstereinteilung aus.

Abbildung 5-3
Diagramm mit Fenstern nach "verheiratet" (JA/NEIN)



Animation. Die Animation ähnelt der Einteilung in Fenster dahin gehend, dass aus den Werten des Animationsfelds mehrere Diagramme erstellt werden. Diese Diagramme werden jedoch nicht gemeinsam angezeigt. Stattdessen können Sie mithilfe der Steuerelemente im Interaktionsmodus die Ausgabe animieren und eine Folge einzelner Diagramme durchblättern. Außerdem ist für die Animation im Gegensatz zur Fenstereinteilung kein kategoriales Feld erforderlich. Sie können ein stetiges Feld angeben, dessen Werte automatisch in Bereiche unterteilt werden. Die Größe des Bereichs lässt sich mithilfe der Animationssteuerelemente im Interaktionsmodus variieren. Nicht bei allen Visualisierungen ist eine Animation möglich.

Abbildung 5-4
Animierter Plot mithilfe einer Variablen mit drei Kategorien – Schieberegler bei niedrigem Blutdruck

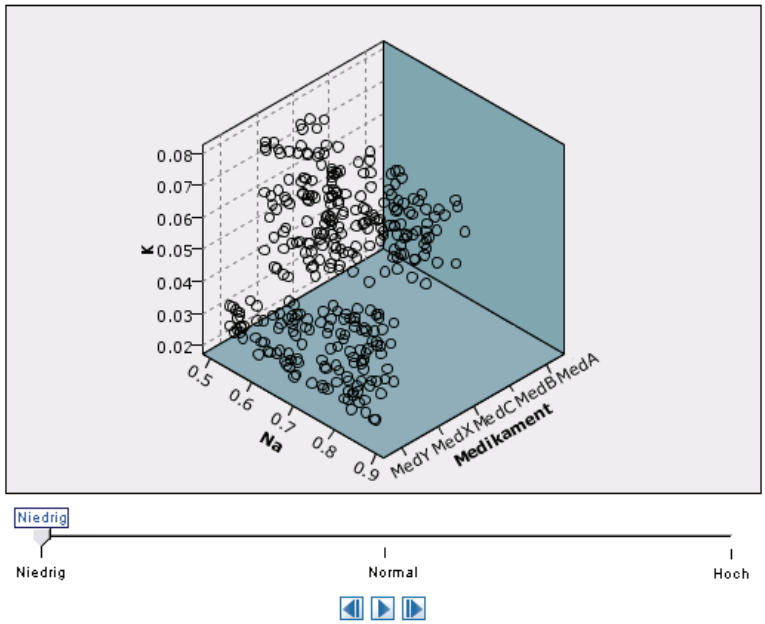


Abbildung 5-5
Animierter Plot mithilfe einer Variablen mit drei Kategorien – Schieberegler bei normalem Blutdruck

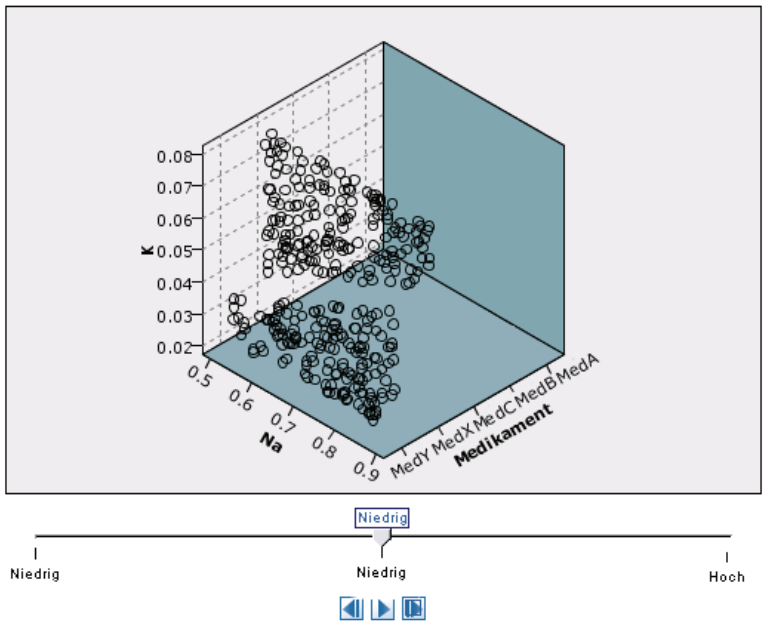
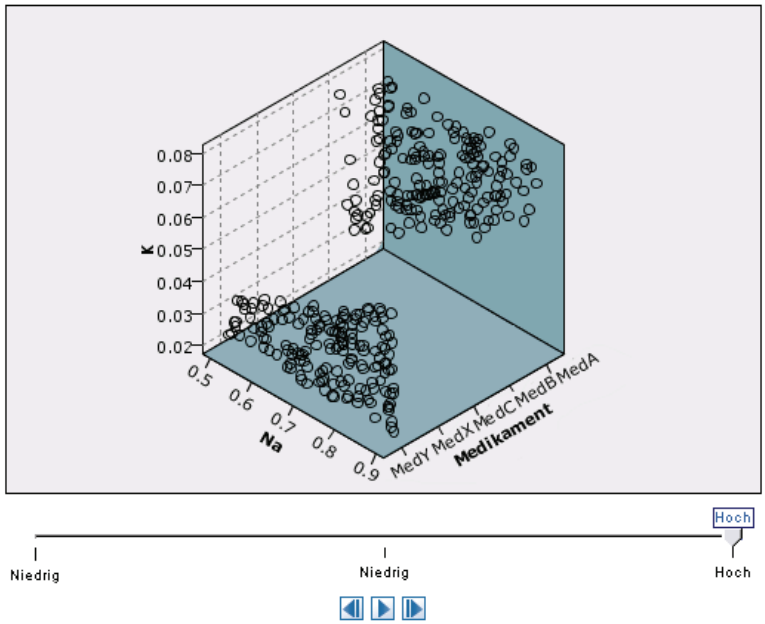


Abbildung 5-6
Animierter Plot mithilfe einer Variablen mit drei Kategorien – Schieberegler bei hohem Blutdruck



Die Registerkarte "Ausgabe"

Bei allen Diagrammtypen können Sie die nachstehenden Optionen für den Dateinamen und die Anzeige der erzeugten Diagramme festlegen.

Hinweis: Für die Diagramme von Verteilungsknoten gelten zusätzliche Einstellungen.

Ausgabename. Bestimmt den Namen des Diagramms, das beim Ausführen des Knotens erstellt wird. Mit Auto wird ein Name auf der Grundlage des Knotens bestimmt, mit dem die Ausgabe erzeugt wird. Optional können Sie auch Angepasst auswählen und einen anderen Namen angeben.

Ausgabe auf Bildschirm. Hiermit lassen Sie das Diagramm in einem neuen Fenster erzeugen und anzeigen.

Ausgabe in Datei. Hiermit wird die Ausgabe als Datei gespeichert.

- **Ausgabediagramm.** Hiermit erstellen Sie Ausgaben in einem Diagrammformat. Nur bei Verteilungsknoten verfügbar.
- **Ausgabetabelle.** Hiermit erstellen Sie Ausgaben in einem Tabellenformat. Nur bei Verteilungsknoten verfügbar.
- **Dateiname.** Geben Sie einen Dateinamen für das erzeugte Diagramm bzw. die erzeugte Tabelle an. Mit der Auslassungsschaltfläche (...) legen Sie eine Datei und einen Pfad fest.
- **Dateityp.** Dient zur Auswahl des Dateityps in der Dropdown-Liste. Für alle Diagrammknoten mit Ausnahme des Verteilungsknotens mit der Option Ausgabetabelle stehen folgende Dateitypen für Diagramme zur Verfügung:

  - Bitmap (*.bmp*)

- PNG (*.png*)
- Ausgabeobjekt (*.cou*)
- JPEG (*.jpg*)
- HTML (*.html*)
- ViZml-Dokument (*.xml*) zur Verwendung in anderen IBM® SPSS® Statistics-Anwendungen.

Für die Option Ausgabetabelle im Verteilungsknoten stehen folgende Dateitypen zur Verfügung:

- Tabulatorgetrennte Daten (*.tab*)
- Kommagetrennte Daten (*.csv*)
- HTML (*.html*)
- Ausgabeobjekt (*.cou*)

Ausgabe paginieren. Beim Speichern der Ausgabe als HTML-Datei wird diese Option zur Verfügung gestellt, damit Sie die Größe der einzelnen HTML-Seiten festlegen können. (Gilt nur für den Verteilungsknoten.)

Zeilen pro Seite. Bei Auswahl von Ausgabe paginieren wird diese Option zur Verfügung gestellt, damit Sie die Länge der einzelnen HTML-Seiten festlegen können. Die Standardeinstellung sind 400 Zeilen. (Gilt nur für den Verteilungsknoten.)

Die Registerkarte "Anmerkungen"

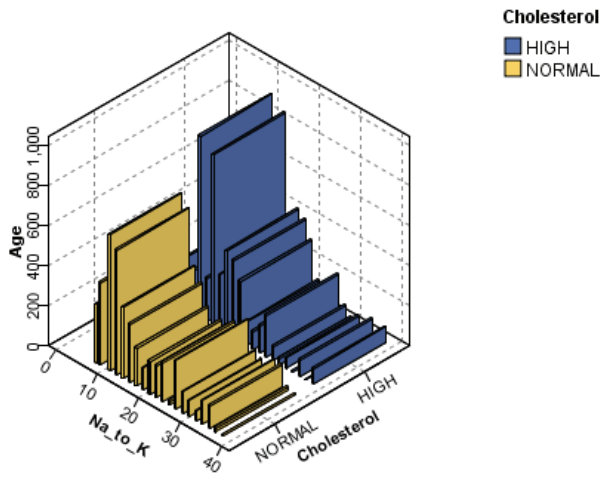
Wird für alle Knoten verwendet. Die Registerkarte bietet Optionen zum Umbenennen von Knoten, zum Anzeigen einer benutzerdefinierten QuickInfo und zum Speichern einer längeren Anmerkung.

3-D-Diagramme

Bei Plots und Sammlungsdiagrammen in IBM® SPSS® Modeler können Daten auf einer dritten Achse dargestellt werden. Auf diese Weise erhalten Sie eine noch größere Flexibilität bei der Visualisierung der Daten zur Auswahl von Untergruppen oder zum Ableiten neuer Felder für die Modellierung.

Nachdem Sie ein 3-D-Diagramm erstellt haben, können Sie darauf klicken und mit der Maus ziehen, um es zu drehen und aus jedem beliebigen Winkel zu betrachten.

Abbildung 5-7
Sammlungsdiagramm mit x-, y- und z-Achse



Für das Erstellen von 3-D-Diagrammen in SPSS Modeler stehen zwei Verfahren zur Auswahl: Daten auf einer dritten Achse plotten (echte 3-D-Diagramme) oder Diagramme mit 3-D-Effekten anzeigen lassen. Beide Verfahren sind sowohl für Plots als auch für Sammlungen verfügbar.

So plotten Sie Daten auf einer dritten Achse:

- Klicken Sie im Dialogfeld des Diagrammknotens auf die Registerkarte Plot.
- Klicken Sie auf die 3-D-Schaltfläche. Die Optionen für die *z*-Achse werden aktiviert.
- Wählen Sie mit der Feldauswahl-Schaltfläche ein Feld für die *z*-Achse aus. In bestimmten Fällen sind hier nur symbolische Felder zulässig. In der Feldauswahl werden die entsprechenden Felder aufgeführt.

So statten Sie ein Diagramm mit 3-D-Effekten aus:

- Erstellen Sie ein Diagramm und klicken Sie im Ausgabefenster auf die Registerkarte Diagramm.
- Klicken Sie auf die 3-D-Schaltfläche. Die Ansicht wechselt zu einem dreidimensionalen Diagramm.

Diagrammtafelknoten

Der Diagrammtafelknoten ermöglicht die Auswahl aus vielen verschiedenen Diagrammausgaben (Balkendiagramme, Kreisdiagramme, Histogramme, Streudiagramme, Hitzekarten usw.) in einem einzigen Knoten. Auf der ersten Registerkarte wählen Sie zunächst die zu untersuchenden Datenfelder aus. Anschließend stellt Ihnen der Knoten eine Reihe von Diagrammtypen zur Auswahl, die für Ihre Daten geeignet sind. Der Knoten filtert automatisch alle Diagrammtypen

heraus, die nicht für die Feldauswahl geeignet sind. Detailliertere bzw. erweiterte Diagrammoptionen können Sie auf der Registerkarte “Detailliert” definieren.

Hinweis: Sie müssen den Diagrammtafelknoten mit einem Stream mit Daten verbinden, um den Knoten bearbeiten oder Diagrammtypen auswählen zu können.

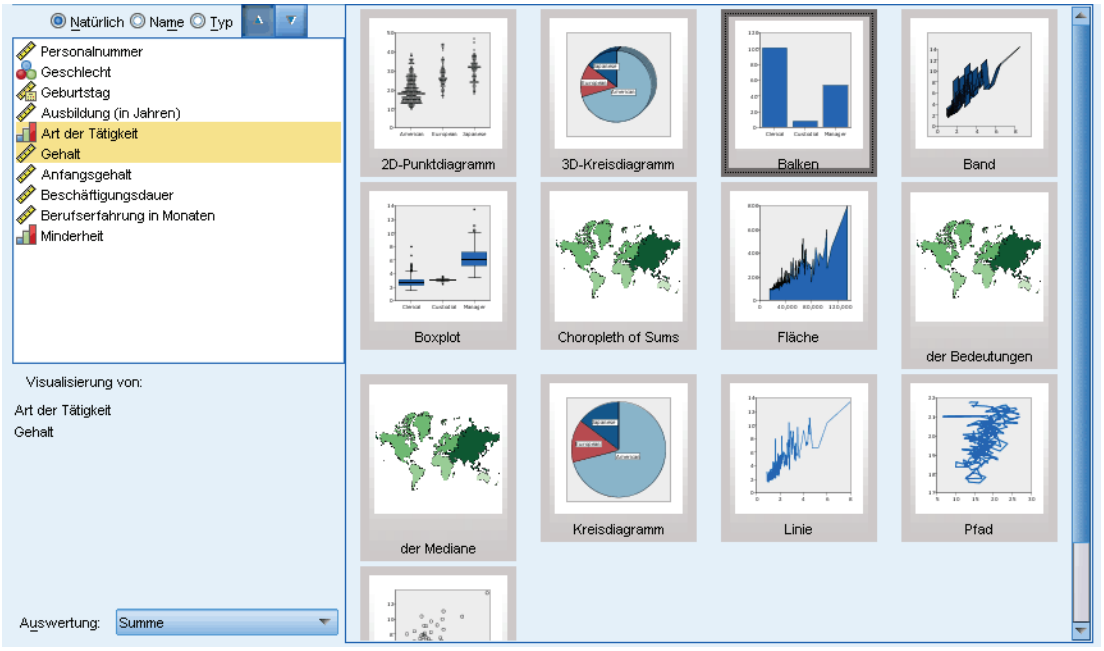
Es gibt zwei Schaltflächen, über die Sie steuern können, welche Visualisierungsvorlagen (sowie Stylesheets und Karten) verfügbar sind:

Verwalten. Verwalten von Visualisierungsvorlagen, Stylesheets und Karten auf dem Computer. Sie können Visualisierungsvorlagen , Stylesheets und Karten auf Ihrem lokalen Rechner importieren, exportieren, umbenennen und löschen. Für weitere Informationen siehe Thema Verwalten von Vorlagen, Stylesheets und Kartendateien auf S. 292.

Speicherort. Ändern des Speicherorts von Visualisierungsvorlagen, Stylesheets und Karten. Der aktuelle Speicherort wird rechts neben der Schaltfläche angezeigt. Für weitere Informationen siehe Thema Einstellen des Speicherorts für Vorlagen, Stylesheets und Karten auf S. 290.

Diagrammtafel – Registerkarte “Einfach”

Abbildung 5-8
Registerkarte “Einfach”



Wenn Sie sich nicht sicher sind, welcher Visualisierungstyp Ihre Daten am besten darstellt, sollten Sie die Registerkarte “Einfach” verwenden. Wenn Sie die Daten auswählen, wird Ihnen die Untermenge der für die Daten geeigneten Visualisierungstypen angezeigt. Beispiele finden Sie unter Grafiktafel Beispiele auf S. 270.

- Wählen Sie mindestens ein Feld/eine Variable in der Liste aus. Mit Strg+Klicken können Sie mehrere Felder auswählen.

Beachten Sie, dass das Messniveau des Felds den Typ der verfügbaren Visualisierungen bestimmt. Sie können das Messniveau ändern, indem Sie das Feld in der Liste durch einen Klick mit der rechten Maustaste auswählen und eine Option wählen. Weitere Informationen zu den verfügbaren Messniveautypen finden Sie unter Feld- bzw. Variablentypen auf S. 257.

- Wählen Sie einen Visualisierungstyp aus. Beschreibungen der verfügbaren Typen finden Sie unter Verfügbare integrierte -Grafiktafel-Visualisierungstypen auf S. 262.

- Bei bestimmten Visualisierungen können Sie eine statistische Funktion (Übersichtsstatistik) auswählen. Je nachdem, ob die Statistik anzahlbasiert ist oder aus einem stetigen Feld berechnet wird, stehen verschiedene Untergruppen mit Statistiken zur Verfügung. Die verfügbaren Statistiken hängen außerdem von der Vorlage selbst ab. Eine vollständige Liste von Statistiken, die ggf. zur Verfügung stehen, finden Sie im Anschluss an den nächsten Schritt.

- Wenn Sie weitere Optionen definieren möchten, beispielsweise optionale Formatierungen und Felder für Fenster, klicken Sie auf Detailliert. Für weitere Informationen siehe Thema Grafiktafel Registerkarte “Detailliert” auf S. 259.

Anhand eines stetigen Felds berechnete statistische Funktionen

- **Mittelwert.** Ein Lagemaß. Die Summe der Ränge, geteilt durch die Zahl der Fälle.
- **Median.** Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).
- **Modalwert.** Der am häufigsten auftretende Wert. Wenn mehrere Werte gleichermaßen die größte Häufigkeit aufweisen, ist jeder von ihnen ein Modalwert.
- **Minimum.** Der kleinste Wert einer numerischen Variablen.
- **Maximum.** Der größte Wert einer numerischen Variablen.
- **Bereich**. Differenz zwischen Mindest- und Höchstwert.
- **Mittelbereich**. Die Mitte des Bereichs, d. h. der Wert, dessen Differenz zum Minimum gleich seiner Differenz zum Maximum ist.
- **Summe.** Die Summe der Werte über alle Fälle mit nichtfehlenden Werten.
- **Kumulierte Summe**. Die kumulierte Summe der Werte. Jedes Grafikelement zeigt die Summe für eine Untergruppe und die Gesamtsumme aller vorherigen Gruppen an.
- **Prozentsumme**. Die Prozentzahl innerhalb jeder Untergruppe basierend auf einem summierten Feld im Vergleich zur Summe über alle Gruppen hinweg.
- **Kumulierte Prozentsumme**. Die kumulative Prozentzahl innerhalb jeder Untergruppe basierend auf einem summierten Feld im Vergleich zur Summe über alle Gruppen hinweg. Jedes Grafikelement zeigt die Prozentzahl für eine Untergruppe und die Gesamtprozentzahl aller vorherigen Gruppen an.
- **Variance.** Ein Maß der Streuung um den Mittelwert. Es ist gleich dem Quotienten aus der Summe der quadrierten Abweichung vom Mittelwert und der um 1 verringerten Fallanzahl. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.

- **Standardabweichung.** Ein Maß für die Streuung um den Mittelwert. In einer Normalverteilung liegen 68% der Fälle innerhalb von einer Standardabweichung des Mittelwerts und 95% der Fälle innerhalb von zwei Standardabweichungen. Wenn beispielsweise für das Alter der Mittelwert 45 und die Standardabweichung 10 beträgt, liegen bei einer Normalverteilung 95 % der Fälle im Bereich zwischen 25 und 65.
- **Standardfehler.** Ein Maß für die Abweichung des Werts einer Teststatistik zwischen Stichproben. Dies ist die Standardabweichung der Stichprobenverteilung einer Statistik. So ist z. B. der Standardfehler des Mittelwerts die Standardabweichung des Stichprobenmittelwerts.
- **Kurtosis.** Ein Maß dafür, wie sich die Beobachtungen um einen zentralen Punkt gruppieren. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremverten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.
- **Schiefe.** Ein Maß für die Asymmetrie einer Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefe-Wert, der mehr als doppelt so groß ist wie sein Standardfehler, für eine Abweichung von der Symmetrie spricht.

Bei folgenden Regionsstatistiken entsteht unter Umständen mehr als ein Grafikelement pro Untergruppe. Bei der Verwendung der Grafikelemente für Intervalle, Bereiche oder Kanten entsteht in einer Regionsstatistik mehr als ein Grafikelement, das den Bereich anzeigt. Alle anderen Grafikelemente erzeugen zwei getrennte Elemente, wobei das eine den Beginn und das andere das Ende des Bereichs anzeigt.

- **Region: Bereich**. Wertebereich zwischen Mindest- und Höchstwert.
- **Region: 95-%-Konfidenzintervall für den Mittelwert**. Ein Wertebereich mit einer 95%igen Wahrscheinlichkeit, die Grundgesamtheit zu enthalten.
- **Region: 95-%-Konfidenzintervall für einzelne Fälle**. Ein Wertebereich mit einer 95%igen Wahrscheinlichkeit, den vorhergesagten Wert angesichts des individuellen Falls zu enthalten.
- **Region: 1 Standardabweichung über/unter dem Mittelwert**. Ein Wertebereich zwischen 1 **Standardabweichung** über und unter dem **Mittelwert**.
- **Region: 1 Standardfehler über/unter dem Mittelwert**. Ein Wertebereich zwischen 1 **Standardfehler** über und unter dem **Mittelwert**.

Anzahlbasierte statistische Funktionen

- **Anzahl**. Die Anzahl der Zeilen/Fälle.
- **Kumulierte Anzahl**. Die kumulierte Anzahl der Zeilen/Fälle. Jedes Grafikelement zeigt die Anzahl für eine Untergruppe und die Gesamtanzahl aller vorherigen Gruppen an.

- **Häufigkeitsprozent**. Die Prozentzahl an Zeilen/Fällen in jeder Untergruppe im Vergleich zur Gesamtzahl an Zeilen/Fällen.
- **Kumulierte Häufigkeitsprozente**. Die kumulierte Prozentzahl an Zeilen/Fällen in jeder Untergruppe im Vergleich zur Gesamtzahl an Zeilen/Fällen. Jedes Grafikelement zeigt die Prozentzahl für eine Untergruppe und die Gesamtprozentzahl aller vorherigen Gruppen an.

Feld- bzw. Variablentypen

In Feldlisten erscheinen neben den Feldern Symbole, die den Feld- und Datentyp anzeigen. Symbole können auch Mehrfachantworten-Sets kennzeichnen.

Messniveau	**Datentyp**			
	Numerisch	Zeichenfolge	Date	Time
Stetig		entfällt		
Sortiertes Set				
Set				

Mehrfachantworten-Set, Set aus kategorialen Variablen	
Mehrfachantworten-Set, Set aus dichotomen Variablen	

Messniveau

Das Messniveau eines Felds ist wichtig, wenn Sie eine Visualisierung erstellen. Im Folgenden finden Sie eine Beschreibung der Messniveaus. Sie können das Messniveau ändern, indem Sie das Feld in der Liste durch Rechtsklicken auswählen und eine Option wählen. In den meisten Fällen müssen Sie nur die beiden breitestgefassten Feldklassifizierungen (kategorial und stetig) berücksichtigen:

Kategorial. Daten mit einer begrenzten Anzahl von eindeutigen Werten bzw. Kategorien (beispielsweise Geschlecht oder Religion). Bei kategorialen Feldern kann es sich um String-Variablen (alphanumerisch) oder um numerische Felder handeln, bei denen numerische Codes zum Darstellen der Kategorien verwendet werden (beispielsweise 0 = *männlich* und 1 = *weiblich*). Auch als qualitative Daten bezeichnet. Sets, sortierte Sets und Flags sind kategoriale Felder.

-
-
-

Stetig. Daten, die auf einer Intervall- oder Verhältnisskala gemessen werden und bei denen die Datenwerte sowohl die Reihenfolge der Werte als auch die Distanz zwischen den Werten festlegen. So ist beispielsweise ein Gehalt von $ 72.195 höher als ein Gehalt von $ 52.398 und die Distanz zwischen den Werten beträgt $ 19.797. Auch als quantitative Daten, Skalendaten oder Daten vom Typ numerischer Bereich bezeichnet.

Kategoriale Felder definieren Kategorien in der Visualisierung, in der Regel zum Zeichnen getrennter Grafikelemente oder zur Gruppierung von Grafikelementen. Stetige Felder werden oft innerhalb von Kategorien kategorialer Felder ausgewertet. So wird beispielsweise in einer Standardvisualisierung, in der das Einkommen nach Geschlecht kategorisiert ist, das durchschnittliche Einkommen von Männern und das durchschnittliche Einkommen von Frauen aufgeführt. Die Rohwerte für stetige Felder können auch in einem Streudiagramm dargestellt werden. So zeigt beispielsweise ein Streudiagramm das aktuelle Gehalt und das Startgehalt für jeden Fall. Ein kategoriales Feld könnte verwendet werden, um die Fälle nach Geschlecht zu gruppieren.

Datentypen

Das Messniveau ist nicht die einzige Feldeigenschaft zur Bestimmung ihres Typs. Ein Feld wird auch als bestimmter Datentyp gespeichert. Mögliche Datentypen sind Zeichenketten (nicht-numerische Daten wie Buchstaben), numerische Werte (reelle Zahlen) und Datumsangaben. Im Gegensatz zum Messniveau kann der Datentyp eines Felds nicht temporär geändert werden. Sie müssen die Speicherart der Daten im Original-Daten-Set ändern.

Mehrfachantworten-Sets

In einigen Datendateien kann außerdem eine besondere Art von “Feld” verwendet werden, die als **Mehrfachantworten-Set** bezeichnet wird. Bei Mehrfachantworten-Sets handelt es sich nicht um “Felder” im üblichen Sinn. Mehrfachantworten-Sets verwenden mehrere Felder, um Antworten auf Fragen aufzuzeichnen, auf welche der Befragte mehr als eine Antwort geben kann. Mehrfachantworten-Sets werden wie kategoriale Felder behandelt und bieten weitestgehend dieselben Möglichkeiten wie kategoriale Felder.

Bei Mehrfachantworten-Sets kann es sich um Sets aus dichotomen Variablen oder um Sets aus kategorialen Variablen handeln.

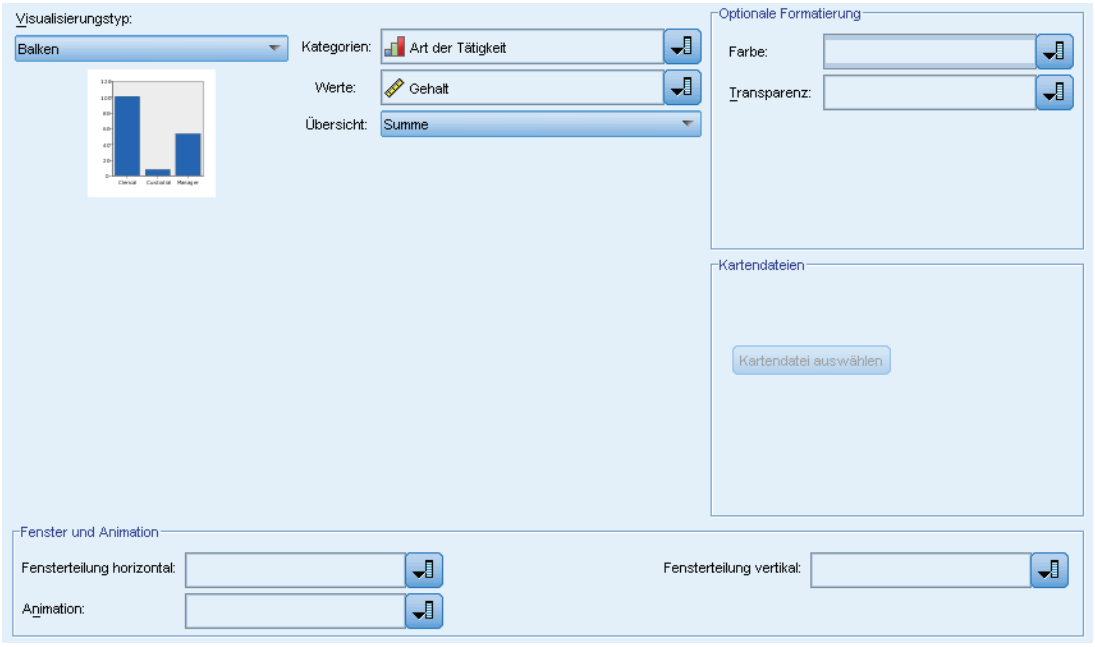
Set aus dichotomen Variablen. Ein Set aus dichotomen Feldern enthält in der Regel mehrere dichotome Felder: Felder mit nur zwei möglichen Werten der Art Ja/Nein, Anwesend/Abwesend, Markiert/Nicht markiert. Auch wenn die Felder nicht unbedingt dichotom sind, werden alle Felder im Set auf die gleiche Weise kodiert.

Beispielsweise gibt eine Umfrage auf die Frage “Welche der folgenden Quellen nutzen Sie für Nachrichten?” fünf mögliche Antworten vor. Der Befragte kann mehrere Antworten wählen, indem er das Kästchen neben jeder Auswahl markiert. Die fünf Antworten werden in der Datendatei zu fünf Felder, die mit 0 für *Nein* (nicht markiert) und 1 für *Ja* (markiert) kodiert werden.

Set aus kategorialen Variablen. Ein Set aus kategorialen Feldern besteht aus mehreren Feldern, die alle auf die gleiche Weise kodiert werden und oft viele mögliche Antwortkategorien enthalten. Man wird beispielsweise in einer Umfrage aufgefordert: "Nennen Sie drei Nationalitäten, die Ihre ethnische Herkunft am besten beschreiben." Hier sind hunderte von Antworten möglich, doch zu Kodierungszwecken ist die Liste auf die 40 häufigsten Nationalitäten beschränkt und alle anderen werden der Kategorie "Andere" zugeordnet. In der Datendatei werden die drei gewählten Nationalitäten zu drei Feldern, wobei jede davon 41 Kategorien enthält (40 kodierte Nationalitäten und eine Kategorie "Andere").

Grafiktafel Registerkarte "Detailliert"

Abbildung 5-9
Registerkarte "Detailliert"



Verwenden Sie die Registerkarte "Detailliert", wenn Sie wissen, welche Art von Visualisierung Sie erstellen möchten, oder wenn Sie optionale Formatierungen, Fenster und/oder eine Animation zu einer Visualisierung hinzufügen möchten. Beispiele finden Sie unter Grafiktafel Beispiele auf S. 270.

- Wenn Sie einen Visualisierungstyp auf der Registerkarte "Einfach" ausgewählt haben, wird dieser angezeigt. Wählen Sie ansonsten einen Typ in der Dropdown-Liste aus. Informationen zu den Visualisierungstypen finden Sie unter Verfügbare integrierte -Grafiktafel-Visualisierungstypen auf S. 262.

- Unmittelbar rechts neben dem Miniaturbild der Visualisierung befinden sich Steuerelemente zur Angabe der für den Visualisierungstyp erforderlichen Felder (Variablen). Sie müssen alle diese Felder angeben.

- Bei bestimmten Visualisierungen können Sie eine statistische Funktion (Übersichtsstatistik) auswählen. In einigen Fällen (z. B. bei Balkendiagrammen) können Sie eine dieser Übersichtsoptionen für die Transparenzformatierung verwenden. Beschreibungen der statistischen Funktionen finden Sie unter Diagrammtafel – Registerkarte “Einfach” auf S. 254.

- Sie können eine oder mehrere der optionalen Formatierungen auswählen. Diese können die Dimensionalität erhöhen, indem sie Ihnen ermöglichen, andere Felder in die Visualisierung aufzunehmen. Beispielsweise können Sie mit einem Feld die Größe der Punkte in einem Streudiagramm variieren. Weitere Informationen zu optionalen Formatierungen finden Sie unter Formatierungen, Überlagerungen, Fenster und Animation auf S. 246. Beachten Sie bitte, dass die Transparenzformatierung nicht über Skripts unterstützt wird.

- Wenn Sie eine Kartenvisualisierung erstellen, enthält die Gruppe Kartendateien die zu verwendenden Kartendateien. Wenn es eine Standardkartendatei gibt, wird sie angezeigt. Klicken Sie zum Ändern der Kartendatei auf Kartendatei auswählen, um das Dialogfeld “Karten auswählen” einzublenden. Sie können die Standardkartendatei auch in diesem Dialogfeld angeben. Für weitere Informationen siehe Thema Auswählen von Kartendateien für Kartenvisualisierungen auf S. 260.

- Sie können eine oder mehrere der Fenstereinteilungs- oder Animationsoptionen auswählen. Weitere Informationen zu Fenstereinteilungs- oder Animationsoptionen finden Sie unter Formatierungen, Überlagerungen, Fenster und Animation auf S. 246.

Auswählen von Kartendateien für Kartenvisualisierungen

Wenn Sie eine Kartenvisualisierungsvorlage auswählen, benötigen Sie eine Kartendatei, die die geografischen Informationen für das Zeichnen der Karte definiert. Wenn es eine Standardkartendatei gibt, wird diese für die Kartenvisualisierung verwendet. Klicken Sie zum Ändern der Kartendatei auf der Registerkarte “Detailliert” auf Kartendatei auswählen, um das Dialogfeld “Karten auswählen” einzublenden.

Das Dialogfeld “Karten auswählen” ermöglicht das Auswählen einer primären und einer Referenzkartendatei. Die Kartendateien definieren die geografischen Informationen für das Zeichnen der Karte. Die Anwendung wird mit einer Gruppe von Standardkartendateien installiert. Wenn es andere ESRI-Shapedateien gibt, die Sie verwenden möchten, müssen Sie die Shapedateien zunächst in SMZ-Dateien umwandeln. Für weitere Informationen siehe Thema Konvertieren und Verteilen von Shapefiles für Karten auf S. 293. Klicken Sie nach Umwandlung der Karte im Dialogfeld “Vorlagenauswahl” auf Verwalten..., um die Karte in das Verwaltungssystem zu laden, damit sie im Dialogfeld “Karten auswählen” zur Verfügung steht.

Es folgen verschiedene Aspekte, die Sie beim Angeben von Kartendateien beachten müssen:

- Für alle Kartenvorlagen ist mindestens eine Kartendatei erforderlich.
- Die Kartendatei verknüpft zumeist ein Kartenschlüsselattribut mit dem Datenschlüssel.

- Wenn die Vorlage keinen Kartenschlüssel benötigt, der zur Verknüpfung mit einem Datenschlüssel dient, sind eine Referenzkartendatei und Felder erforderlich, die Koordinaten (wie geografische Länge und Breite) für Zeichnungselemente der Referenzkarte angeben.
- Überlagerungskartenvorlagen benötigen zwei Karten: eine primäre Kartendatei und eine Referenzkartendatei. Die Referenzkarte wird zuerst gezeichnet, damit sie sich hinter der primären Kartendatei befindet.

Weitere Informationen zur Kartenterminologie, z. B. Attribute und Merkmale, finden Sie unter Wichtige Konzepte im Zusammenhang mit Karten auf S. 294.

Kartendatei. Sie können eine beliebige im Verwaltungssystem vorhandene Kartendatei auswählen. Dazu zählen vorinstallierte und von Ihnen importierte Kartendateien. Weitere Informationen zum Verwalten von Kartendateien finden Sie unter Verwalten von Vorlagen, Stylesheets und Kartendateien auf S. 292.

Kartenschlüssel. Geben Sie das Attribut an, das Sie als Schlüssel zum Verknüpfen der Kartendatei mit dem Datenschlüssel verwenden möchten.

Diese Kartendatei und Einstellungen als Standard speichern. Aktivieren Sie dieses Kontrollkästchen, wenn Sie die ausgewählte Kartendatei als Standard verwenden möchten. Wenn Sie eine Standardkartendatei angegeben haben, müssen Sie nicht bei jeder Erstellung einer Kartenvisualisierung eine Kartendatei angeben.

Datenschlüssel. In diesem Steuerelement wird derselbe Wert wie auf der Registerkarte “Detailliert” im Dialogfeld “Vorlagenauswahl” angezeigt. Der Wert wird hier angezeigt, damit Sie den Schlüssel entsprechend einer bestimmten ausgewählten Kartendatei ändern können.

Alle Kartenmerkmale in der Visualisierung anzeigen. Bei Aktivierung dieser Option werden alle Merkmale in der Karte in der Visualisierung wiedergegeben, auch wenn es keinen übereinstimmenden Datenschlüsselwert gibt. Wenn Sie nur die Merkmale anzeigen möchten, für die Daten vorhanden sind, deaktivieren Sie diese Option. Durch Kartenschlüssel aus der Liste Nicht übereinstimmende Kartenschlüssel angegebene Merkmale werden nicht in der Visualisierung wiedergegeben.

Karten- und Datenwerte vergleichen. Der Karten- und der Datenschlüssel werden verknüpft, um die Kartenvisualisierung zu erstellen. Die Werte dieser beiden Schlüssel müssen übereinstimmen, da andernfalls keine Kartenvisualisierung erstellt werden kann. Klicken Sie auf Vergleichen, um zu prüfen, ob die Datenschlüssel- und Kartenschlüsselwerte übereinstimmen. Das angezeigte Symbol informiert Sie über den Status des Vergleichs. Diese Symbole werden nachfolgend beschrieben. Wenn es nach dem Vergleich Datenschlüsselwerte ohne entsprechende Kartenschlüsselwerte gibt, werden die Datenschlüsselwerte in der Liste Nicht übereinstimmende Datenschlüssel angezeigt. In der Liste Nicht übereinstimmende Kartenschlüssel können Sie außerdem sehen, zu welchen Kartenschlüsselwerten es keine übereinstimmenden Datenschlüsselwerte gibt. Wenn Alle Kartenmerkmale in der Visualisierung anzeigen nicht aktiviert ist, werden durch diese Kartenschlüsselwerte angegebene Merkmale nicht wiedergegeben.

Tabelle 5-1
Vergleichssymbole

Symbol	Beschreibung
	Es ist kein Vergleich erfolgt. Dies ist der Standardstatus, bevor Sie auf Vergleichen klicken. Sie sollten achtsam fortfahren, da Sie nicht wissen, ob die Datenschlüssel- und Kartenschlüsselwerte übereinstimmen.
	Ein Vergleich ist erfolgt und die Datenschlüssel- und Kartenschlüsselwerte stimmen vollständig überein. Für jeden Wert des Datenschlüssels gibt es ein übereinstimmendes vom Kartenschlüssel bestimmtes Merkmal.
	Ein Vergleich ist erfolgt und verschiedene Datenschlüssel- und Kartenschlüsselwerte stimmen nicht überein. Für verschiedene Datenschlüsselwerte gibt es kein übereinstimmendes vom Kartenschlüssel bestimmtes Merkmal. Sie sollten achtsam fortfahren. Falls Sie fortfahren, enthält die Kartenvisualisierung nicht alle Datenwerte.
	Ein Vergleich ist erfolgt und Datenschlüssel- und Kartenschlüsselwerte stimmen nicht überein. Sie müssen einen anderen Daten- oder Kartenschlüssel auswählen, da andernfalls keine Karte wiedergegeben wird.

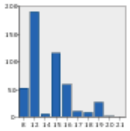
Verfügbare integrierte -Grafiktafel-Visualisierungstypen

Sie können mehrere unterschiedliche Visualisierungstypen erstellen. Alle folgenden integrierten Typen sind auf den Registerkarten “Basis” und “Detailliert” verfügbar. Einige der Beschreibungen für die Vorlagen (insbesondere die Kartenvorlagen) geben die Felder (Variablen) an, die mithilfe von Sondertext auf der Registerkarte “Detailliert” festgelegt wurden.

Tabelle 5-2
Verfügbare Grafiktypen

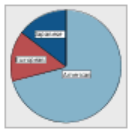
Balken
Berechnet eine Auswertungsstatistik für ein kontinuierliches numerisches Feld und zeigt die Ergebnisse für jede Kategorie eines kategorialen Felds in Form von Balken an.

Voraussetzungen: Ein kategoriales Feld und ein kontinuierliches Feld.



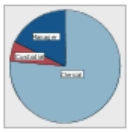
Balken der Häufigkeiten
Zeigt den Anteil der Zeilen/Fälle in jeder Kategorie eines kategorialen Felds als Balken an. Sie können dieses Diagramm auch über den Verteilungs-Diagrammknoten herstellen. Dieser Knoten bietet einige zusätzliche Optionen. Für weitere Informationen siehe Thema Verteilungsknoten auf S. 312.

Voraussetzungen: Ein einzelnes kategoriales Feld.



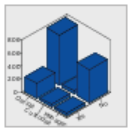
Kreis
Berechnet die Summe eines kontinuierlichen numerischen Felds und zeigt den Anteil dieser Summe in jeder Kategorie eines kategorialen Felds in Form eines Kreissegments an.

Voraussetzungen: Ein kategoriales Feld und ein kontinuierliches Feld.



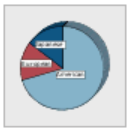
Kreisdiagramm der Häufigkeiten
Zeigt den Anteil der Zeilen/Fälle in jeder Kategorie eines kategorialen Felds als Kreissegmente an.

Voraussetzungen: Ein einzelnes kategoriales Feld.



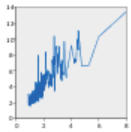
3D-Balken
Berechnet eine Auswertungsstatistik für ein kontinuierliches numerisches Feld und zeigt die Ergebnisse für den Schnittpunkt von Kategorien zwischen zwei kategorialen Feldern an.

Voraussetzungen: Zwei kategoriale Felder und ein kontinuierliches Feld.



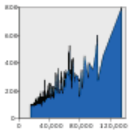
3D-Kreisdiagramm
Mit Ausnahme eines zusätzlichen 3D-Effekts identisch mit dem Kreisdiagramm.

Voraussetzungen: Ein kategoriales Feld und ein kontinuierliches Feld.



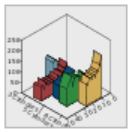
Linie
Berechnet eine Auswertungsstatistik für ein Feld für jeden Wert eines anderen Felds und verbindet die Werte durch eine Linie. Sie können dieses Diagramm auch über den Diagrammknoten “Diagramme” herstellen. Dieser Knoten bietet einige zusätzliche Optionen. Für weitere Informationen siehe Thema Plotknoten auf S. 303.

Voraussetzungen: Zwei Felder beliebigen Typs.



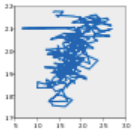
Fläche
Berechnet eine Auswertungsstatistik für ein Feld für jeden Wert eines anderen Felds und verbindet die Werte durch eine Fläche. Der Unterschied zwischen einem Linien- und einem Flächendiagramm ist minimal, da die Fläche durch eine Linie dargestellt wird, unter der der Bereich farbig markiert ist. Wenn Sie jedoch eine Farbformatierung verwenden, wird die Linie einfach getrennt und der Bereich gestapelt.

Voraussetzungen: Zwei Felder beliebigen Typs.



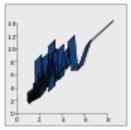
3D-Fläche
Zeigt die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an, indem die Werte durch ein kategoriales Feld getrennt werden. Für jede Kategorie wird ein Flächenelement erstellt.

Voraussetzungen: Ein kategoriales Feld und zwei Felder beliebigen Typs.



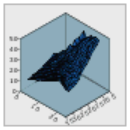
Pfad
Zeigt die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an, indem die Werte in der Reihenfolge, in der sie im ursprünglichen Daten-Set auftreten, mit einer Linie verbunden werden. Die Einhaltung der Reihenfolge ist der wesentliche Unterschied zwischen einem Pfad- und einem Liniendiagramm.

Voraussetzungen: Zwei Felder beliebigen Typs.



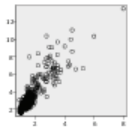
Band
Berechnet eine Auswertungsstatistik für ein Feld für jeden Wert eines anderen Felds und verbindet die Werte durch ein Band. Ein Band ist im Wesentlichen eine Linie mit 3D-Effekten. Es handelt sich nicht um ein echtes 3D-Diagramm.

Voraussetzungen: Zwei Felder beliebigen Typs.



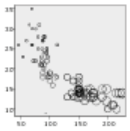
Oberfläche
Zeigt die Werte von drei Feldern im Verhältnis zueinander an, indem die Werte mit einer Oberfläche verbunden werden.

Voraussetzungen: Drei Felder beliebigen Typs.



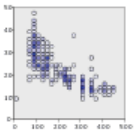
Streudiagramm
Zeigt die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an. Dieses Diagramm kann den Zusammenhang zwischen den Feldern (falls vorhanden) verdeutlichen. Sie können Streudiagramme auch über den Diagrammknoten "Diagramme" herstellen. Dieser Knoten bietet einige zusätzliche Optionen. Für weitere Informationen siehe Thema Plotknoten auf S. 303.

Voraussetzungen: Zwei Felder beliebigen Typs.



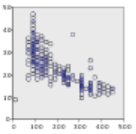
Blasendiagramm
Wie das allgemeine Streudiagramm zeigt das Blasendiagramm die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an. Der Unterschied liegt darin, dass die Werte eines dritten Felds verwendet werden, um die Größe der einzelnen Diagramme zu variieren.

Voraussetzungen: Drei Felder beliebigen Typs.



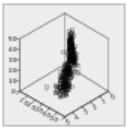
In Klassen unterteiltes Streudiagramm
Wie das allgemeine Streudiagramm zeigt das Blasendiagramm die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an. Der Unterschied liegt darin, dass ähnliche Werte in Klassen zusammengefasst werden und die Farb- und Größenformatierung verwendet wird, um die Anzahl an Fällen in den einzelnen Klassen anzuzeigen.

Voraussetzungen: Zwei kontinuierliche Felder.



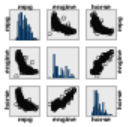
In Hex-Klassen unterteiltes Streudiagramm
Siehe die Beschreibung für das in Klassen unterteilte Streudiagramm. Der Unterschied liegt in der Form der zugrunde liegenden Klassen, die die Form von Sechsecken und nicht von Kreisen aufweisen. Das resultierende, in Hex-Klassen unterteilte Streudiagramm ist dem in Klassen unterteilten Streudiagramm ähnlich. Die Anzahl an Werten in jeder Klasse unterscheidet sich jedoch in den einzelnen Diagrammen aufgrund der zugrunde liegenden Klassen.

Voraussetzungen: Zwei kontinuierliche Felder.



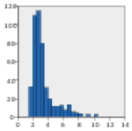
3D-Streudiagramm
Zeigt die Werte von drei Feldern im Verhältnis zueinander an. Dieses Diagramm kann den Zusammenhang zwischen den Feldern (falls vorhanden) verdeutlichen. Sie können 3D-Streudiagramme auch über den Diagrammknoten "Diagramme" herstellen. Dieser Knoten bietet einige zusätzliche Optionen. Für weitere Informationen siehe Thema Plotknoten auf S. 303.

Voraussetzungen: Drei Felder beliebigen Typs.

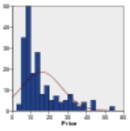


Streudiagramm-Matrix (SPLOM)
Zeigt für jedes Feld die Werte eines Felds im Verhältnis zu den Werten eines anderen Felds an. Eine SPLOM entspricht einer Tabelle von Streudiagrammen. Die SPLOM enthält zudem ein Histogramm für jedes Feld.

Voraussetzungen: Mindestens zwei kontinuierliche Felder.



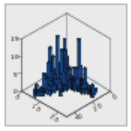
Histogramm
Zeigt die Häufigkeitsverteilung eines Felds an. Mit einem Histogramm können Sie den Verteilungstyp bestimmen und feststellen, ob die Verteilung schief ist. Sie können dieses Diagramm auch über den Diagrammknoten "Histogramm" herstellen. Dieser Knoten bietet einige zusätzliche Optionen. Für weitere Informationen siehe Thema Histogramm – Registerkarte "Plot" auf S. 318.



Histogramm mit Normalverteilung
Zeigt die Häufigkeitsverteilung eines kontinuierlichen Felds mit einer überlagerten Normalverteilungskurve an.

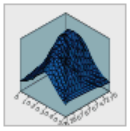
Voraussetzungen: Ein einzelnes kontinuierliches Feld.

Voraussetzungen: Ein einzelnes Feld beliebigen Typs.



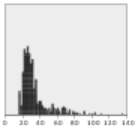
3D-Histogramm
Zeigt die Häufigkeitsverteilung von zwei kontinuierlichen Feldern an.

Voraussetzungen: Zwei kontinuierliche Felder.



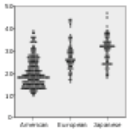
3D-Dichte
Zeigt die Häufigkeitsverteilung von zwei kontinuierlichen Feldern an. Dieses Diagramm ist dem 3D-Histogramm ähnlich. Der einzige Unterschied besteht darin, dass anstatt von Balken die Oberfläche für die Anzeige der Verteilung verwendet wird.

Voraussetzungen: Zwei kontinuierliche Felder.



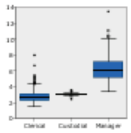
Punktdiagramm
Zeigt die einzelnen Fälle/Zeilen an und stapelt sie am richtigen Datenpunkt auf der *x*-Achse. Dieses Diagramm zeigt wie das Histogramm die Verteilung der Daten an. Der Unterschied besteht darin, dass jeder Fall bzw. jede Zeile und nicht die aggregierten Häufigkeiten für eine bestimmte Klasse (einen Wertebereich) angezeigt werden.

Voraussetzungen: Ein einzelnes Feld beliebigen Typs.



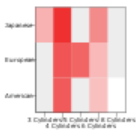
2D-Punktdiagramm
Zeigt die einzelnen Fälle/Zeilen an und stapelt sie für jede Kategorie eines kategorialen Felds am richtigen Datenpunkt auf der *y*-Achse.

Voraussetzungen: Ein kategoriales Feld und ein kontinuierliches Feld.



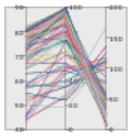
Boxplot
Berechnet die fünf Statistiken (Minimum, erstes Quartil, Median, drittes Quartil und Maximum) für ein kontinuierliches Feld für jede Kategorie eines kategorialen Felds. Die Ergebnisse werden als Boxplot-/Schema-Elemente dargestellt. Mit den Boxplots können Sie feststellen, wie die Verteilung kontinuierlicher Daten innerhalb der Kategorien variiert.

Voraussetzungen: Ein kategoriales Feld und ein kontinuierliches Feld.



Verteilung
Berechnet den Mittelwert für ein kontinuierliches Feld für den Schnittpunkt von Kategorien zwischen zwei kategorialen Feldern.

Voraussetzungen: Zwei kategoriale Felder und ein kontinuierliches Feld.



Parallel
Erstellt parallele Achsen für jedes Feld und zieht für jede Zeile bzw. jeden Fall in den Daten eine Linie durch den Feldwert.

Voraussetzungen: Mindestens zwei kontinuierliche Felder.

Choroplethenkarte der Häufigkeiten
Berechnet die Anzahl für jede Kategorie eines kategorialen Felds (Datenschlüssel) und zeichnet eine Karte, bei der Farbsättigung zur Darstellung der Häufigkeiten in den Kartenstrukturen verwendet wird, die den Kategorien entsprechen.

Voraussetzungen: Ein kategoriales Feld. Eine Karte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt.

Choroplethenkarte der Mittelwerte/Mediane/Summen
Berechnet Mittelwert, Median oder Summe eines kategorialen Felds (Farbe) für jede Kategorie eines kategorialen Felds (Datenschlüssel) und zeichnet eine Karte, bei der Farbsättigung zur Darstellung der berechneten Statistiken in den Kartenstrukturen verwendet wird, die den Kategorien entsprechen.

Voraussetzungen: Ein kategoriales Feld und ein kontinuierliches Feld. Eine Karte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt.

Flächenkartogramm der Werte###
Zeichnet eine Karte, bei der Farbe zur Darstellung der Werte eines kategorialen Felds (Farbe) für die Kartenstrukturen verwendet wird, die Werten entsprechen, die durch ein anderes kategoriales Feld (Datenschlüssel) definiert sind. Wenn mehrere kategoriale Werte des Felds "Farbe" für die einzelnen Strukturen vorliegen, wird der Modalwert verwendet.

Voraussetzungen: Zwei kategoriale Felder. Eine Karte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt.

Koordinaten auf einer Choroplethenkarte der Häufigkeiten
Ähnlich wie "Choroplethenkarte der Häufigkeiten", mit dem Unterschied, dass zwei weitere kontinuierliche Felder (Längengrad und Breitengrad) vorhanden sind, die Koordinaten zum Zeichnen von Punkten auf der Choroplethenkarte angeben.

Voraussetzungen: Ein kategoriales Feld und zwei kontinuierliche Felder. Eine Karte, deren Schlüssel mit den

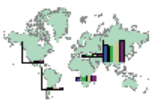
Koordinaten auf einer Choroplethenkarte der Mittelwerte/Mediane/Summen
Ähnlich wie "Choroplethenkarte der Mittelwerte/Mediane/Summen", mit dem Unterschied, dass zwei weitere kontinuierliche Felder (Längengrad und Breitengrad) vorhanden sind, die Koordinaten zum Zeichnen von Punkten auf der Choroplethenkarte angeben.

Voraussetzungen: Ein kategoriales Feld und drei kontinuierliche Felder. Eine Karte, deren Schlüssel mit

Datenschlüssel-Kategorien übereinstimmt.

Koordinaten auf einem Flächenkartogramm der Werte###
Ähnlich wie "Flächenkartogramm der Werte###", mit dem Unterschied, dass zwei weitere kontinuierliche Felder (Längengrad und Breitengrad) vorhanden sind, die Koordinaten zum Zeichnen von Punkten auf der Choroplethenkarte angeben.

Voraussetzungen: Zwei kategoriale Felder und zwei kontinuierliche Felder. Eine Karte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt.



Balken auf einer Karte
Berechnet eine Auswertungsstatistik für ein kontinuierliches Feld (Werte) und zeigt die Ergebnisse für jede Kategorie eines kategorialen Felds (Kategorien) für jede Kartenstruktur (Datenschlüssel) als Balkendiagramme in der Mitte der einzelnen Kartenstrukturen an.

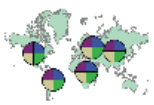
Voraussetzungen: Zwei kategoriale Felder und ein kontinuierliches Feld. Eine Karte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt.

Kreisdiagramm auf einer Karte
Berechnet die Summe eines kontinuierlichen Felds (Werte) in den einzelnen Kategorien eines kategorialen Felds (Kategorien) für jede Kartenstruktur (Datenschlüssel) und zeichnet eine Karte sowie die Summen als Segmente eines Kreisdiagramms in der Mitte der einzelnen Kartenstrukturen.

den Datenschlüssel-Kategorien übereinstimmt.

Balken mit Zählerwerten auf einer Karte
Berechnet den Anteil an Zeilen/Fällen in den einzelnen Kategorien eines kategorialen Felds (Kategorien) für jede Kartenstruktur (Datenschlüssel) und zeichnet eine Karte und die Balkendiagramme in der Mitte der einzelnen Kartenstrukturen.

Voraussetzungen: Zwei kategoriale Felder. Eine Karte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt.



Kreisdiagramm mit Zählerwerten auf einer Karte
Zeigt den Anteil an Zeilen/Fällen in den einzelnen Kategorien eines kategorialen Felds (Kategorien) für jede Kartenstruktur (Datenschlüssel) an und zeichnet eine Karte sowie die Anteile als Segmente eines Kreisdiagramms in der Mitte der einzelnen Kartenstrukturen.

Voraussetzungen: Zwei kategoriale Felder. Eine Karte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt.

Liniendiagramm auf einer Karte
Berechnet eine Auswertungsstatistik für ein kontinuierliches Feld (Werte) für jeden Wert eines anderen Felds (X) für jede Kartenstruktur (Datenschlüssel) und zeichnet eine Karte sowie die Liniendiagramme, die die Werte verbinden, in der Mitte der einzelnen Kartenstrukturen.

Voraussetzungen: Zwei kategoriale Felder und ein kontinuierliches Feld. Eine Karte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt.

Koordinaten auf einer Bezugskarte
Zeichnet eine Karte und Punkte mithilfe kontinuierlicher Felder (Längengrad und Breitengrad), die Koordinaten für die Punkte angeben.

Voraussetzungen: Zwei Bereichsfelder. Eine Kartendatei.

Pfeile auf einer Bezugskarte
Zeichnet eine Karte und Pfeile mithilfe kontinuierlicher Felder, die die Startpunkte (Start Längengrad und Start Breitengrad) und Endpunkte (Ende Längengrad und Ende Breitengrad) für die einzelnen Pfeile angeben. Jeder Datensatz/Fall in den Daten führt zu einem Pfeil auf der Karte.

Voraussetzungen: Vier kontinuierliche Felder. Eine Kartendatei.

Punktüberlagerungskarte
Zeichnet eine Referenzkarte und überlagert diese mit einer weiteren Punktkarte, wobei die Farbe der Punktstrukturen durch ein kategoriales Feld (Farbe) festgelegt ist.

Voraussetzungen: Zwei kategoriale Felder. Eine Punktkarte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt. Eine Referenzkartendatei.

Polygonüberlagerungskarte
Zeichnet eine Referenzkarte und überlagert diese mit einer weiteren Polygonkarte, wobei die Farbe der Polygonstrukturen durch ein kategoriales Feld (Farbe) festgelegt ist.

Voraussetzungen: Zwei kategoriale Felder. Eine Polygonkarte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt. Eine Referenzkartendatei.

Linienüberlagerungskarte
Zeichnet eine Referenzkarte und überlagert diese mit einer weiteren Linienkarte, wobei die Farbe der Linienstrukturen durch ein kategoriales Feld (Farbe) festgelegt ist.

Voraussetzungen: Zwei kategoriale Felder. Eine Linienkarte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt. Eine Referenzkartendatei.

Voraussetzungen: Ein kategoriales Feld und zwei Felder beliebigen Typs. Eine Karte, deren Schlüssel mit den Datenschlüssel-Kategorien übereinstimmt.

Erstellen von Kartenvisualisierungen

Bei vielen Visualisierungen müssen nur zwei Elemente ausgewählt werden: die relevanten Felder (Variablen) und eine Vorlage zur Visualisierung dieser Felder. Es sind keine weiteren Entscheidungen oder Aktionen erforderlich. Bei Kartenvisualisierungen ist mindesten ein weiterer Schritt nötig: Auswahl einer Kartendatei, die die geografischen Informationen für die Kartenvisualisierung definiert.

Die Grundschritte zur Erstellung einer einfachen Karte lauten wie folgt:

- Wählen Sie die relevanten Felder auf der Registerkarte “Einfach” aus. Informationen dazu, welcher Typ und welche Anzahl von Feldern für verschiedene Kartenvisualisierungen erforderlich sind, finden Sie unter Verfügbare integrierte -Grafiktafel-Visualisierungstypen auf S. 262.
- Wählen Sie eine Kartenvorlage aus.
- Klicken Sie auf die Registerkarte “Detailliert”.
- Vergewissern Sie sich, dass Datenschlüssel und die anderen erforderlichen Dropdown-Listen auf die richtigen Felder gesetzt sind.
- Klicken Sie in der Gruppe “Kartendateien” auf Kartendatei auswählen.
- Wählen Sie im Dialogfeld “Karten auswählen” die Kartendatei und den Kartenschlüssel aus. Die Werte des Kartenschlüssels müssen mit den unter Datenschlüssel für das Feld angegebenen Werten übereinstimmen. Mit der Schaltfläche Vergleichen können diese Werte verglichen werden. Wenn Sie eine Überlagerungskartenvorlage auswählen, müssen Sie auch eine Referenzkarte auswählen. Die Referenzkarte ist nicht mit den Daten verknüpft. Sie wird als Hintergrund für die Hauptkarte verwendet. Weitere Informationen zum Dialogfeld “Karten auswählen” finden Sie unter Auswählen von Kartendateien für Kartenvisualisierungen auf S. 260.
- Klicken Sie auf OK, um das Dialogfeld “Karten auswählen” zu schließen.
- Klicken Sie in der Grafiktafel-Vorlagenauswahl auf Ausführen, um die Kartenvisualisierung zu erstellen.

Grafiktafel Beispiele

Dieser Abschnitt enthält einige unterschiedliche Beispiele zur Veranschaulichung der verfügbaren Optionen. Die Beispiele liefern zudem Informationen für die Interpretation der resultierenden Visualisierungen.

In diesen Beispielen wird der Stream *graphboard.str* verwendet, der auf die Datendateien *employee_data.sav*, *customer_subset.sav* und *worldsales.sav* verweist. Diese Dateien finden Sie im Ordner *Demos* jeder IBM® SPSS® Modeler Client-Installation. Der Zugriff über die Programmgruppe “SPSS Modeler” ist im Start-Menü von Windows möglich. Die Datei *graphboard.str* befindet sich im Ordner *streams*.

Es wird empfohlen, die Beispiele in der vorgegebenen Reihenfolge zu lesen. Die nachfolgenden Beispiele bauen auf den vorherigen Beispielen auf.

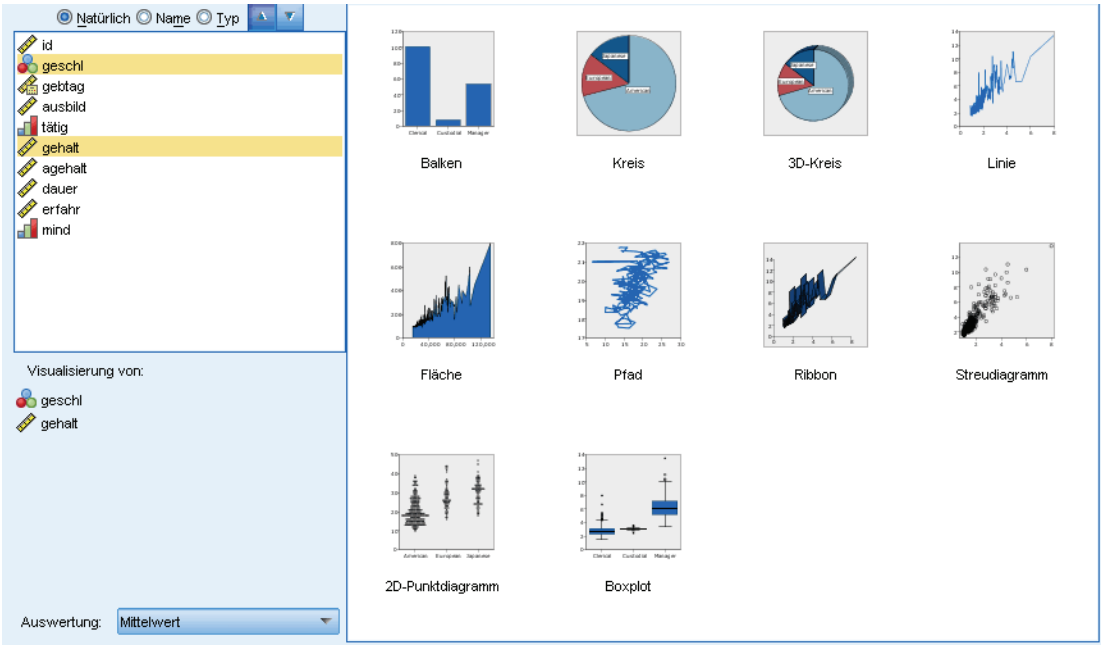
Beispiel: Balkendiagramm mit Auswertungsstatistik

Wir erstellen ein Balkendiagramm, das ein kontinuierliches numerisches Feld bzw. eine kontinuierliche numerische Variable für jede Kategorie eines Sets bzw. einer kategorialen Variable zusammenfasst. Insbesondere erstellen wir ein Balkendiagramm, das das mittlere Gehalt für Männder und Frauen darstellt.

In diesem und einigen folgenden Beispielen wird die Datei *Employee data* verwendet, bei der es sich um ein hypothetisches Daten-Set mit Informationen über die Mitarbeiter eines Unternehmens handelt.

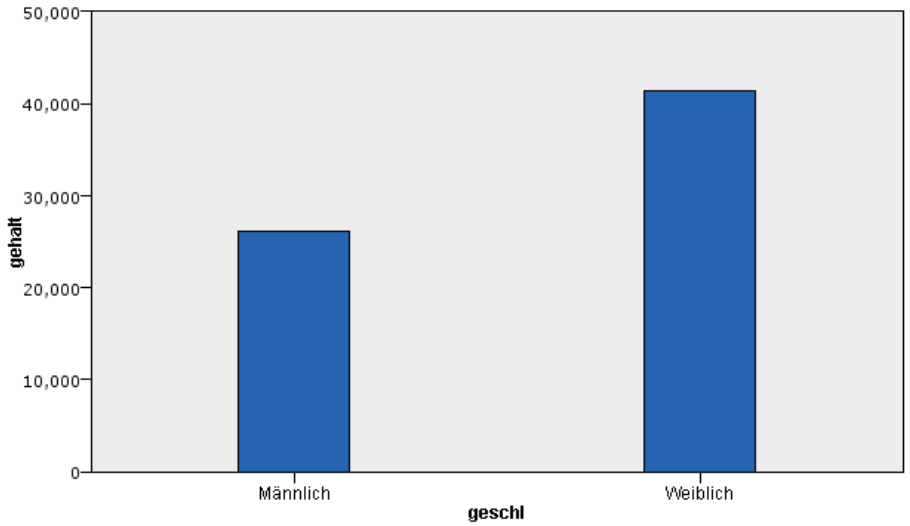
- Fügen Sie einen Statistics-Quellknoten hinzu, der auf *employee_data.sav* verweist.
- Fügen Sie einen Grafiktafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
- Wählen Sie auf der Registerkarte "Basis" die Optionen *Gender* (Geschlecht) und *Current Salary* (Aktuelles Gehalt). (Wenn Sie bei gedrückter Strg-Taste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)
- Wählen Sie Balken.
- Wählen Sie aus der Dropdown-Liste "Auswertung" den Eintrag Mittelwert aus.

Abbildung 5-11
Auswahl auf der Registerkarte "Basis", Balkendiagramm mit Auswertungsstatistik



- Klicken Sie auf Ausführen.
- Klicken Sie in der eingeblendeten Anzeige auf die Symbolleistenschaltfläche "Feld- und Wertelabels anzeigen" (die zweite in der Zweiergruppe in der Mitte der Symbolleiste).

Abbildung 5-12
Balkendiagramm mit Auswertungsstatistik



Wir stellen Folgendes fest:

- Basierend auf der Höhe der Balken ist klar, dass das mittlere Gehalt von Männern über dem von Frauen liegt.

Beispiel: Gruppiertes Balkendiagramm mit Auswertungsstatistik

Wir erstellen nun ein gruppiertes Balkendiagramm, um festzustellen, ob der Unterschied im mittleren Gehalt zwischen Männern und Frauen von der Art der Tätigkeit abhängig ist. Vielleicht arbeiten Frauen im Durchschnitt mehr als Männer in bestimmten Tätigkeitsarten.

Hinweis: In diesem Beispiel wird die Datendatei *Employee data* verwendet.

▶ Fügen Sie einen Grafiktafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.

▶ Wählen Sie auf der Registerkarte “Basis” die Optionen *Employment Category* (Art der Tätigkeit) und *Current Salary* (Aktuelles Gehalt). (Wenn Sie bei gedrückter Strg-Taste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)

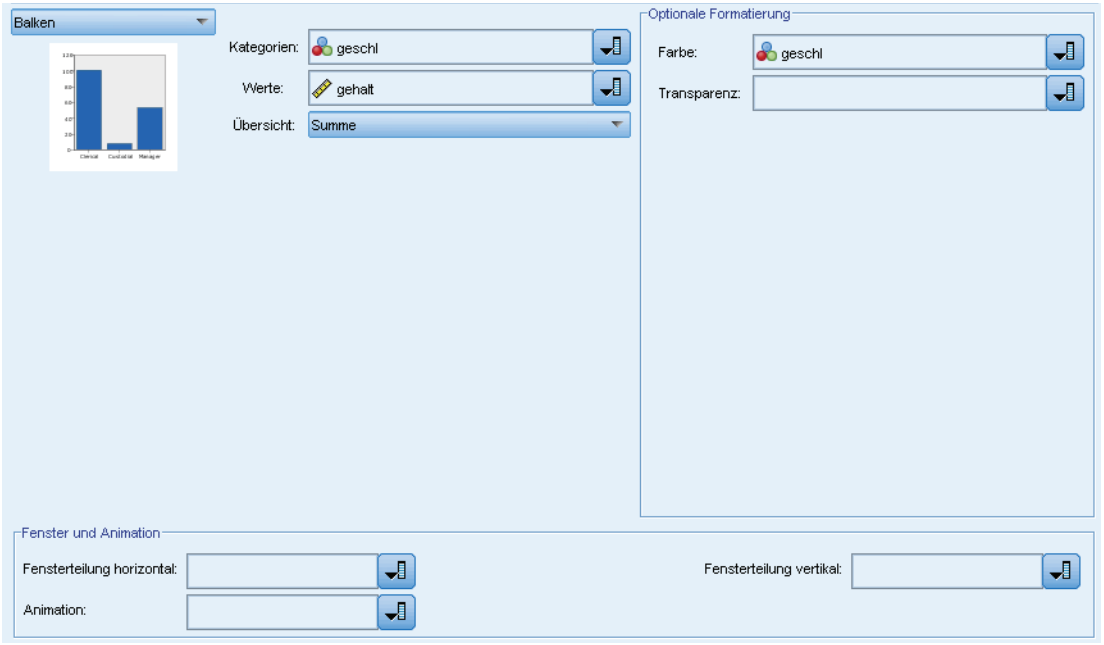
▶ Wählen Sie Balken.

▶ Wählen Sie aus der Liste “Auswertung” den Eintrag Mittelwert aus.

▶ Klicken Sie auf die Registerkarte “Detailliert”. Beachten Sie, dass Ihre Auswahl auf der vorherigen Registerkarte hier berücksichtigt wird.

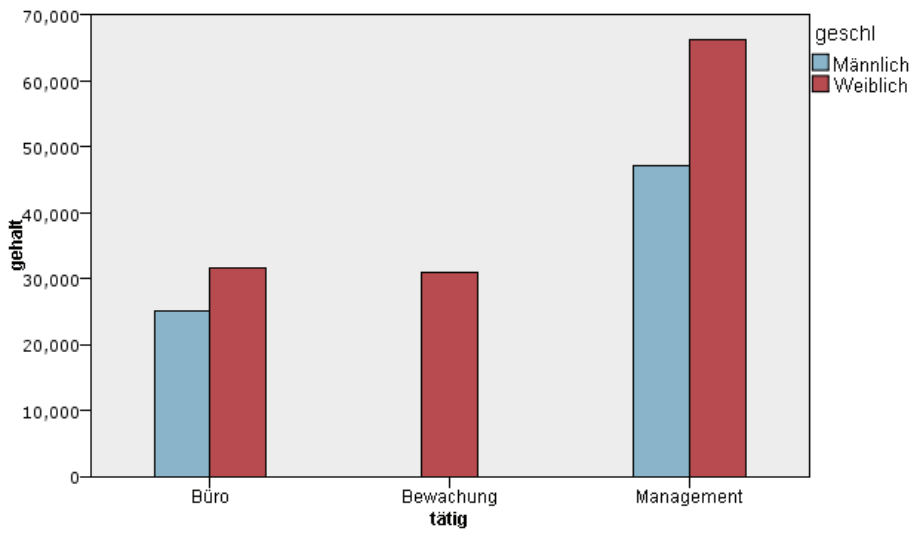
▶ Wählen Sie in der Gruppe “Optionale Formatierungen” die Option *Geschlecht* aus der Dropdown-Liste “Farbe” aus.

Abbildung 5-13
Auswahl auf der Registerkarte "Detailliert", gruppiertes Balkendiagramm



▶ Klicken Sie auf Ausführen.

Abbildung 5-14
Gruppiertes Balkendiagramm



Wir stellen Folgendes fest:

- Der Unterschied im mittleren Gehalt für die einzelnen Arten der Tätigkeiten scheint nicht genauso groß zu sein wie im Balkendiagramm, in dem das mittlere Gehalt für alle Männer und Frauen verglichen wurde. Vielleicht gibt es eine unterschiedliche Anzahl an

Männern und Frauen in den einzelnen Gruppen. Das könnten wir überprüfen, indem wir ein Balkendiagramm der Häufigkeiten erstellen.

- Unabhängig von der Art der Tätigkeit ist das mittlere Gehalt für Männer immer größer als das der Frauen.

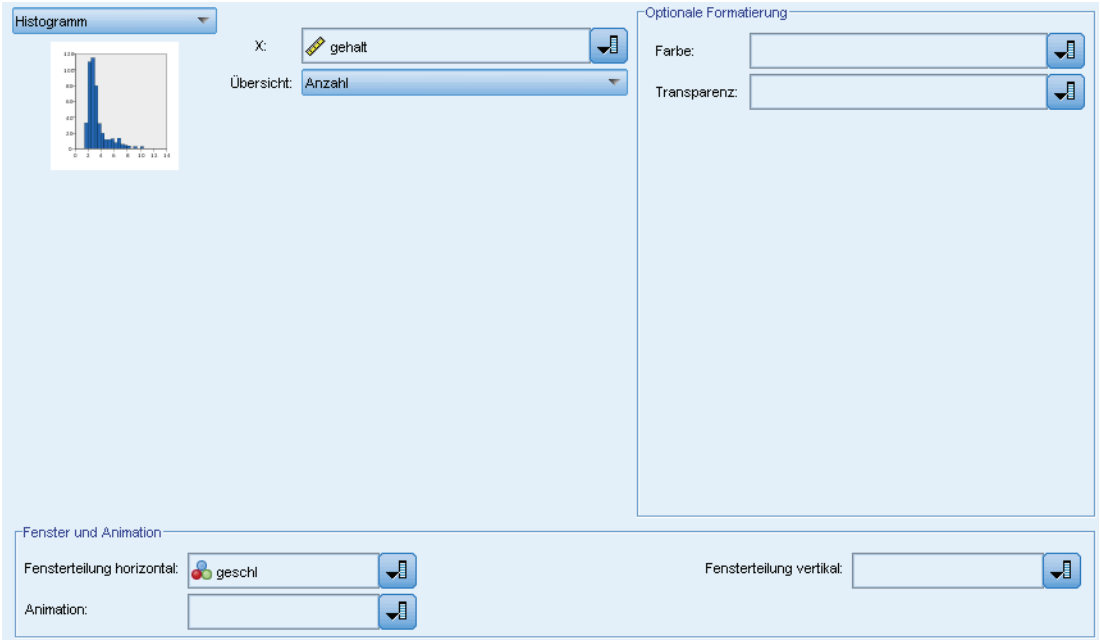
Beispiel: Unterteiltes Histogramm

Wir erstellen ein Histogramm, das nach Geschlechtern unterteilt ist, um die Häufigkeitsverteilung der Gehälter von Männern und Frauen vergleichen zu können. Die Häufigkeitsverteilung zeigt, wie viele Fälle bzw. Zeilen innerhalb eines bestimmten Gehaltsbereichs liegen. Mit dem unterteilten Histogramm können wir den Unterschied bei den Gehältern von Männern und Frauen genauer analysieren.

Hinweis: In diesem Beispiel wird die Datendatei *Employee data* verwendet.

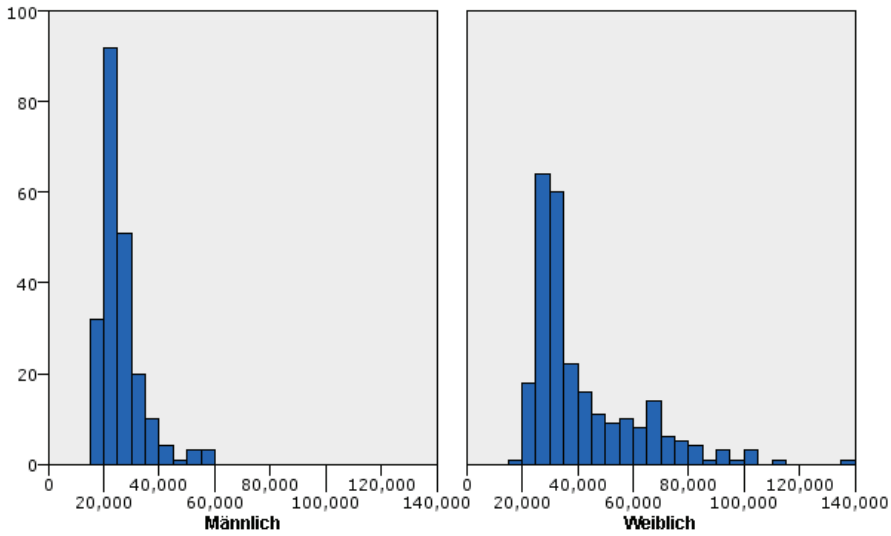
- Fügen Sie einen Grafiktafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
- Wählen Sie auf der Registerkarte "Basis" die Option *Current Salary* (Aktuelles Gehalt).
- Wählen Sie Histogramm.
- Klicken Sie auf die Registerkarte "Detailliert".
- Wählen Sie in der Gruppe "Aufteilungen und Animation" die Option *gender* (Geschlecht) aus der Dropdown-Liste "Aufteilen nach" aus.

Abbildung 5-15
Auswahl auf der Registerkarte "Detailliert", unterteiltes Histogramm



- Klicken Sie auf Ausführen.

Abbildung 5-16
Unterteiltes Histogramm



Wir stellen Folgendes fest:

- Keine der beiden Häufigkeitsverteilungen ist eine Normalverteilung. Das heißt, die Histogramme stellen keine Glockenkurve dar, wie es bei einer Normalverteilung der Fall wäre.
- Die höheren Balken befinden sich auf der linken Seite des Diagramms. Das bedeutet, dass sowohl mehr Männer als auch mehr Frauen geringere als höhere Gehälter haben.
- Die Häufigkeitsverteilungen der Gehälter der Männer und der Frauen sind nicht identisch. Beachten Sie die Form der Histogramme. Es gibt mehr Männer, die höhere Gehälter erhalten, als Frauen, die höhere Gehälter erhalten.

Beispiel: Unterteiltes Punktdiagramm

Wie ein Histogramm zeigt auch ein Punktdiagramm die Verteilung eines kontinuierlichen numerischen Bereichs an. Im Gegensatz zu Histogrammen, die die Häufigkeiten für klassierte Datenbereiche darstellen, zeigen Punktdiagramme jede Zeile bzw. jeden Fall in den Daten an. Daher bietet ein Punktdiagramm im Vergleich zu Histogrammen eine höhere Granularität. Für die Analyse von Häufigkeitsverteilungen könnte ein Punktdiagramm sogar der geeignetere Ausgangspunkt sein.

Hinweis: In diesem Beispiel wird die Datendatei *Employee data* verwendet.

▶ Fügen Sie einen Grafiktafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.

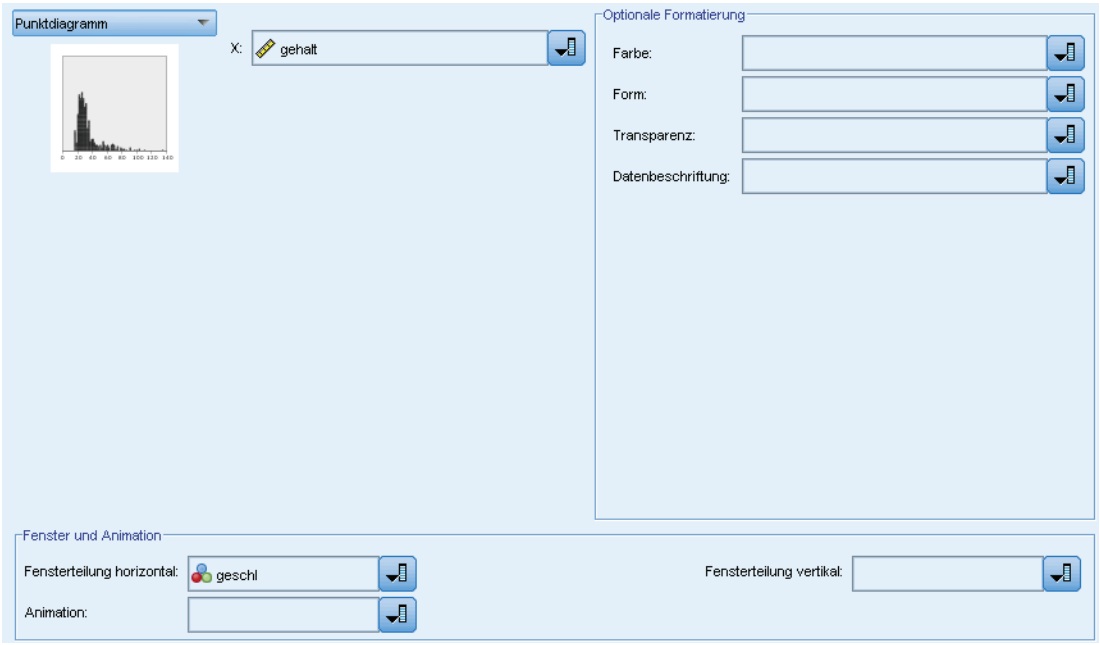
▶ Wählen Sie auf der Registerkarte “Basis” die Option *Current Salary* (Aktuelles Gehalt).

▶ Wählen Sie die Option Punktdiagramm.

▶ Klicken Sie auf die Registerkarte “Detailliert”.

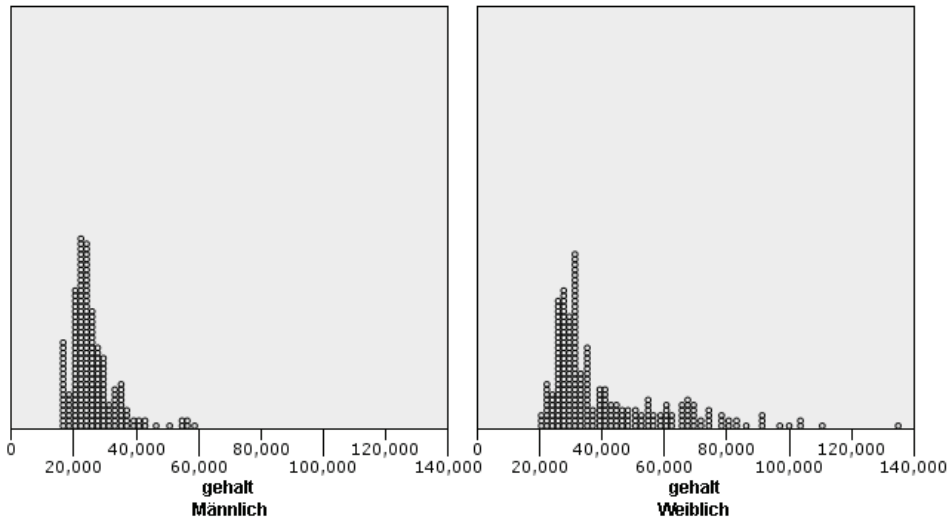
- Wählen Sie in der Gruppe "Aufteilungen und Animation" die Option *gender* (Geschlecht) aus der Dropdown-Liste "Aufteilen nach" aus.

Abbildung 5-17
Auswahl auf der Registerkarte "Detailliert", unterteiltes Punktdiagramm



- Klicken Sie auf Ausführen.
- Maximieren Sie das angezeigte Ausgabefenster, um das Diagramm besser sehen zu können.

Abbildung 5-18
Unterteiltes Punktdiagramm



Im Vergleich zum Histogramm (siehe Beispiel: Unterteiltes Histogramm auf S. 274) stellen wir Folgendes fest:

- Die Spitze bei 20.000, die sich im Histogramm für Frauen ergab, ist im Punktdiagramm weniger deutlich ausgeprägt. Es sind viele Fälle und Zeilen um diesen Wert konzentriert, die meisten von ihnen liegen jedoch näher an 25.000. Dieses Granularitätsniveau ist im Histogramm nicht ersichtlich.
- Obwohl das Histogramm für Männer darauf hindeutet, dass das mittlere Gehalt für Männer nach 40.000 gleichmäßig abnimmt, zeigt das Punktdiagramm, dass die Verteilung nach diesem Wert bis 80.000 relativ einheitlich ist. Bei jedem Gehaltswert in diesem Bereich gibt es drei oder mehr Männer, die dieses Gehalt beziehen.

Beispiel: Boxplot

Ein Boxplot ist eine weitere sinnvolle Visualisierung, um darzustellen, wie die Daten verteilt sind. Ein Boxplot enthält mehrere statistische Messgrößen, die wir nach der Erstellung der Visualisierung kennenlernen werden.

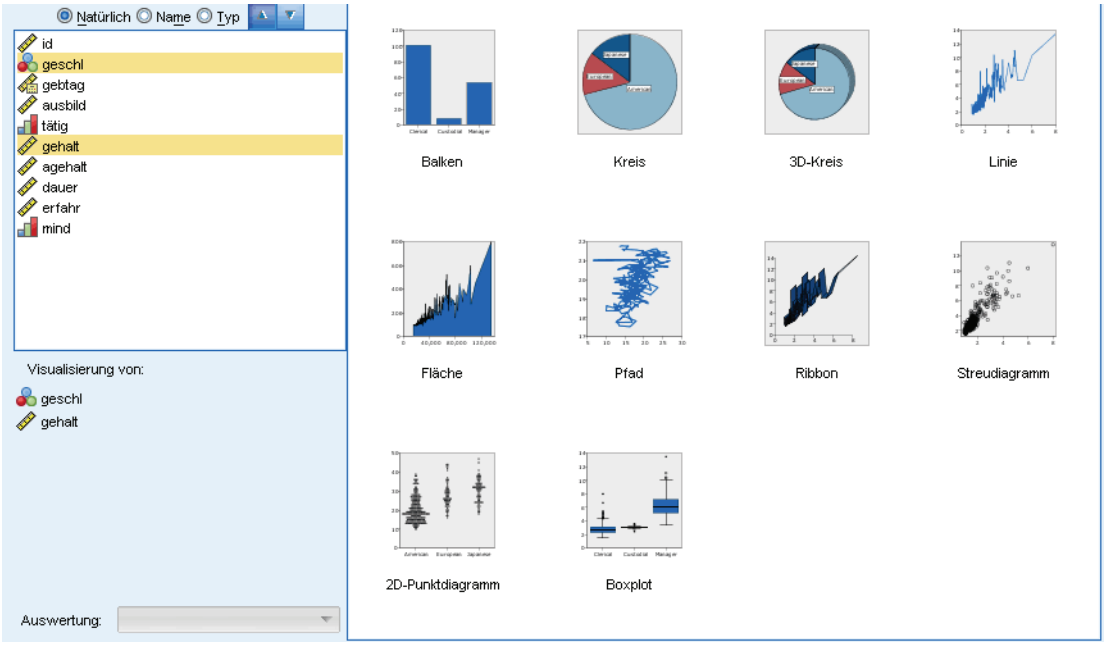
Hinweis: In diesem Beispiel wird die Datendatei *Employee data* verwendet.

▶ Fügen Sie einen Grafiktafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.

▶ Wählen Sie auf der Registerkarte "Basis" die Optionen *Gender* (Geschlecht) und *Current Salary* (Aktuelles Gehalt). (Wenn Sie bei gedrückter Strg-Taste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)

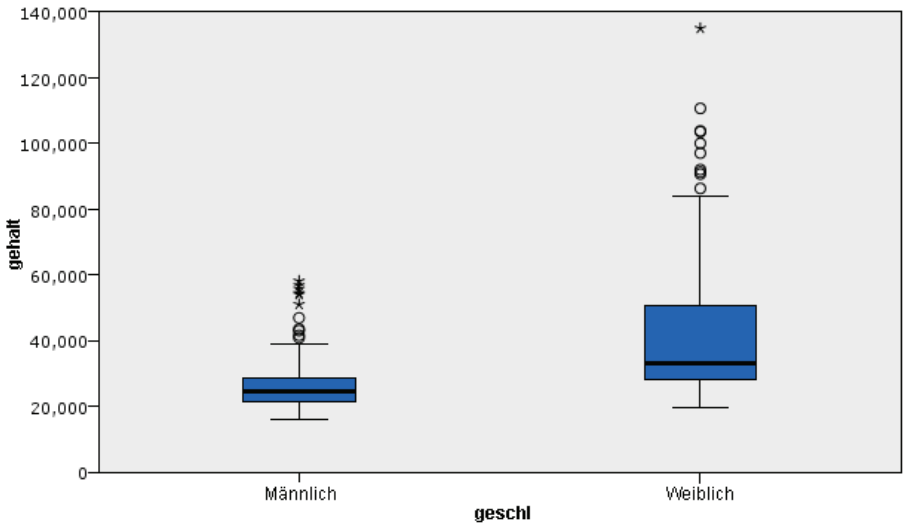
▶ Wählen Sie Boxplot.

Abbildung 5-19
Auswahl auf der Registerkarte "Basis", Boxplot



▶ Klicken Sie auf Ausführen.

Abbildung 5-20
Boxplot



Machen wir uns zunächst mit den einzelnen Bereichen des Boxplots vertraut:

- Die dunkle Linie in der Mitte der Boxen ist der Median des Gehalts (*salary*), Die Hälfte der Fälle bzw. Zeilen besitzt einen höheren Wert als der Median und die andere Hälfte einen geringeren Wert. Wie der Mittelwert ist der Median eine Messgröße für Lagemaße. Im Gegensatz zum Mittelwert haben Fälle bzw. Zeilen mit Extremwerten weniger Einfluss auf den Median. In diesem Beispiel ist der Median kleiner als der Mittelwert (siehe Beispiel: Balkendiagramm mit Auswertungsstatistik auf S. 271). Der Unterschied zwischen dem Mittelwert und dem Median deutet an, dass einige Fälle bzw. Zeilen mit Extremwerten den Mittelwert anheben. Das heißt, es gibt ein paar Angestellte, die große Gehälter beziehen.
- Im unteren Bereich der Box wird das 25. Perzentil dargestellt. 25 Prozent der Fälle/Zeilen haben Werte unter dem 25. Perzentil. Im oberen Bereich der Box wird das 75. Perzentil dargestellt. 25 Prozent der Fälle/Zeilen haben Werte über dem 75. Perzentil. Das bedeutet, dass 50 % der Fälle/Zeilen innerhalb der Box liegen. Die Box ist für Frauen wesentlich kürzer als für Männer. Das deutet darauf hin, dass das Gehalt (*salary*) bei Frauen weniger variiert als bei Männern. Der obere und untere Bereich der Box werden häufig als **Hinges** bezeichnet.
- Die T-Balken, die von den Boxen ausgehen, werden als **Fühler** oder **Whisker** bezeichnet. Die Länge beträgt das 1,5-Fache der Höhe der Box oder falls keine Fälle bzw. Zeilen mit Werten in diesem Bereich vorhanden sind, wird die Länge durch den maximalen bzw. minimalen Wert festgelegt. Bei einer Normalverteilung der Daten wird erwartet, dass circa 95 % der Daten innerhalb der Fühler liegen. In diesem Beispiel sind die Fühler bei Frauen kürzer als bei den Männern. Auch das deutet darauf hin, dass das Gehalt (*salary*) bei Frauen weniger variiert als bei Männern.
- Die Punkte sind **Ausreißer**. Ausreißer sind Werte, die nicht innerhalb der Fühler liegen. Ausreißer sind Extremwerte. Die Sternchen sind **extreme Ausreißer**. Das sind all jene Fälle/Zeilen, deren Werte mehr als dreimal so groß sind wie die Höhe der Boxen. Es sind mehrere Ausreißer bei Frauen und Männern vorhanden. Berücksichtigen Sie, dass der

Mittelwert größer als der Median ist. Der höhere Mittelwert wird von diesen Ausreißern verursacht.

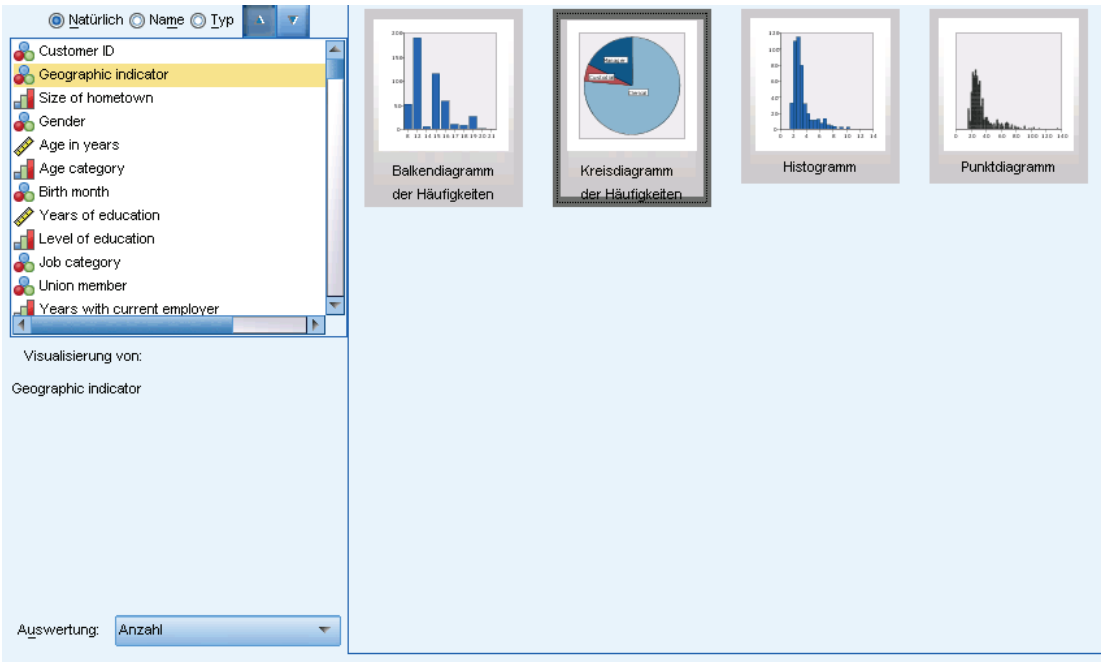
Beispiel: Kreisdiagramm

Wir verwenden nun ein anderes Daten-Set, um andere Visualisierungstypen kennenzulernen. Das Daten-Set *customer_subset* ist eine hypothetische Datendatei mit Informationen über Kunden.

Zunächst erstellen wir ein Kreisdiagramm, um zu ermitteln, welche Anteile der Kunden in verschiedenen geografischen Regionen zu finden sind.

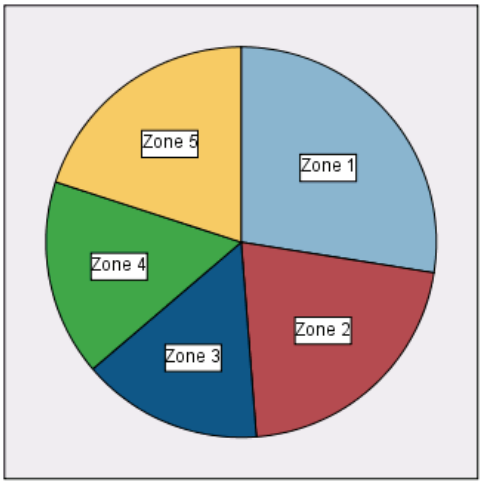
- Fügen Sie einen Statistics-Quellknoten hinzu, der auf *customer_subset.sav* verweist.
- Fügen Sie einen Grafiktafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
- Wählen Sie auf der Registerkarte "Basis" die Variable *Geographic indicator* (Geografischer Indikator).
- Wählen Sie Kreisdiagramm der Häufigkeiten.

Abbildung 5-21
Auswahl auf der Registerkarte "Basis", Kreisdiagramm



- Klicken Sie auf Ausführen.

Abbildung 5-22
Kreisdiagramm



Wir stellen Folgendes fest:

- In Zone 1 leben mehr Kunden als in allen anderen Zonen.
- Die Kunden sind gleichmäßig auf die anderen Zonen verteilt.

Beispiel: Verteilung

Wir erstellen nun eine kategoriale Verteilung, um das mittlere Einkommen für Kunden in unterschiedlichen geografischen Regionen und Altersgruppen zu überprüfen.

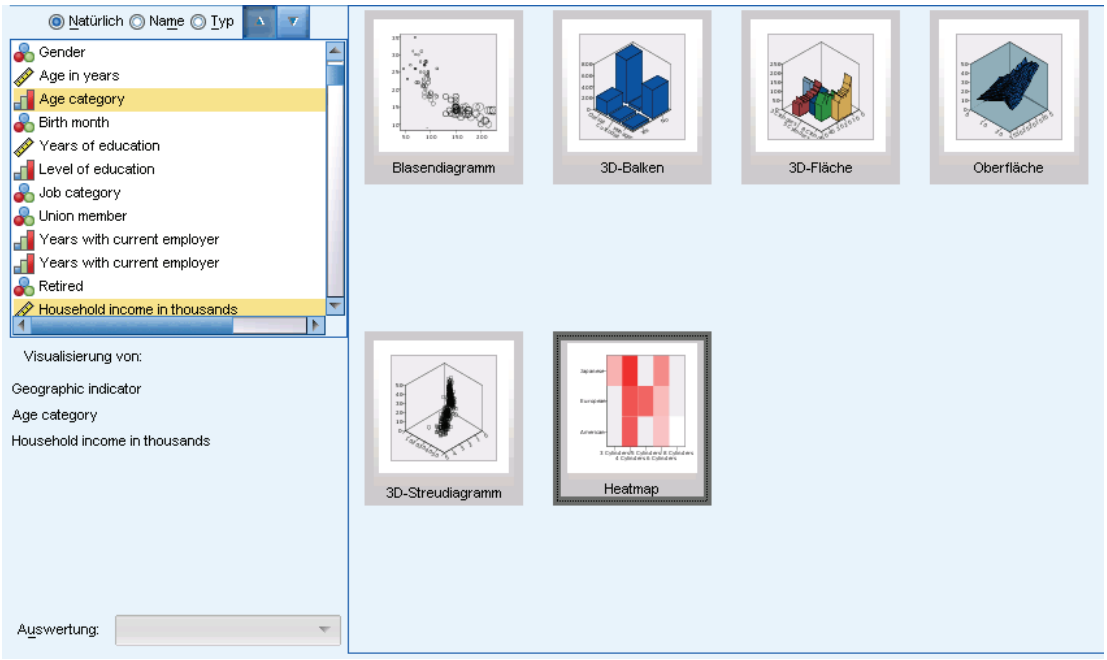
Hinweis: Für dieses Beispiel wird die Datei *customer_subset* verwendet.

▶ Fügen Sie einen Grafiktafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.

▶ Wählen Sie auf der Registerkarte “Basis” die Optionen *Geographic indicator* (Geografischer indikator), *Age category* (Alterskategorie) und *Household income in thousands* (Haushaltseinkommen in Tausend), aus, und zwar in dieser Reihenfolge. (Wenn Sie bei gedrückter Strg-Taste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)

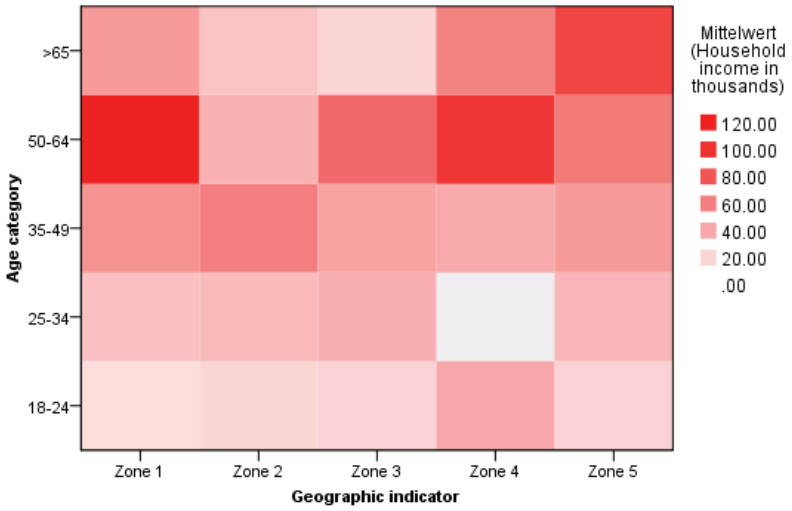
▶ Wählen Sie Verteilung.

Abbildung 5-23
Auswahl auf der Registerkarte "Basis", Verteilung



- Klicken Sie auf Ausführen.
- Klicken Sie im angezeigten Ausgabefenster, auf die Symbolleistenschaltfläche "Feld- und Wertelabels anzeigen" (die rechte der beiden Schaltflächen in der Mitte der Symbolleiste).

Abbildung 5-24
Kategoriale Verteilung



Wir stellen Folgendes fest:

- Eine Verteilung verhält sich wie eine Tabelle, in der anstelle von Zahlen Farben verwendet werden, um die Werte der Zellen darzustellen. Ein kräftiges Rot steht für den höchsten Wert, während Grau den niedrigsten Wert darstellt. Der Wert der einzelnen Zellen ist der Mittelwert des stetigen Felds bzw. der stetigen Variablen für jedes Kategorienpaar.
- Mit Ausnahme der Zonen 2 und 5 hat die Kundengruppe, deren Alter zwischen 50 und 64 liegt, ein höheres mittleres Haushaltseinkommen als die anderen Gruppen.
- In Zone 4 gibt es keine Kunden im Alter von 25 bis 34.

Beispiel: Streudiagramm-Matrix (SPLOM)

Wir erstellen eine Streudiagramm-Matrix aus mehreren unterschiedlichen Variablen, um feststellen zu können, ob Zusammenhänge zwischen den Variablen im Daten-Set bestehen.

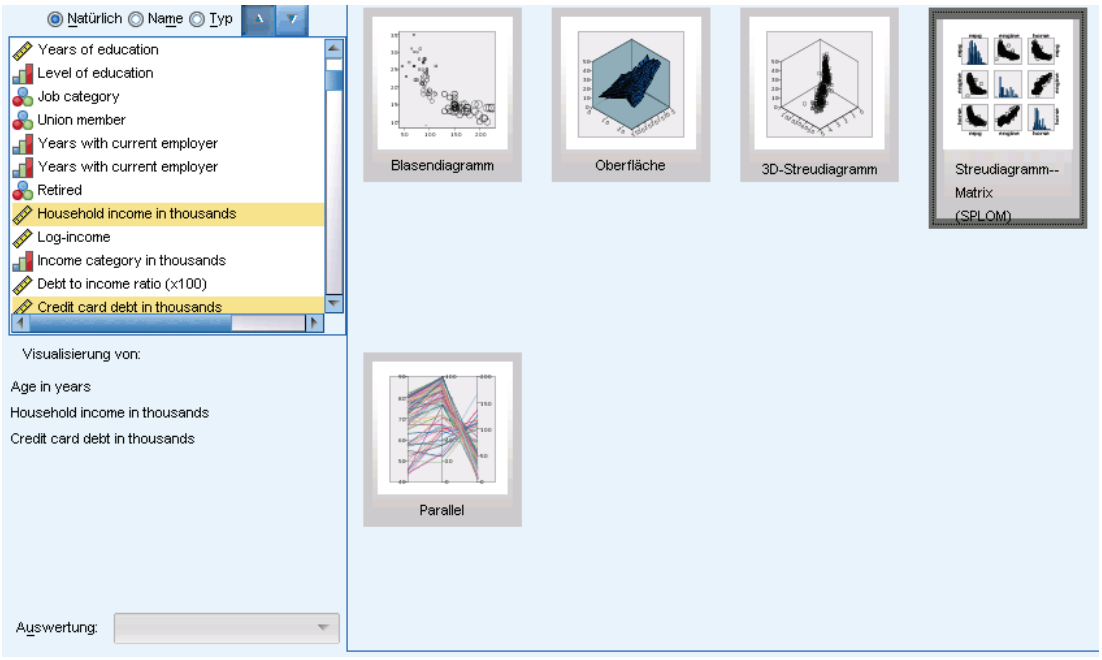
Hinweis: Für dieses Beispiel wird die Datei *customer_subset* verwendet.

▶ Fügen Sie einen Grafiktafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.

▶ Wählen Sie auf der Registerkarte "Basis" die Optionen *Age in years* (Alter in Jahren), *Household income in thousands* (Haushaltseinkommen in Tausend) und *Credit card debt in thousands* (Schulden auf Kreditkarte in Tausend). (Wenn Sie bei gedrückter Strg-Taste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)

▶ Wählen Sie SPLOM.

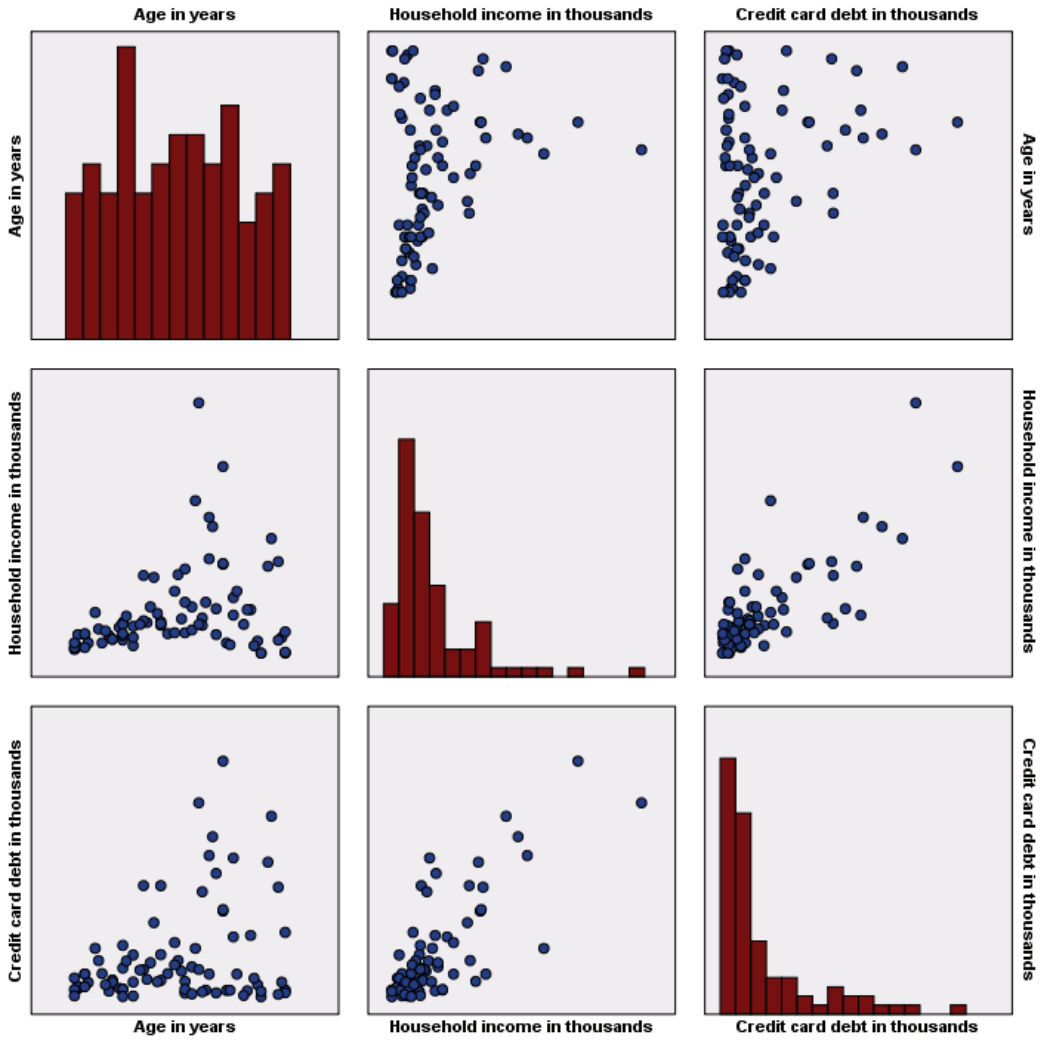
Abbildung 5-25
Auswahl auf der Registerkarte "Basis", SPLOM



▶ Klicken Sie auf Ausführen.

▶ Maximieren Sie das Ausgabefenster, um die Matrix besser sehen zu können.

Abbildung 5-26
Streudiagramm-Matrix (SPLOM)



Wir stellen Folgendes fest:

- Die auf der Diagonale angezeigten Histogramme stellen die Verteilung jeder Variablen in der SPLOM dar. Das Histogramm für *age* (Alter) wird in der oberen linken Zelle, das für *income* (Einkommen) in der mittleren Zelle und das für *creddebt* (Kreditkartenschulden) in der Zelle unten rechts dargestellt. Keine der Variablen weist eine Normalverteilung auf. Das heißt, keines der Histogramme ähnelt einer Glockenkurve. Beachten Sie auch, dass die Histogramme für *income* (Einkommen) und *creddebt* (Kreditkartenschulden) positiv schief sind.

- Es scheint keine Beziehung zwischen *age* (Alter) und den anderen Variablen zu geben.
- Zwischen *income* (Einkommen) und *creddebt* (Kreditkartenschulden) besteht ein lineares Verhältnis. Das heißt, die Kreditkartenschulden (*creddebt*) steigen, wenn das Einkommen (*income*) zunimmt. Gegebenenfalls können eigene Streudiagramme für diese Variablen und andere zugehörige Variablen erstellt werden, um die Beziehungen genauer zu untersuchen.

Beispiel: Choroplethenkarten (Farbkarten) von Summen

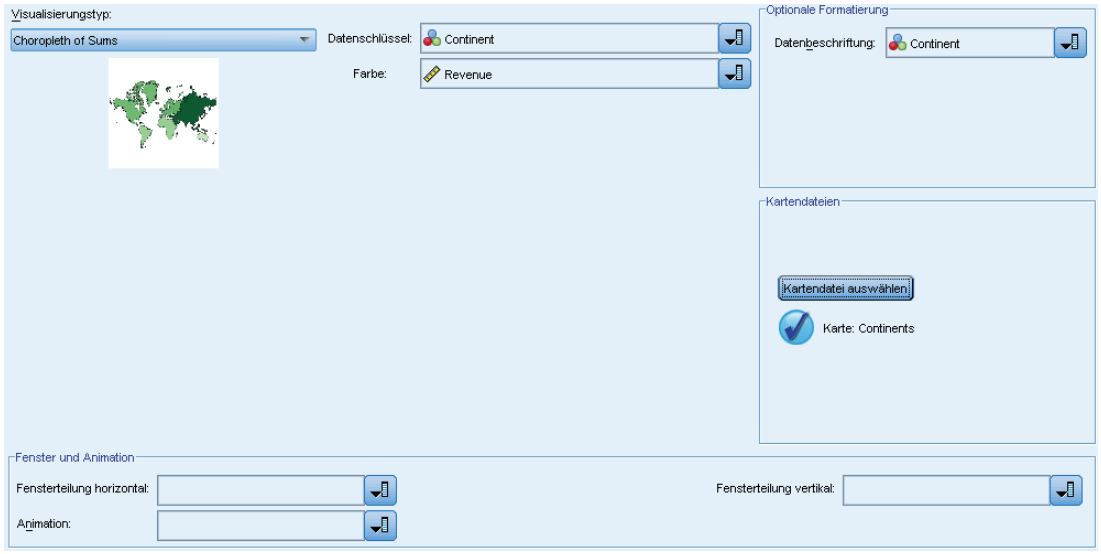
Nun erstellen wir eine Kartenvisualisierung. Im anschließenden Beispiel erstellen wir dann eine Variation dieser Visualisierung. Beim Daten-Set handelt es sich um *worldsales*. Dieses ist eine hypothetische Datendatei, die Verkaufserlöse nach Kontinent und Produkt enthält.

- ▶ Fügen Sie einen Grafiktafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.
- ▶ Wählen Sie auf der Registerkarte "Basis" die Optionen *Continent* (Kontinent) und *Revenue* (Ertrag) aus. (Wenn Sie bei gedrückter Strg-Taste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)
- ▶ Wählen Sie Choroplethenkarte von Summen aus.
- ▶ Klicken Sie auf die Registerkarte "Detailliert".
- ▶ Wählen Sie in der Gruppe "Optionale Formatierungen" die Option *Continent* (Kontinent) aus der Dropdown-Liste der Datenbeschriftungen aus.
- ▶ Klicken Sie in der Gruppe "Kartendateien" auf Kartendatei auswählen.
- ▶ Stellen Sie sicher, dass Karte im Dialogfeld "Karten auswählen" mit *Continents* (Kontinente) und Kartenschlüssel mit *CONTINENT* (KONTINENT) festgelegt ist.
- ▶ Klicken Sie in den Gruppen "Karten- und Datenwerte vergleichen" auf Vergleichen, um sicherzustellen, dass die Kartenschlüssel mit den Datenschlüsseln übereinstimmen. In diesem Beispiel weisen alle Datenschlüsselwerte entsprechende Kartenschlüssel und Funktionen auf. Außerdem wird angezeigt, dass für Ozeanien keine Daten vorliegen.

Abbildung 5-27
Dialogfeld "Karten auswählen"

▶ Klicken Sie im Dialogfeld "Karten auswählen" auf OK.

Abbildung 5-28
Auswahl auf der Registerkarte "Basis," Choroplethenkarte der Summe



▶ Klicken Sie auf Ausführen.

Abbildung 5-29
Choroplethenkarte der Summe

Mit dieser Kartenvisualisierung können wir mühelos erkennen, dass der Ertrag in Nordamerika am höchsten und in Südamerika sowie in Afrika am niedrigsten ist. Jeder Kontinent ist beschriftet, da wir für die Datenbeschriftungsformatierung die Option *Continent* (Kontinent) ausgewählt haben.

Beispiel: Balkendiagramme auf einer Karte

In diesem Beispiel wird verdeutlicht, wie sich der Ertrag auf jedem Kontinent nach Produkt aufteilen lässt.

Hinweis: In diesem Beispiel wird die Datendatei *worldsales* verwendet.

▶ Fügen Sie einen Grafiktafelknoten hinzu und öffnen Sie ihn zur Bearbeitung.

▶ Wählen Sie auf der Registerkarte “Basis” die Optionen *Continent* (Kontinent), *Product* (Produkt) und *Revenue* (Ertrag) aus. (Wenn Sie bei gedrückter Strg-Taste klicken, können Sie mehrere Felder bzw. Variablen auswählen.)

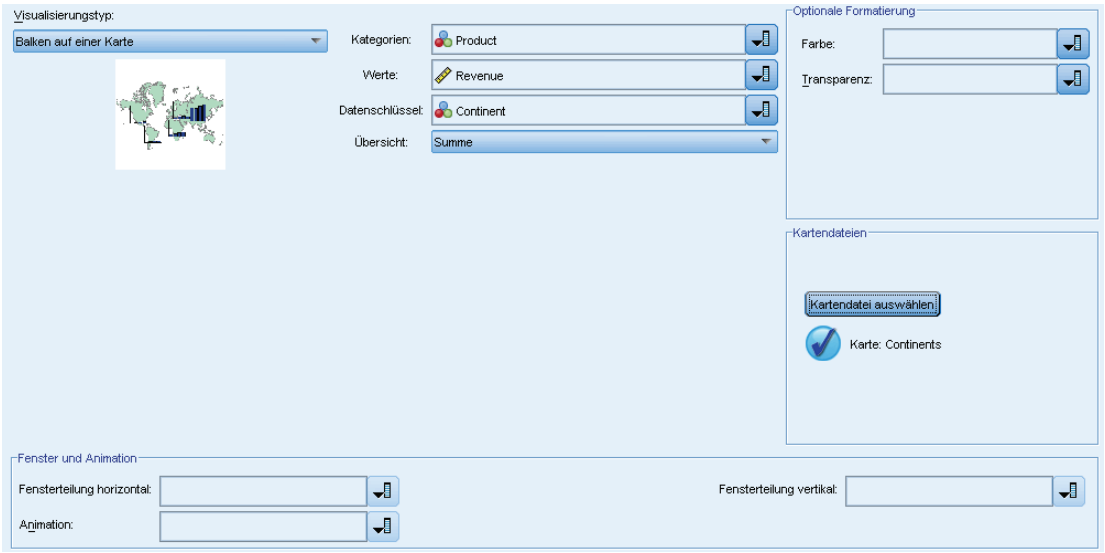
▶ Wählen Sie Balken auf einer Karte aus.

▶ Klicken Sie auf die Registerkarte “Detailliert”.

Beim Verwenden mehrerer Felder eines bestimmten Typs ist es wichtig zu prüfen, dass jedes Feld dem richtigen Abschnitt zugewiesen ist.

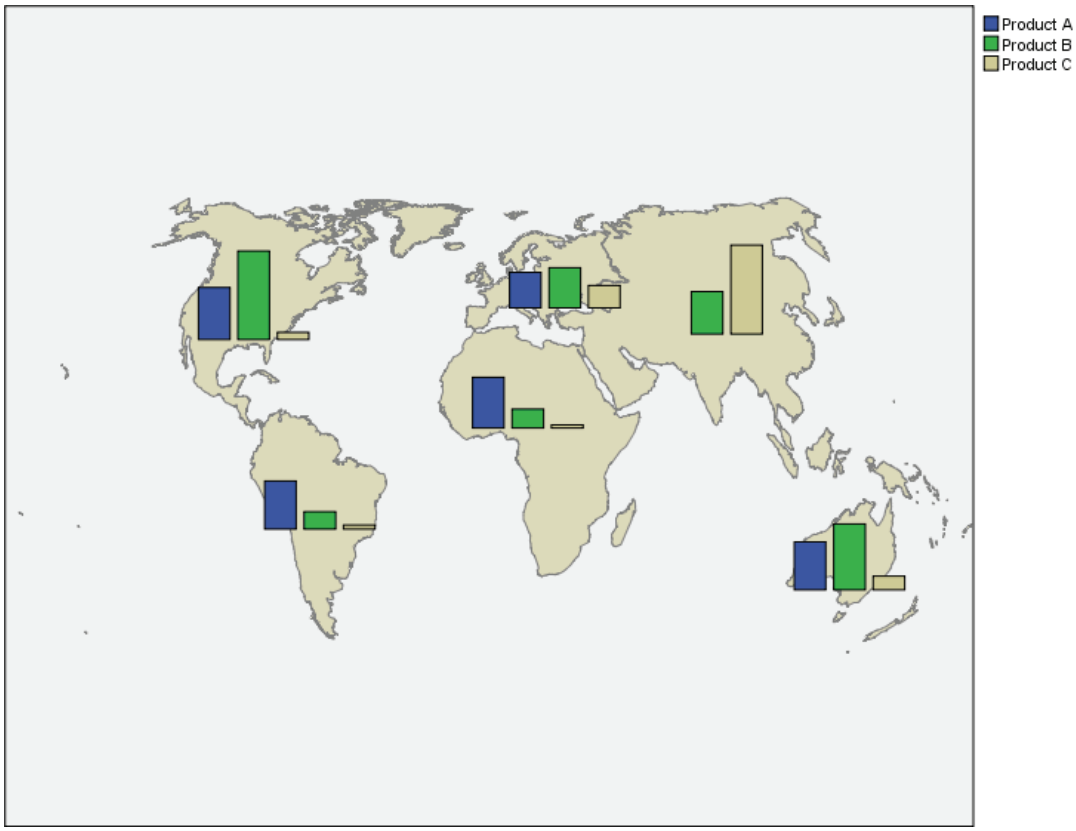
- Wählen Sie in der Dropdown-Liste “Kategorien” den Eintrag *Product* (Produkt) aus.
- Wählen Sie in der Dropdown-Liste “Werte” den Eintrag *Revenue* (Ertrag) aus.
- Wählen Sie in der Dropdown-Liste “Datenschlüssel” den Eintrag *Continent* (Kontinent) aus.
- Wählen Sie in der Dropdown-Liste “Zusammenfassung” den Eintrag *Sum* (Summe) aus.
- Klicken Sie in der Gruppe “Kartendateien” auf Kartendatei auswählen.
- Stellen Sie sicher, dass Karte im Dialogfeld “Karten auswählen” mit *Continents* (Kontinente) und Kartenschlüssel mit *CONTINENT* (KONTINENT) festgelegt ist.
- Klicken Sie in den Gruppen “Karten- und Datenwerte vergleichen” auf Vergleichen, um sicherzustellen, dass die Kartenschlüssel mit den Datenschlüsseln übereinstimmen. In diesem Beispiel weisen alle Datenschlüsselwerte entsprechende Kartenschlüssel und Funktionen auf. Außerdem wird angezeigt, dass für Ozeanien keine Daten vorliegen.
- Klicken Sie im Dialogfeld “Karten auswählen” auf OK.

Abbildung 5-30
Auswahl auf der Registerkarte “Basis”, Balken auf einer Karte



- Klicken Sie auf Ausführen.
- Maximieren Sie das angezeigte Ausgabefenster, um die Anzeige besser sehen zu können.

Abbildung 5-31
Balkendiagramme auf einer Karte



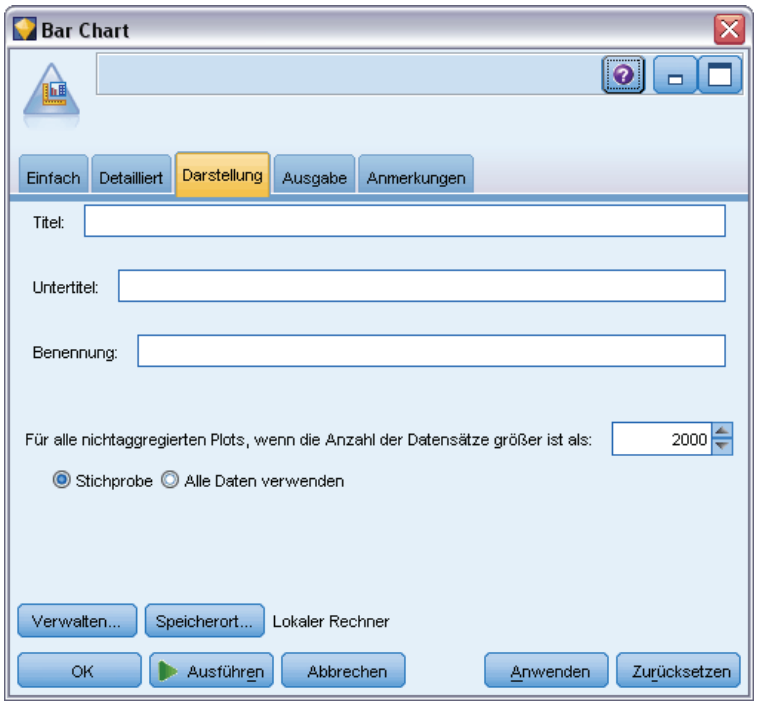
Wir stellen Folgendes fest:

- Die Verteilung des Gesamtertrags über Produkte hinweg ist in Südamerika und Afrika sehr ähnlich.
- *Product C* (Produkt C) erzeugt mit Ausnahme von Asien am wenigsten Ertrag.
- *Product A* (Produkt A) erbringt in Asien keinen bzw. nur minimalen Ertrag.

Diagrammtafel – Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Abbildung 5-32
Einstellungen auf der Registerkarte "Darstellung" für einen Diagrammtafelknoten



Allgemeine Darstellungsoptionen

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Benennung. Dient zur Eingabe des Texts, der zur Benennung des Diagramms verwendet werden soll.

Stichprobenziehung. Geben Sie eine Methode für umfangreichere Daten-Sets an. Sie können wahlweise eine maximal zulässige Größe für das Daten-Set angeben oder den Standardwert für die Anzahl an Datensätzen verwenden. Bei umfangreichen Daten-Sets steigt die Leistung, wenn Sie die Option Stichprobe aktivieren. Alternativ können Sie mit Alle Daten verwenden alle Datenpunkte gleichzeitig plotten lassen; dies kann sich jedoch beträchtlich auf die Leistung der Software auswirken.

Optionen für das Stylesheet-Erscheinungsbild

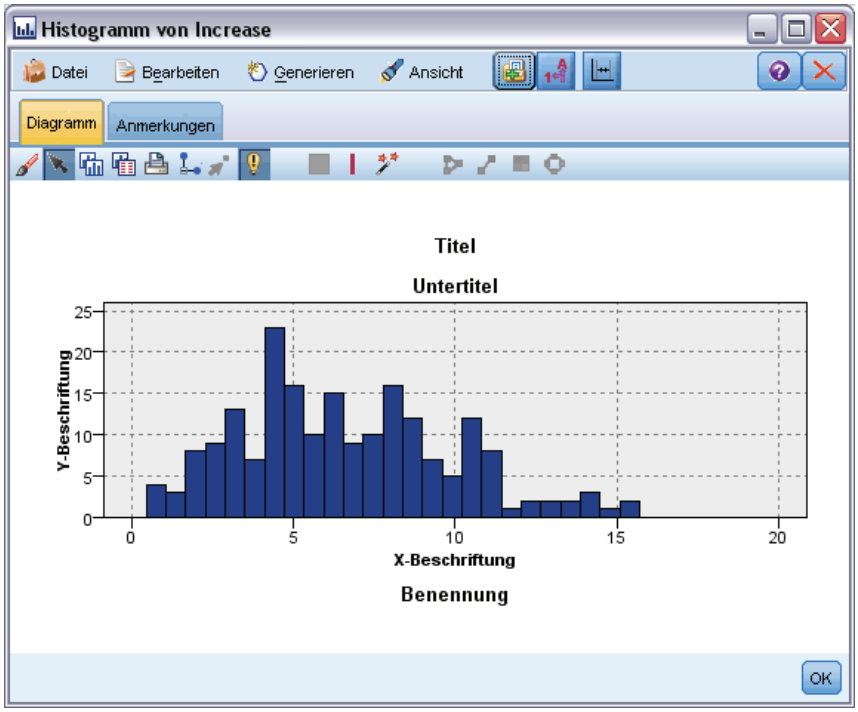
Es gibt zwei Schaltflächen, über die Sie steuern können, welche Visualisierungsvorlagen (sowie Stylesheets und Karten) verfügbar sind:

Verwalten. Verwalten von Visualisierungsvorlagen, Stylesheets und Karten auf dem Computer. Sie können Visualisierungsvorlagen , Stylesheets und Karten auf Ihrem lokalen Rechner importieren, exportieren, umbenennen und löschen. Für weitere Informationen siehe Thema Verwalten von Vorlagen, Stylesheets und Kartendateien auf S. 292.

Speicherort. Ändern des Speicherorts von Visualisierungsvorlagen, Stylesheets und Karten. Der aktuelle Speicherort wird rechts neben der Schaltfläche angezeigt. Für weitere Informationen siehe Thema Einstellen des Speicherorts für Vorlagen, Stylesheets und Karten auf S. 290.

Das folgende Beispiel zeigt, wo sich die Darstellungsoptionen in einem Diagramm befinden. (*Anmerkung*: Es ist nicht bei jedem Diagramm jede Option möglich.)

Abbildung 5-33
Position verschiedener Diagramm-Darstellungsoptionen



Einstellen des Speicherorts für Vorlagen, Stylesheets und Karten

Visualisierungsvorlagen, -Stylesheets und Kartendateien werden in einem bestimmten lokalen Ordner im IBM® SPSS® Collaboration and Deployment Services Repository gespeichert. Beim Auswählen von Vorlagen, Stylesheets und Karten werden in diesem Ordner nur die vordefinierten Elemente angezeigt. Wenn alle Vorlagen, Stylesheets und Kartendateien an einem Ort gespeichert werden, können IBM SPSS-Anwendungen leicht darauf zugreifen. Informationen zum Hinzufügen weiterer Vorlagen, Stylesheets und Karten zu diesem Speicherort finden Sie unter Verwalten von Vorlagen, Stylesheets und Kartendateien auf S. 292.

So legen Sie den Speicherort für Vorlagen, Stylesheets und Kartendateien fest

- Klicken Sie im Dialogfeld einer Vorlage, eines Stylesheets oder einer Karte auf Speicherort..., um das Dialogfeld “Vorlagen, Stylesheets und Karten” anzuzeigen.

▶ Wählen Sie eine Option für den Standardpfad für Vorlagen, Stylesheets und Kartendateien:

Lokaler Rechner. Vorlagen, Stylesheets und Kartendateien werden in einem bestimmten Ordner auf Ihrem lokalen Computer gespeichert. Unter Windows XP ist dieser Ordner *C:\Dokumente und Einstellungen\<Benutzer>\Anwendungsdaten\SPSSInc\Graphboard*. Dieser Ordner kann nicht geändert werden.

IBM® SPSS® Collaboration and Deployment Services Repository. Vorlagen, Stylesheets und Kartendateien werden in einem vom Benutzer angegebenen Ordner im IBM SPSS Collaboration and Deployment Services Repository gespeichert. Um den Ordner zu betrachten, klicken Sie auf Ordner. Weitere Informationen finden Sie unter Verwenden von IBM SPSS Collaboration and Deployment Services Repository als Speicherort für Vorlagen, Stylesheets und Kartendateien auf S. 291.

▶ Klicken Sie auf OK.

Verwenden von IBM SPSS Collaboration and Deployment Services Repository als Speicherort für Vorlagen, Stylesheets und Kartendateien

Visualisierungs-Vorlagen und -Stylesheets können im IBM® SPSS® Collaboration and Deployment Services Repository gespeichert werden. Der Speicherort ist ein bestimmter Ordner im IBM SPSS Collaboration and Deployment Services Repository. Wird dieser Ordner als Standardpfad angegeben, können alle Vorlagen, Stylesheets und Kartendateien in diesem Ordner ausgewählt werden.

So legen Sie einen Ordner in IBM SPSS Collaboration and Deployment Services Repository als Speicherort für Vorlagen, Stylesheets und Kartendateien fest

▶ Klicken Sie in einem Dialogfeld mit der Schaltfläche “Speicherort” auf Speicherort....

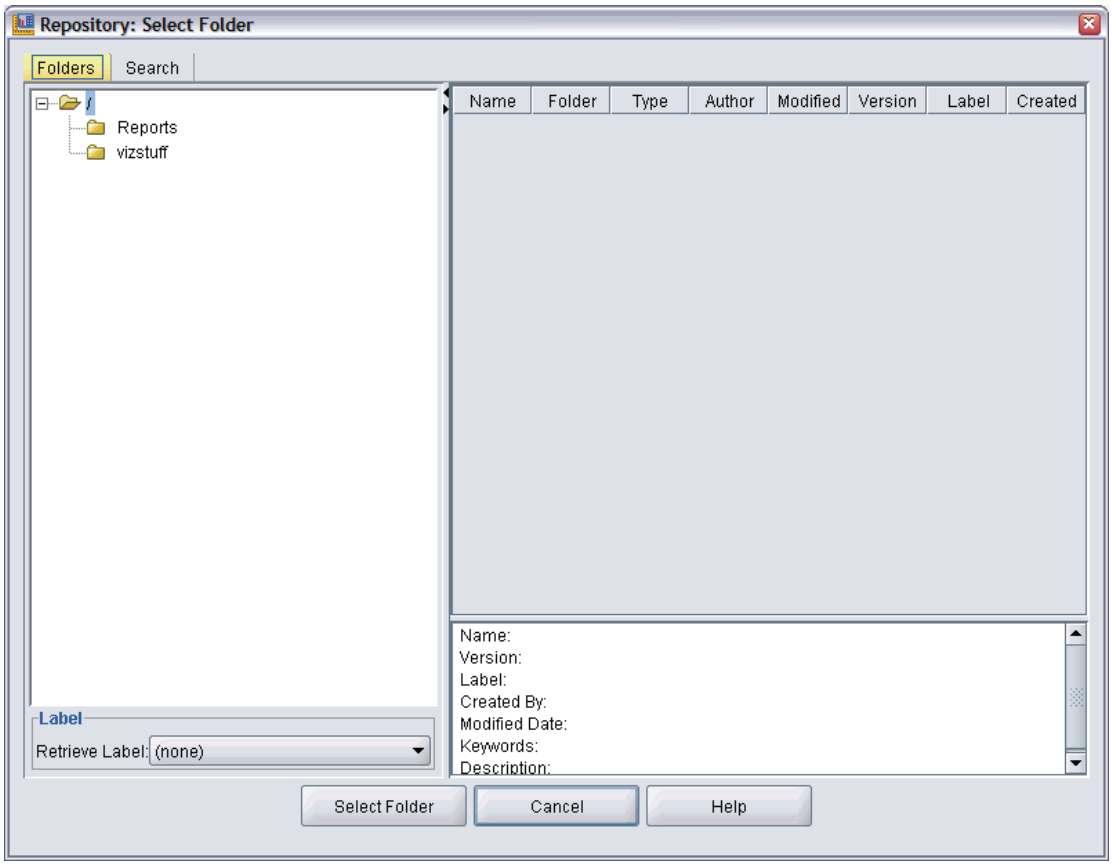
▶ Wählen Sie IBM® SPSS® Collaboration and Deployment Services Repository aus.

▶ Klicken Sie auf Ordner.

Hinweis: Wenn Sie noch nicht mit dem IBM SPSS Collaboration and Deployment Services Repository verbunden sind, werden Sie aufgefordert, Ihre Verbindungsdaten einzugeben.

▶ Wählen Sie im Dialogfeld “Ordner auswählen” den Ordner aus, in dem Vorlagen, Stylesheets und Kartendateien gespeichert werden.

Abbildung 5-34
Dialogfeld "Ordner auswählen"



- Wählen Sie bei Bedarf eine Beschriftung aus Beschriftung abrufen. Es werden nur Vorlagen, Stylesheets und Kartendateien mit dieser Beschriftung angezeigt.
- Wenn Sie einen Ordner mit einer bestimmten Vorlage oder Kartendatei oder einem bestimmten Stylesheet suchen, können Sie die Vorlage, das Stylesheet oder die Kartendatei mithilfe der Registerkarte "Suchen" ausfindig machen. Im Dialogfeld "Ordner auswählen" wird automatisch der Ordner ausgewählt, in dem sich die gefundene Vorlage oder Kartendatei oder das Stylesheet befindet.
- Klicken Sie auf Ordner auswählen.

Verwalten von Vorlagen, Stylesheets und Kartendateien

Mithilfe des Dialogfelds "Vorlagen, Stylesheets und Kartendateien verwalten" können Sie die Vorlagen, Stylesheets und Karten im lokalen Speicherort Ihres Computers verwalten. In diesem Dialogfeld können Sie Visualisierungsvorlagen, Stylesheets und Kartendateien im lokalen Speicherort des Computers importieren, exportieren, umbenennen und löschen.

- Klicken Sie in einem der Dialogfelder, in dem Sie Vorlagen, Stylesheets oder Karten auswählen, auf Verwalten....

Dialogfeld "Vorlagen, Stylesheets und Karten verwalten"

Auf der Registerkarte "Vorlagen" werden alle lokalen Vorlagen aufgelistet. Auf der Registerkarte "Stylesheets" werden alle lokalen Stylesheets samt Beispielvisualisierungen mit Beispieldaten aufgelistet. Sie können ein Stylesheet auswählen, um dessen Stile auf die Beispiel-Visualisierungen anzuwenden. Für weitere Informationen siehe Thema Stylesheets anwenden auf S. 393.Auf der Registerkarte "Karten" werden alle lokalen Kartendateien aufgelistet. Diese Registerkarte dient auch zum Anzeigen der Kartenschlüssel samt Beispielwerten, eines Kommentars (sofern bei Erstellung der Karte angegeben) und einer Vorschau der Karte.

Die folgenden Schaltflächen können auf beiden Registerkarten verwendet werden, unabhängig davon, welche aktuell aktiv ist.

Importieren. Dient zum Importieren einer Visualisierungsvorlage, eines Stylesheets oder einer Kartendatei aus dem Dateisystem. Durch den Import einer Vorlage, eines Stylesheets oder einer Kartendatei kann die IBM SPSS-Anwendung darauf zugreifen. Wenn Ihnen ein anderer Benutzer eine Vorlage, ein Stylesheet oder eine Kartendatei gesendet hat, müssen Sie die Datei importieren, bevor Sie sie in Ihrer Anwendung verwenden.

Exportieren. Dient zum Exportieren einer Visualisierungsvorlage, eines Stylesheets oder einer Kartendatei aus dem Dateisystem. Exportieren Sie eine Vorlage, ein Stylesheet oder eine Kartendatei, wenn Sie sie an einen anderen Benutzer senden möchten.

Umbenennen. Dient zum Umbenennen der ausgewählten Visualisierungsvorlage, des Stylesheets oder der Kartendatei. Sie können den Namen nicht in einen Namen ändern, der bereits verwendet wird.

Kartenschlüssel exportieren. Exportiert die Kartenschlüssel als Datei mit kommagetrennten Werten (CSV). Diese Schaltfläche ist nur auf der Registerkarte "Karte" aktiviert.

Löschen. Dient zum Löschen der ausgewählten Visualisierungsvorlagen, Stylesheets oder Kartendateien. Über STRG-Klicken können Sie mehrere Visualisierungsvorlagen, Stylesheets oder Kartendateien auswählen. Löschvorgänge können nicht rückgängig gemacht werden. Gehen Sie daher mit Bedacht vor.

Konvertieren und Verteilen von Shapefiles für Karten

Mit der Grafiktafel-Vorlagenauswahl können Sie Kartenvisualisierungen aus der Kombination aus einer Visualisierungsvorlage und einer SMZ-Datei erstellen. SMZ-Dateien ähneln ESRI Shapefiles (Dateiformat SHP) dahingehend, dass sie die geografischen Informationen zum Zeichnen einer Karte (z. B. Ländergrenzen) enthalten. Sie sind jedoch für Kartenvisualisierungen optimiert. Die Grafiktafel-Vorlagenauswahl ist mit einer ausgewählten Anzahl an SMZ-Dateien vorinstalliert. Wenn Sie bereits ein ESRI Shapefile besitzen, das Sie für Kartenvisualisierungen verwenden möchten, müssen Sie zunächst das Shapefile mithilfe des Dienstprogramms zur Konvertierung von Karten in eine SMZ-Datei konvertieren. Das Dienstprogramm zur Konvertierung von Karten unterstützt ESRI Shapefiles mit Punkten, Polylinien oder Polygonen (Shape-Typen 1, 3 und 5), die eine einzige Schicht enthalten.

Zusätzlich zur Konvertierung von ESRI Shapefiles können Sie mit dem Dienstprogramm zur Konvertierung von Karten auch den Detaillierungsgrad der Karte bearbeiten, Strukturbeschriftungen ändern, Strukturen zusammenführen und Strukturen verschieben sowie zahlreiche weitere optionale Änderungen vornehmen. Außerdem können Sie mit dem Dienstprogramm zur Konvertierung von Karten auch bestehende SMZ-Dateien (einschließlich der vorinstallierten) bearbeiten.

Bearbeiten vorinstallierter SMZ-Dateien

- Exportieren Sie die SMZ-Datei aus dem Verwaltungssystem. Für weitere Informationen siehe Thema Verwalten von Vorlagen, Stylesheets und Kartendateien auf S. 292.
- Verwenden Sie das Dienstprogramm zur Konvertierung von Karten zum Öffnen und Bearbeiten der exportierten SMZ-Datei. Es wird empfohlen, die Datei unter einem anderen Namen zu speichern. Für weitere Informationen siehe Thema Verwenden des Dienstprogramms zur Konvertierung von Karten auf S. 295.
- Importieren Sie die geänderte SMZ-Datei aus dem Verwaltungssystem. Für weitere Informationen siehe Thema Verwalten von Vorlagen, Stylesheets und Kartendateien auf S. 292.

Zusätzliche Ressourcen für Kartendateien

Geodaten im Dateiformat SHP, die für Ihre kartenbezogenen Anforderungen verwendet werden können, wird von vielen privaten und öffentlichen Quellen angeboten. Informieren Sie sich auf den Websites Ihrer jeweiligen Regierung, wenn Sie nach kostenlosen Daten suchen. Viele der Vorlagen dieses Produkts basieren auf öffentlich verfügbaren Daten, die von GeoCommons (*http://www.geocommons.com*) und der Volkszählungsbehörde U.S. Census Bureau (*http://www.census.gov*) bezogen wurden. Eine weitere Quelle für Geodaten auf sämtlichen Ebenen in den USA ist die wissenschaftliche Behörde U.S. Geological Survey (*http://www.geodata.gov*).

WICHTIGER HINWEIS: Informationen zu Produkten von Drittanbietern wurden von den Anbietern des jeweiligen Produkts, aus deren veröffentlichten Ankündigungen oder anderen, öffentlich verfügbaren Quellen bezogen. IBM hat diese Produkte nicht getestet und kann die Genauigkeit bezüglich Leistung, Kompatibilität oder anderen Behauptungen nicht bestätigen, die sich auf Drittanbieter-Produkte beziehen. Fragen bezüglich der Funktionen von Drittanbieter-Produkten sollten an die Anbieter der jeweiligen Produkte gerichtet werden. Jegliche Verweise auf Drittanbieter-Websites in dieser Information werden nur der Vollständigkeit halber bereitgestellt und dienen nicht als Befürwortung dieser. Das Material auf diesen Websites ist kein Bestandteil des Materials dieses IBM-Programms, sofern dies nicht in einer Datei mit Hinweisen zu diesem IBM-Programm anders vermerkt ist. Die Verwendung des Materials dieser Websites erfolgt auf eigene Gefahr.

Wichtige Konzepte im Zusammenhang mit Karten

Sie sollten sich mit einigen wichtigen Konzepten im Zusammenhang mit Shapefiles vertraut machen, um das Dienstprogramm zur Konvertierung von Karten effektiver nutzen zu können.

Ein **Shapefile** stellt die geografischen Informationen bereit, die zum Zeichnen einer Karte erforderlich sind. Es gibt drei Typen von Shapefiles, die vom Dienstprogramm zur Konvertierung von Karten unterstützt werden:

- **Punkt.** Das Shapefile gibt die Position von Punkten (z. B. Städte) an.
- **Polylinie.** Das Shapefile gibt Wegverläufe und deren Position (z. B. Flüsse) an.
- **Polygon.** Das Shapefile gibt begrenzte Regionen und ihre Position (z. B. Länder) an.

Am häufigsten werden Polygon-Shapefiles verwendet. Choroplethenkarten werden aus Poygon-Shapefiles erstellt. Bei Choroplethenkarten werden Farben zur Darstellung von Werten innerhalb einzelner Polygone (Regionen) verwendet. Punkt- und Polylinien-Shapefiles werden üblicherweise auf ein Polygon-Shapefile gelegt (Überlagerung). Ein Beispiel hierfür ist ein Punkt-Shapefile mit deutschen Städten, das ein Shapefile der deutschen Bundesländer überlagert.

Shapefiles bestehen aus **Strukturen**. Strukturen sind die einzelnen geografischen Elemente. Beispiele für Strukturen sind Länder, Bundesländer, Städte usw. Das Shapefile enthält auch Daten zu den Strukturen. Diese Daten werden in **Attributen** gespeichert. Attribute ähneln Feldern oder Variablen in einer Datendatei. Es gibt mindestens ein Attribut, das der **Kartenschlüssel** für die Struktur ist. Beim Kartenschlüssel kann es sich um eine Beschriftung, wie beispielsweise den Namen eines Landes oder Bundeslandes handeln. Der Kartenschlüssel ist das Element, das Sie mit einer Variablen bzw. einem Feld in einer Datendatei verknüpfen, um eine Kartenvisualisierung zu erstellen.

Beachten Sie, dass in der SMZ-Datei nur das Schlüsselattribut bzw. die Schlüsselattribute beibehalten werden können. Das Dienstprogramm zur Konvertierung von Karten unterstützt nicht die Speicherung zusätzlicher Attribute. Daher müssen Sie mehrere SMZ-Dateien erstellen, wenn Sie eine Zusammenfassung in verschiedenen Schichten durchführen möchten. Wenn Sie beispielsweise Bundesländer und Landkreise zusammenfassen möchten, benötigen Sie getrennte SMZ-Dateien: eine mit einem Schlüssel für die Bundesländer und eine mit einem Schlüssel für die Landkreise.

Verwenden des Dienstprogramms zur Konvertierung von Karten

So starten Sie das Dienstprogramm zur Konvertierung von Karten:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Werkzeuge > Dienstprogramm zur Konvertierung von Karten

Das Dienstprogramm zur Konvertierung von Karten enthält vier Hauptbildschirme (Schritte). Einer der Schritte beinhaltet auch Unterschritte für detailliertere Festlegungen zur Bearbeitung der Kartendatei.

Schritt 1 – Ziel und Quellendatei auswählen

Zunächst müssen Sie eine Kartendatei als Quelle und ein Ziel für die konvertierte Kartendatei auswählen. Sie benötigen sowohl die *.shp* als auch die *.dbf*-Datei für das Shapefile.

Wählen Sie die zu konvertierende SHP- (ESRI) oder SMZ-Datei aus Navigieren Sie zu einer bestehenden Kartendatei auf Ihrem Computer. Dies ist die Datei, in die die Konvertierung erfolgt und die als SMZ-Datei gespeichert wird. Die *.dbf*-Datei für das Shapefile *muss* im selben Verzeichnis gespeichert werden und einen Basisdateinamen tragen, der mit dem Namen der *.shp*-Datei übereinstimmt. Die *.dbf*-Datei ist erforderlich, da sie Attributinformationen für die *.shp*-Datei enthält.

Geben Sie Ziel und Dateinamen für die konvertierte Kartendatei an. Geben Sie einen Pfad und einen Dateinamen für die SMZ-Datei ein, die aus der ursprünglichen Karten-Quellendatei erstellt wird.

Schritt 2 - Kartenschlüssel auswählen

Nun wählen Sie aus, welche Kartenschlüssel in die SMZ-Datei aufgenommen werden sollen. Sie können anschließend einige Optionen ändern, die beeinflussen, wie die Karte gerendert wird. Die anschließenden Schritte im Dienstprogramm zur Konvertierung von Karten beinhalten eine Vorschau der Karte. Die ausgewählten Rendering-Optionen werden zur Erzeugung der Kartenvorschau verwendet.

Primären Kartenschlüssel auswählen. Wählen Sie das Attribut aus, das als primärer Schlüssel zur Angabe und Beschriftung von Strukturen in der Karte dient. Der primäre Schlüssel einer Weltkarte könnte beispielsweise das Attribut für die Ländernamen sein. Der primäre Schlüssel verknüpft außerdem die Daten mit den Kartenstrukturen. Achten Sie also darauf, dass die Werte (Beschriftungen) des ausgewählten Attributs mit den Werten in Ihren Daten übereinstimmen. Bei der Auswahl eines Attributs werden Beispielbeschriftungen angezeigt. Wenn Sie diese Beschriftungen ändern möchten, können Sie dies in einem späteren Schritt tun.

Einzuschließende Alternativschlüssel auswählen. Markieren Sie neben dem primären Kartenschlüssel auch alle anderen Schlüsselattribute, die in die generierte SMZ-Datei aufgenommen werden sollen. So können beispielsweise manche Attribute übersetzte Beschriftungen enthalten. Wenn Sie Daten erwarten, die in anderen Sprachen codiert sind, ist es sinnvoll, diese Attribute beizubehalten. Beachten Sie, dass Sie nur diejenigen zusätzlichen Schlüssel auswählen können, die für dieselben Strukturen stehen wie der primäre Schlüssel. Wenn es sich beim primären Schlüssel beispielsweise um die vollständigen Namen der US-Bundesstaaten handelt, können Sie nur diejenigen Alternativschlüssel auswählen, die ebenfalls für US-Bundesstaaten stehen, also beispielsweise die Kürzel der Bundesstaaten.

Karte automatisch glätten. Shapefiles mit Polygonen enthalten üblicherweise zu viele Datenpunkte und zu viele Details für statistische Kartenvisualisierungen. Die überschüssigen Details können ablenkend wirken und die Leistung beeinträchtigen. Durch Glätten können Sie den Detaillierungsgrad verringern und die Karte verallgemeinern. Die Karte sieht dadurch prägnanter aus lässt sich schneller rendern. Wenn die Karte automatisch geglättet wird, beträgt der maximale Winkel 15 Grad und der beizubehaltende Bereich beträgt 99. Informationen zu diesen Einstellungen finden Sie unter Karte glätten auf S. 297. Beachten Sie, dass Sie die Gelegenheit haben, später in einem anderen Schritt eine weitere Glättung durchzuführen.

Grenzen zwischen sich berührenden Polygonen der gleichen Struktur entfernen. Einige Strukturen können Unterstrukturen enthalten, deren Grenzen innerhalb der relevanten Hauptstrukturen liegen. Beispielsweise kann eine Weltkarte der Kontinente interne Grenzen für die Länder auf den einzelnen Kontinenten aufweisen. Wenn Sie diese Option auswählen, werden die internen

Grenzen nicht in der Karte angezeigt. Beim Beispiel mit der Weltkarte der Kontinente werden durch die Auswahl dieser Option die Ländergrenzen entfernt, während die Grenzen der Kontinente beibehalten werden.

Schritt 3 - Karte bearbeiten

Nachdem Sie die grundlegenden Optionen für die Karte angegeben haben, können Sie konkretere Optionen bearbeiten. Diese Änderungen sind optional. Dieser Schritt des Dienstprogramms zur Konvertierung von Karten führt Sie durch die zugehörigen Aufgaben und zeigt eine Vorschau der Karte an, mit der Sie die Änderungen überprüfen können. Je nach Shapefile-Typ (Punkt, Polylinie oder Polygon) und Koordinatensystem stehen einige Aufgaben möglicherweise nicht zur Verfügung.

Für alle Aufgaben werden auf der linken Seite des Dienstprogramms zur Konvertierung von Karten die folgenden allgemeinen Steuerelemente angezeigt:

Beschriftungen auf Karte anzeigen. Standardmäßig werden in der Vorschau keine Strukturbeschriftungen angezeigt. Sie können auswählen, dass die Beschriftungen angezeigt werden sollen. Die Beschriftungen erleichtern zwar möglicherweise die Identifizierung der Strukturen, sie können jedoch die Direktauswahl auf der Vorschaukarte behindern. Aktivieren Sie diese Option, wenn Sie sie benötigen, beispielsweise für die Bearbeitung der Strukturbeschriftungen.

Farben für Vorschaukarte auswählen. Standardmäßig werden auf der Vorschaukarte Bereiche mit einer Volltonfarbe angezeigt. Alle Strukturen haben dieselbe Farbe. Sie können festlegen, dass den einzelnen Kartenstrukturen eine Reihe verschiedener Farben zugewiesen wird. Diese Option kann die Unterscheidung verschiedener Strukturen auf der Karte erleichtern. Sie ist besonders hilfreich, wenn Sie Strukturen zusammenführen und sehen möchten, wie die neuen Strukturen in der Vorschau dargestellt werden.

Für alle Aufgaben wird außerdem auf der rechten Seite des Dienstprogramms zur Konvertierung von Karten das folgende allgemeine Steuerelement angezeigt:

Rückgängig. Wenn Sie eine ungewollte Änderung vorgenommen haben, können Sie durch Klicken auf Rückgängig den vorherigen Zustand wiederherstellen. Es können maximal 100 Änderungen rückgängig gemacht werden.

Karte glätten

Shapefiles mit Polygonen enthalten üblicherweise zu viele Datenpunkte und zu viele Details für statistische Kartenvisualisierungen. Die überschüssigen Details können ablenkend wirken und die Leistung beeinträchtigen. Durch Glätten können Sie den Detaillierungsgrad verringern und die Karte verallgemeinern. Die Karte sieht dadurch prägnanter aus lässt sich schneller rendern. Diese Option steht nicht für Punkt- und Polylinienkarten zur Verfügung.

Max. Winkel. Der maximale Winkel, dessen Wert zwischen 1 und 20 liegen muss, gibt die Toleranz für die Glättung von Punktesets an, die nahezu linear sind. Ein größerer Wert bietet größere Toleranz für die lineare Glättung. Dabei wird anschließend eine größere Anzahl an Punkten verworfen, was zu einer generalisierten Karte führt. Zur Anwendung der linearen

Glättung überprüft das Dienstprogramm zur Konvertierung von Karten den Innenwinkel, der jeweils durch drei Punkte auf der Karte gebildet wird. Wenn 180 minus dem Winkel kleiner ist als der angegebene Wert, verwirft das Dienstprogramm zur Konvertierung von Karten den mittleren Punkt. Anders ausgedrückt: das Dienstprogramm zur Konvertierung von Karten überprüft, ob die von den drei Punkten gebildete Linie annähernd gerade ist. Wenn dies der Fall ist, behandelt das Dienstprogramm zur Konvertierung von Karten die Linie als Gerade zwischen den Endpunkten und verwirft den mittleren Punkt.

Beizubehaltender Prozentsatz. Der beizubehaltende Prozentsatz, bei dem es sich um einen Wert zwischen 90 und 100 handeln muss, legt fest, welche Menge an Landbereich beim Glätten der Karte beibehalten werden soll. Diese Option betrifft nur diejenigen Strukturen, die mehrere Polygone aufweisen, wie es der Fall ist, wenn eine Struktur Inseln beinhaltet. Wenn der Gesamtbereich einer Struktur minus eines Polygons größer ist als der angegebene Prozentsatz des ursprünglichen Bereichs, verwirft das Dienstprogramm zur Konvertierung von Karten das Polygon aus der Karte. Das Dienstprogramm zur Konvertierung von Karten entfernt niemals sämtliche Polygone für die Struktur. Es bleibt also stets mindestens ein Polygon für die Struktur erhalten, unabhängig vom angewendeten Glättungsumfang.

Klicken Sie nach der Auswahl des maximalen Winkels und des beizubehaltenden Prozentsatzes auf Zuweisen. Die Vorschau wird mit den Glättungsänderungen aktualisiert. Wenn die Karte weiterer Glättung bedarf, wiederholen Sie den Vorgang, bis der gewünschte Glättungsgrad erreicht ist. Beachten Sie, dass die mögliche Glättung begrenzt ist. Beim wiederholten Glätten wird irgendwann ein Punkt erreicht, an dem keine weitere Glättung auf die Karte angewendet werden kann.

Strukturbeschriftungen bearbeiten

Sie können die Strukturbeschriftungen nach Bedarf bearbeiten (beispielsweise so, dass sie den erwarteten Daten entsprechen) und auch die Position der Beschriftungen auf der Karte ändern. Auch wenn Sie nicht glauben, dass sie die Beschriftungen ändern müssen, sollten Sie sie überprüfen, bevor Sie Visualisierungen aus der Karte erstellen. Da in der Vorschau Beschriftungen nicht standardmäßig angezeigt werden, kann es auch sinnvoll sein, die Option Beschriftungen auf Karte anzeigen auszuwählen, um sie sichtbar zu machen.

Schlüssel. Wählen Sie den Schlüssel aus, der die zu überprüfenden/bearbeitenden Strukturbeschriftungen enthält.

Strukturen. In dieser Liste werden die im ausgewählten Schlüssel enthaltenen Strukturen angezeigt. Zur Bearbeitung der Beschriftung doppelklicken Sie auf die Liste. Wenn Beschriftungen in der Karte angezeigt werden, können Sie auch direkt in der Kartenvorschau auf die Strukturbeschriftungen doppelklicken. Wenn Sie die Beschriftungen mit einer Datendatei vergleichen möchten, klicken Sie auf Vergleichen.

X/Y. Diese Textfelder geben den aktuellen Mittelpunkt der Beschriftung für die ausgewählte Struktur in der Karte an. Die Einheiten werden in den Koordinaten der Karte angezeigt. Dabei kann es sich um lokale kartesische Koordinaten handeln (z. B. das State Plane Coordinate System für die USA) oder um geografische Koordinaten (wobei X den Längengrad und Y den Breitengrad angibt). Geben Sie Koordinaten für die neue Position der Beschriftung an. Wenn Beschriftungen

angezeigt werden, können Sie die Beschriftungen auch durch Klicken und Ziehen auf der Karte verschieben. Die Textfelder werden entsprechend der neuen Position aktualisiert.

Vergleichen. Wenn Sie über eine Datendatei mit Datenwerten verfügen, die den Strukturbeschriftungen für einen bestimmten Schlüssel entsprechen sollen, klicken Sie auf Vergleichen, um das Dialogfeld "Mit externer Datenquelle vergleichen" anzuzeigen. In diesem Dialogfeld können Sie die Datendatei öffnen und ihre Werte direkt mit den Strukturbeschriftungen des Kartenschlüssels vergleichen.

Dialogfeld "Mit externer Datenquelle vergleichen"

Im Dialogfeld "Mit externer Datenquelle vergleichen" können Sie eine Datei mit tabulatorgetrennten Werten (Erweiterung *.txt*) oder eine Datei mit kommagetrennten Werten (Erweiterung *.csv*) öffnen. Wenn die Datei geöffnet ist, können Sie in der Datendatei ein Feld auswählen, das mit den Strukturbeschriftungen in einem bestimmten Kartenschlüssel verglichen werden soll. Anschließend können Sie etwaige Diskrepanzen in der Kartendatei korrigieren.

Felder in der Datendatei. Wählen Sie das Feld aus, dessen Werte mit den Strukturbeschriftungen verglichen werden sollen. Wenn die erste Zeile der *.txt*- bzw. *.csv*-Datei deskriptive Beschriftungen für die einzelnen Felder enthält, klicken Sie auf Erste Zeile als Spaltenbeschriftungen verwenden. Andernfalls werden die einzelnen Felder durch ihre Position in der Datendatei angegeben (z. B. "Spalte 1", "Spalte 2" usw.).

Schlüssel für Vergleich. Wählen Sie den Kartenschlüssel aus, dessen Strukturbeschriftungen mit den Feldwerten der Datendatei verglichen werden sollen.

Vergleichen. Klicken Sie, wenn Sie mit dem Vergleich der Werte beginnen möchten.

Vergleichsergebnisse. Standardmäßig werden in der Tabelle "Vergleichsergebnisse" nur die nicht zugeordneten Feldwerte in der Datendatei aufgelistet. Die Anwendung versucht, eine zugehörige Strukturbeschriftung zu finden, normalerweise durch Prüfung auf eingefügte oder fehlende Leerzeichen. Klicken Sie auf die Dropdown-Liste in der Spalte *Kartenbeschriftung*, um die Strukturbeschriftung in der Kartendatei mit dem angezeigten Feldwert abzugleichen. Wenn Ihre Kartendatei keine entsprechende Strukturbeschriftung enthält, wählen Sie die Option *Ohne Zuordnung belassen* aus. Wenn Sie alle Feldwerte anzeigen möchten, auch diejenigen, die bereits mit einer Strukturbeschriftung übereinstimmen, deaktivieren Sie die Option Nur nicht zugeordnete Fälle anzeigen. Dies kann sinnvoll sein, um bestimmte Zuordnungen außer Kraft zu setzen.

Sie können jede Struktur nur einmal für die Zuordnung zu einem Feldwert verwenden. Wenn Sie mehrere Strukturen einem einzelnen Feldwert zuordnen möchten, können Sie die Strukturen zusammenführen und anschließend die neue, zusammengeführte Struktur dem Feldwert zuordnen. Weitere Informationen zum Zusammenführen von Strukturen finden Sie unter Strukturen zusammenführen auf S. 300.

Strukturen zusammenführen

Das Zusammenführen von Strukturen ist nützlich, um größere Regionen in einer Karte zu erstellen. Wenn Sie beispielsweise eine Karte der deutschen Bundesländer konvertieren, können Sie die Bundesländer (die Strukturen in diesem Beispiel) zu größeren Nord-, Süd-, Ost- und Westregionen zusammenführen.

Schlüssel. Wählen Sie den Kartenschlüssel aus, der die Strukturbeschriftungen enthält, mit denen Sie die zusammenzuführenden Strukturen identifizieren können.

Strukturen. Klicken Sie auf die erste zusammenzuführende Struktur. Klicken Sie mit gedrückter Strg-Taste auf die anderen Strukturen, die zusammengeführt werden sollen. Beachten Sie, dass die Strukturen auch in der Kartenvorschau ausgewählt werden. Neben der Auswahl aus der Liste haben Sie auch die Möglichkeit, direkt in der Kartenvorschau auf die Strukturen zu klicken und dann beim Klicken die Strg-Taste gedrückt zu halten.

Klicken Sie nach der Auswahl der zusammenzuführenden Strukturen auf Zusammenführen, um das Dialogfeld "Zusammengeführte Struktur benennen" anzuzeigen, in dem Sie eine Beschriftung auf die neue Struktur anwenden können. Sie können nach dem Zusammenführen der Strukturen auch die Option Farben für Vorschaukarte auswählen aktivieren, um sich zu vergewissern, dass die Ergebnisse Ihren Erwartungen entsprechen.

Nach dem Zusammenführen der Strukturen können Sie auch die Beschriftung für die neue Struktur verschieben. Dies ist in der Aufgabe *Strukturbeschriftungen bearbeiten* möglich. Für weitere Informationen siehe Thema Strukturbeschriftungen bearbeiten auf S. 298.

Dialogfeld "Zusammengeführte Struktur benennen"

Im Dialogfeld "Zusammengeführte Struktur benennen" können Sie der neu zusammengeführten Struktur Beschriftungen zuweisen.

In der Tabelle "Beschriftungen" werden Informationen für die einzelnen Schlüssel in der Kartendatei angezeigt. Außerdem können Sie dort den Schlüsseln jeweils eine Beschriftung zuweisen.

Neue Beschriftung. Geben Sie eine neue Beschriftung für die zusammengeführte Struktur ein, die dem betreffenden Kartenschlüssel zugewiesen werden soll.

Schlüssel. Der Kartenschlüssel, dem Sie die neue Beschriftung zuweisen.

Alte Beschriftungen. Die Beschriftungen für die Strukturen, die zu der neuen Struktur zusammengeführt werden.

Grenzen zwischen sich berührenden Polygonen entfernen. Aktivieren Sie diese Option, um die Grenzen aus den zusammengeführten Strukturen zu entfernen. Wenn Sie beispielsweise Bundesländer zu geografischen Regionen zusammengeführt haben, können Sie mit dieser Option die Grenzen zwischen den einzelnen Bundesländern entfernen.

Strukturen verschieben

Sie können Strukturen in der Karte verschieben. Dies kann nützlich sein, wenn Sie Strukturen zusammenbringen möchten, beispielsweise das Festlandsterritorium eines Landes und die zugehörigen abgelegenen Inseln.

Schlüssel. Wählen Sie den Kartenschlüssel aus, der die Strukturbeschriftungen enthält, mit denen Sie die zu verschiebenden Strukturen identifizieren können.

Strukturen. Klicken Sie auf die zu verschiebende Struktur. Beachten Sie, dass die Struktur in der Kartenvorschau ausgewählt wird. Sie können auch direkt in der Kartenvorschau auf die Struktur klicken.

X/Y. Diese Textfelder geben den aktuellen Mittelpunkt der Struktur in der Karte an. Die Einheiten werden in den Koordinaten der Karte angezeigt. Dabei kann es sich um lokale kartesische Koordinaten handeln (z. B. das State Plane Coordinate System für die USA) oder um geografische Koordinaten (wobei `X` den Längengrad und `Y` den Breitengrad angibt). Geben Sie die Koordinaten für die neue Position der Struktur an. Sie können Strukturen auch durch Klicken und Ziehen auf der Karte verschieben. Die Textfelder werden entsprechend der neuen Position aktualisiert.

Strukturen löschen

Sie können unerwünschte Strukturen aus der Karte löschen. Dies kann nützlich sein, wenn Sie die Karte überschaubarer gestalten möchten, indem Sie Strukturen löschen, die bei der Kartenvisualisierung nicht von Belang sind.

Schlüssel. Wählen Sie den Kartenschlüssel aus, der die Strukturbeschriftungen enthält, mit denen Sie die zu löschenden Strukturen identifizieren können.

Strukturen. Klicken Sie auf die zu löschende Struktur. Wenn Sie mehrere Strukturen gleichzeitig löschen möchten, klicken Sie mit gedrückter Strg-Taste auf die weiteren Strukturen. Beachten Sie, dass die Strukturen auch in der Kartenvorschau ausgewählt werden. Neben der Auswahl aus der Liste haben Sie auch die Möglichkeit, direkt in der Kartenvorschau auf die Strukturen zu klicken und dann beim Klicken die Strg-Taste gedrückt zu halten.

Einzelne Elemente löschen

Neben dem Löschen ganzer Strukturen können Sie einige der einzelnen Elemente löschen, aus denen sich die Strukturen zusammensetzen, beispielsweise Seen und kleine Inseln. Diese Option steht nicht für Punktkarten zur Verfügung.

Elemente. Klicken Sie auf die zu löschenden Elemente. Wenn Sie mehrere Elemente gleichzeitig löschen möchten, klicken Sie mit gedrückter Strg-Taste auf die weiteren Elemente. Beachten Sie, dass die Elemente auch in der Kartenvorschau ausgewählt werden. Neben der Auswahl aus der Liste haben Sie auch die Möglichkeit, direkt in der Kartenvorschau auf die Elemente zu klicken und dann beim Klicken die Strg-Taste gedrückt zu halten. Da die Liste der Elementnamen nicht selbsterklärend ist (den einzelnen Elementen wird jeweils eine Nummer innerhalb der Struktur zugewiesen), sollten Sie die Auswahl in der Kartenvorschau überprüfen, um sich zu vergewissern, dass Sie die gewünschten Elemente ausgewählt haben.

Projektion festlegen

Die Kartenprojektion gibt an, wie die dreidimensionale Erde in zwei Dimensionen dargestellt wird. Projektionen verursachen stets Verzerrungen. Allerdings sind, je nachdem, ob eine Weltkarte betrachtet wird oder eine regional begrenzte Karte, einige Projektionen besser geeignet als andere. Außerdem wird bei einigen Projektionen die Form der ursprünglichen Strukturen beibehalten. Projektionen, bei denen die Form beibehalten wird, sind konforme Projektionen. Diese Option steht nur für Karten mit geografischen Koordinaten (Längen- und Breitengrade) zur Verfügung.

Im Gegensatz zu anderen Optionen im Dienstprogramm zur Konvertierung von Karten kann die Projektion auch nach der Erstellung einer Kartenvisualisierung geändert werden.

Projektion. Wählen Sie eine Kartenprojektion aus. Bei der Erstellung einer Weltkarte oder einer Karte für eine Erdhalbkugel sollten Sie die Projektionstypen *Lokal*, *Mercator* oder *Winkel-Tripel* verwenden. Für kleinere Gebiete sollten Sie die Projektionstypen *Lokal*, *Lambert, konisch, konform* oder *Mercator, diagonal* verwenden. Bei allen Projektionen wird der WGS83-Ellipsoid als Bezugshöhe verwendet.

- Die Projektion vom Typ **Lokal** wird immer dann verwendet, wenn die Karte mit einem lokalen Koordinatensystem erstellt wurde, beispielsweise dem State Plane Coordinate System für die USA. Diese Koordinatensysteme werden statt durch geografische Koordinaten (Längen- und Breitengrade) durch kartesische Koordinaten definiert. Beim Projektionstyp “Lokal” befinden sich die horizontalen und vertikalen Linien in gleichmäßigen Abständen in einem kartesischen Koordinatensystem. Projektionen vom Typ “Lokal” sind nicht konform.
- Die Projektion vom Typ **Mercator** ist eine konforme Projektion für Weltkarten. Die horizontalen und vertikalen Linien sind gerade und stehen immer im rechten Winkel zueinander. Beachten Sie, dass sich die Mercator-Projektion ins Unendliche ausdehnt, wenn sie sich dem Nord- bzw. Südpol nähert. Daher kann sie nicht verwendet werden, wenn auf der Karte der Nord- bzw. Südpol enthalten ist. Die Verzerrung wird umso größer, je mehr sich die Karte diesen Grenzen nähert.
- Die Projektion vom Typ **Winkel-Tripel** ist eine nicht konforme Projektion für Weltkarten. Sie ist zwar nicht konform, bietet jedoch einen guten Ausgleich zwischen Form und Größe. Abgesehen von Äquator und Nullmeridian sind alle Linien gekrümmt. Wenn auf Ihrer Weltkarte der Nord- bzw. Südpol enthalten ist, ist dies eine gute Wahl für die Projektion.
- Wie der Name andeutet, handelt es sich beim Projektionstyp **Lambert, konisch, konform** (Lambertsche Schnittkegelprojektion) um eine konforme Projektion. Diese wird für Karten von Kontinenten oder kleineren Landmassen verwendet, deren Ost-West-Ausdehnung größer ist als die Nord-Süd-Ausdehnung.
- Der Projektionstyp **Mercator, diagonal** ist eine weitere konforme Projektion für Karten von Kontinenten oder kleineren Landmassen. Diese Projektion eignet sich besonders für Landmassen, bei denen die Nord-Süd-Ausdehnung größer ist als die Ost-West-Ausdehnung.

Schritt 4 - Fertigstellen

Nun können Sie einen Kommentar hinzufügen, der die Kartendatei beschreibt, und eine Beispieldatendatei aus den Kartenschlüsseln erstellen.

Kartenschlüssel. Wenn die Kartendatei mehrere Schlüssel enthält, wählen Sie den Kartenschlüssel aus, dessen Strukturbeschriftungen in der Vorschau angezeigt werden soll. Wenn Sie eine Datendatei aus der Karte erstellen, werden diese Beschriftungen als Datenwerte verwendet.

Kommentar. Geben Sie einen Kommentar ein, der die Karte beschreibt oder zusätzliche Informationen angibt, die für Ihre Benutzer relevant sein könnten, beispielsweise die Quellen für die ursprünglichen Shapefiles. Der Kommentar wird im Verwaltungssystem der Grafiktafel-Vorlagenauswahl angezeigt.

Aus Strukturbeschriftungen Datenset erstellen. Aktivieren Sie diese Option, wenn Sie eine Textdatendatei aus den angezeigten Strukturbeschriftungen erstellen möchten. Durch Klicken auf Durchsuchen… können Sie einen Speicherort und einen Dateinamen angeben. Wenn Sie die Erweiterung *.txt* hinzufügen, wird die Datei als Datei mit tabulatorgetrennten Werten gespeichert. Wenn Sie die Erweiterung *.csv* hinzufügen, wird die Datei als Datei mit kommagetrennten Werten gespeichert. Wenn Sie keine Angabe machen, wird standardmäßig die Erweiterung CSV verwendet.

Verteilen der Kartendateien

Im ersten Schritt des Dienstprogramms zur Konvertierung von Karten haben Sie ein Verzeichnis zur Speicherung der konvertierten SMZ-Datei ausgewählt. Außerdem haben Sie möglicherweise ausgewählt, dass die Karte zum Verwaltungssystem der Grafiktafel-Vorlagenauswahl hinzugefügt werden soll. Wenn Sie sich für die Speicherung im Verwaltungssystem entschieden haben, steht Ihnen die Karte in jedem IBM SPSS-Produkt zur Verfügung, das auf demselben Computer ausgeführt wird.

Um die Karte an andere Benutzer zu verteilen, müssen Sie diesen die SMZ-Datei zukommen lassen. Die Benutzer können dann mit dem Verwaltungssystem die Karte importieren. Sie können einfach die Datei senden, deren Speicherort Sie in Schritt 1 angegeben haben. Wenn Sie eine Datei senden möchten, die sich im Verwaltungssystem befindet, müssen Sie sie zuerst exportieren:

- Klicken Sie in der Vorlagenauswahl auf Verwalten...
- Klicken Sie auf die Registerkarte “Karte”.
- Wählen Sie die Karte aus, die Sie verteilen möchten.
- Klicken Sie auf Exportieren... und wählen Sie ein Verzeichnis aus, in dem die Datei gespeichert werden soll.

Nun können Sie die Kartendatei an andere Benutzer senden. Die Benutzer müssen diesen Vorgang umkehren und die Karte in das Verwaltungssystem importieren.

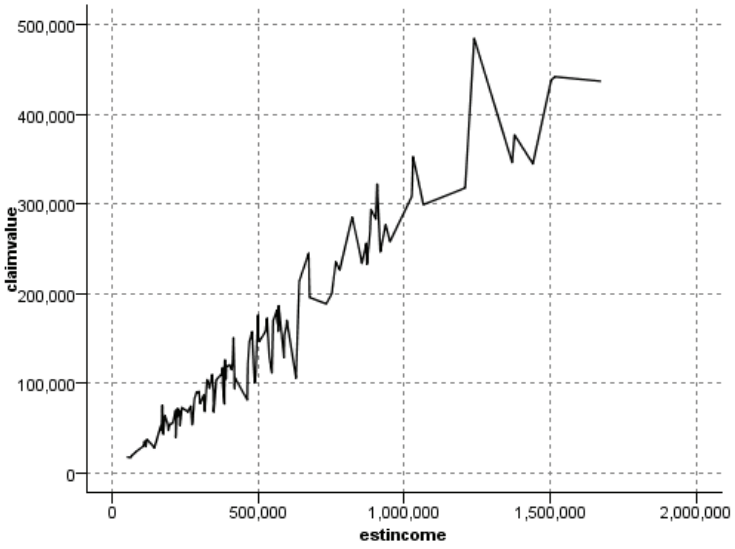
Plotknoten

Plotknoten zeigen die Beziehung zwischen numerischen Feldern. Sie können einen Plot mithilfe von Punkten (auch als Streudiagramm bezeichnet) oder mit Linien erstellen. Mit einem X-Modus im Dialogfeld stehen drei Arten von Linienplots zur Verfügung.

X-Modus = Sortieren

Beim X-Modus Sortieren werden die Daten nach den Werten für das Feld sortiert, das auf der *x*-Achse geplottet wird. So entsteht eine einzelne Linie, die von links nach rechts im Diagramm verläuft. Wenn Sie ein nominales Feld als Überlagerung verwenden, entstehen mehrere Linien mit verschiedenen Farbtönen, die von links nach rechts im Diagramm verlaufen.

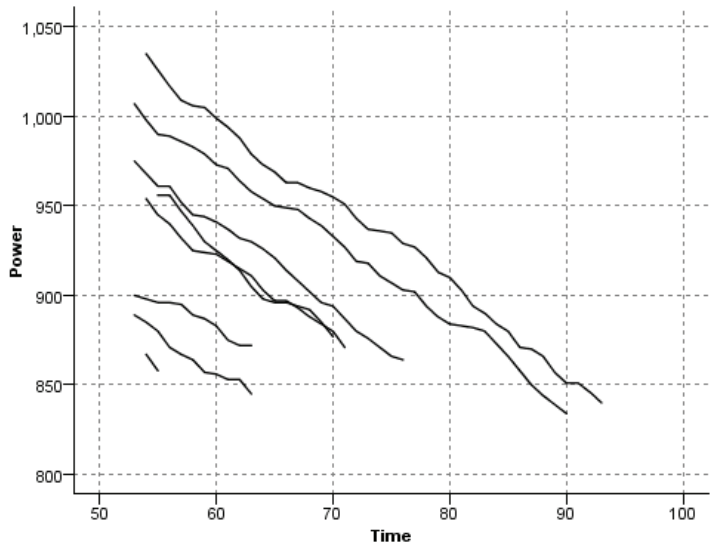
Abbildung 5-35
Linienplot mit X-Modus "Sortieren"



X-Modus = Überlagern

Beim X-Modus Überlagern werden mehrere Linienplots in einem einzigen Diagramm erstellt. Die Daten werden beim Überlagerungsplot nicht sortiert. Solange die Werte auf der *x*-Achse steigen, werden die Daten auf einer einzelnen Linie geplottet. Sobald die Werte fallen, beginnt eine neue Linie. Beispiel: Wenn *x* von 0 auf 100 steigt, werden die *y*-Werte auf einer einzelnen Linie aufgetragen. Sobald *x* unter 100 fällt, wird eine neue Linie zusätzlich zur ersten Linie geplottet. Der fertige Plot umfasst ggf. verschiedene Plots, mit denen mehrere Serien von *y*-Werten bequem miteinander verglichen werden können. Diese Art von Plot eignet sich für Daten mit einer Zeitraum-Komponente, z. B. für den Strombedarf über mehrere aufeinander folgende 24-Stunden-Zeiträume.

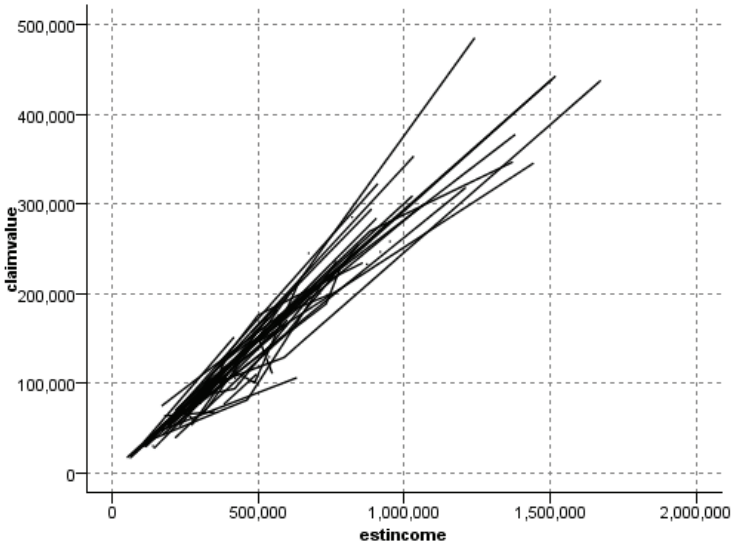
Abbildung 5-36
Linienplot mit X-Modus "Überlagern"



X-Modus = Wie gelesen

Beim X-Modus Wie gelesen werden die *x*- und *y*-Werte so geplottet, wie sie aus der Datenquelle gelesen wurden. Diese Option eignet sich für Daten mit einer Zeitreihen-Komponente, bei der Sie sich für Trends oder Muster interessieren, die sich aus der Reihenfolge der Daten ergeben. Unter Umständen sollten Sie die Daten sortieren, bevor Sie diese Art von Plot erstellen. Außerdem ist es möglich, zwei ähnliche Plots mit dem X-Modus Sortieren und Wie gelesen zu vergleichen, um so zu ermitteln, inwieweit ein Muster von der Sortierung abhängig ist.

Abbildung 5-37
Linienplot, oben mit X-Modus "Sortieren" dargestellt, nun erneut mit X-Modus "Wie gelesen" ausgeführt

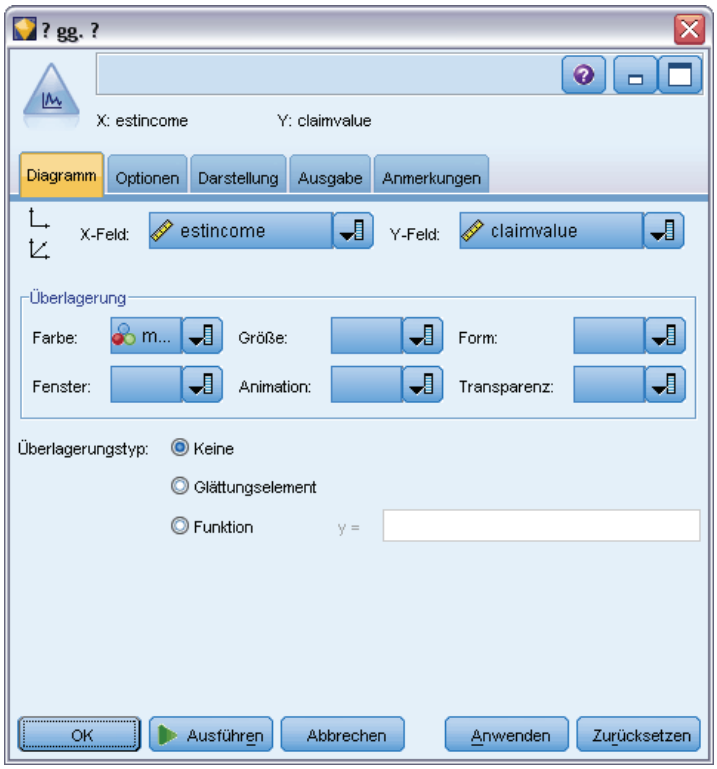


Sie können auch den Diagrammtafelknoten zur Erstellung von Streudiagrammen und Liniendiagrammen verwenden. In diesem Knoten stehen jedoch mehr Optionen zur Auswahl. Für weitere Informationen siehe Thema Verfügbare integrierte -Grafiktafel-Visualisierungstypen auf S. 262.

Registerkarte des Plotknotens

Plots zeigen die Werte eines *Y*-Felds gegen die Werte eines *X*-Felds. Häufig entsprechen diese Felder einer abhängigen Variablen bzw. einer unabhängigen Variablen.

Abbildung 5-38
Einstellungen auf der Registerkarte "Plot" für Plotknoten



X-Feld. Wählen Sie ein Feld in der Liste aus, das auf der horizontalen *x*-Achse dargestellt werden soll.

Y-Feld. Wählen Sie ein Feld in der Liste aus, das auf der vertikalen *y*-Achse dargestellt werden soll.

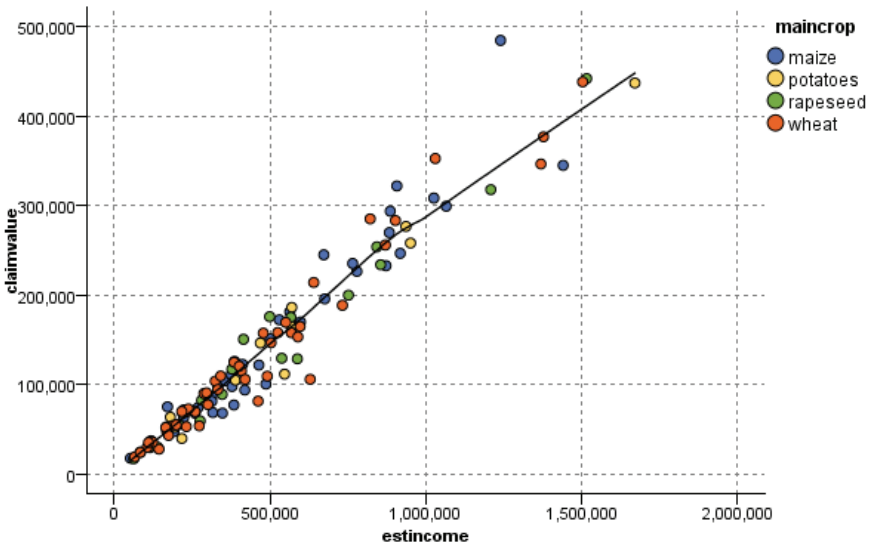
Z-Feld. Wenn Sie auf die Schaltfläche für das 3-D-Diagramm klicken, steht ein Feld in der Liste zur Auswahl, das auf der *z*-Achse dargestellt werden kann.

Überlagerung. Die Kategorien für die Datenwerte können auf unterschiedliche Weise dargestellt werden. Verwenden Sie beispielsweise *maincrop* (Hauptfeldfrucht) als Farbüberlagerung, um so die Werte *estincome* (Geschätztes Einkommen) und *claimvalue* (Förderungswert) für die Hauptfeldfrucht darzustellen, die von den Anspruchstellern gezogen wird. Für weitere Informationen siehe Thema Formatierungen, Überlagerungen, Fenster und Animation auf S. 246.

Überlagerungstyp. Gibt an, ob eine Überlagerungsfunktion oder ein Glättungselement angezeigt werden soll. Die Smoother-(Glättungs-) und die Überlagerungsfunktion werden immer als Funktion von y berechnet.

- **Keine.** Es wird keine Überlagerung angezeigt.
- **Glättungselement.** Zeigt eine geglättete Anpassungslinie an, die mithilfe einer Regression mit lokal gewichteten iterativen robusten kleinsten Quadraten (LOESS) berechnet wurde. Bei dieser Methode wird im Grunde eine Reihe von Regressionen berechnet, wobei sich jede auf einen kleinen Bereich innerhalb des Plots konzentriert. Dies führt zu einer Reihe "lokaler" Regressionslinien, die anschließend zu einer glatten Kurve zusammengefügt werden.

Abbildung 5-39
Plot mit LOESS-Smoother-Überlagerung



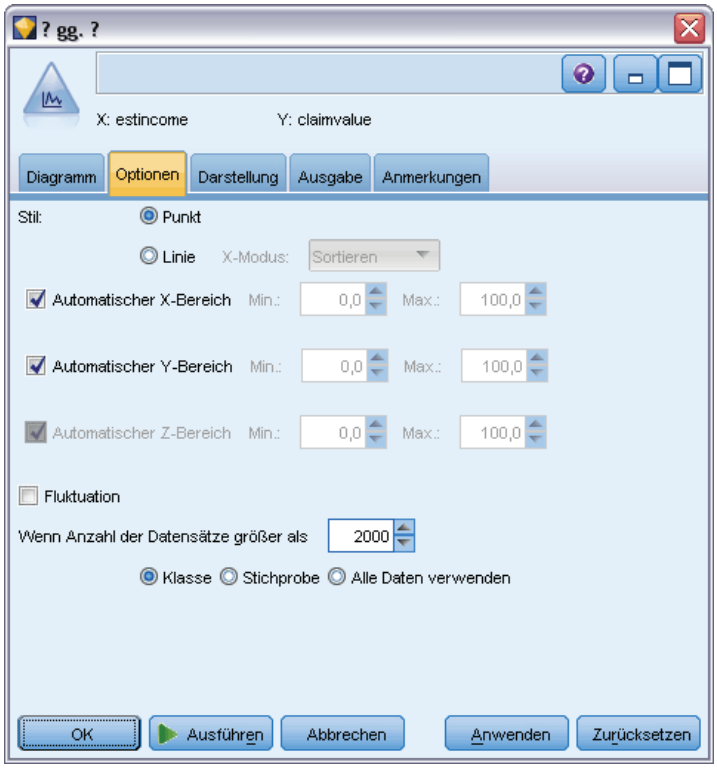
- **Funktion.** Hiermit geben Sie eine bekannte Funktion an, die mit tatsächlichen Werten verglichen werden soll. Um beispielsweise die Istwerte mit den vorhergesagten Werten zu vergleichen, plotten Sie die Funktion y = x als Überlagerung. Geben Sie eine Funktion für y = im Textfeld an. Die Standardfunktion lautet y = x; Sie können jedoch auch andere Funktionen festlegen, z. B. quadratische Funktionen oder beliebige Ausdrücke im Hinblick auf x.

  Hinweis: Überlagerungsfunktionen sind für Fenster- und Animationsdiagramme nicht verfügbar.

Sobald Sie die Optionen für einen Plot festgelegt haben, können Sie den Plot direkt aus dem Dialogfeld heraus starten. Klicken Sie hierzu auf Ausführen. Auf der Registerkarte “Optionen” können Sie jedoch zusätzliche Optionen angeben, z. B. Klassieren, X-Modus oder Stil.

Plot – Registerkarte "Optionen"

Abbildung 5-40
Einstellungen auf der Registerkarte "Optionen" für Plotknoten



Stil. Wählen Sie Punkt oder Linie als Plotstil aus. Bei Auswahl von Linie wird das Steuerelement X-Modus aktiviert. Bei Auswahl von Punkt wird ein Pluszeichen (+) als Standard-Punktform verwendet. Nach der Erstellung des Diagramms können Sie die Punktform und ihre Größe ändern.

X-Modus. Bei Linienplots definieren Sie den Stil mithilfe eines X-Modus. Wählen Sie Sortieren, Überlagern oder Wie gelesen. Bei Überlagerung und Wie gelesen sollten Sie eine maximal zulässige Daten-Set-Größe angeben, die für die Stichprobennahme der ersten *n* Datensätze verwendet wird. Ansonsten werden die standardmäßig festgelegten 2,000 Datensätze verwendet.

Automatischer X-Bereich. Hiermit geben Sie an, dass der gesamte Wertebereich in den Daten entlang dieser Achse verwendet werden soll. Um nur eine explizite Untergruppe von Werten auf der Grundlage der angegebenen Werte für Min und Max zu verwenden, deaktivieren Sie diese Option. Geben Sie die gewünschten Werte ein oder stellen Sie sie mit den Pfeilen ein. Automatische Bereiche sind standardmäßig aktiviert, um so den raschen Aufbau der Diagramme zu gewährleisten.

Automatischer Y-Bereich. Hiermit geben Sie an, dass der gesamte Wertebereich in den Daten entlang dieser Achse verwendet werden soll. Um nur eine explizite Untergruppe von Werten auf der Grundlage der angegebenen Werte für Min und Max zu verwenden, deaktivieren Sie diese Option. Geben Sie die gewünschten Werte ein oder stellen Sie sie mit den Pfeilen ein. Automatische Bereiche sind standardmäßig aktiviert, um so den raschen Aufbau der Diagramme zu gewährleisten.

Automatischer Z-Bereich. Nur wenn ein 3-D-Diagramm auf der Registerkarte "Plot" angegeben wurde. Hiermit geben Sie an, dass der gesamte Wertebereich in den Daten entlang dieser Achse verwendet werden soll. Um nur eine explizite Untergruppe von Werten auf der Grundlage der angegebenen Werte für Min und Max zu verwenden, deaktivieren Sie diese Option. Geben Sie die gewünschten Werte ein oder stellen Sie sie mit den Pfeilen ein. Automatische Bereiche sind standardmäßig aktiviert, um so den raschen Aufbau der Diagramme zu gewährleisten.

Fluktuation. Auch als **Bewegung** bezeichnet. Die Fluktuation eignet sich für Punktplots von Daten-Sets, in denen sich zahlreiche Werte wiederholen. Um eine deutlichere Verteilung der Werte zu erzielen, können Sie mit der Fluktuation die Punkte zufällig um den tatsächlichen Wert herum verteilen.

Hinweis für Benutzer früherer Versionen von ClementineSPSS Modeler: In dieser IBM® SPSS® Modeler-Version wird für den in einem Plot verwendeten Fluktuationswert eine andere Metrik verwendet. In früheren Versionen bestand der Wert aus einer tatsächlichen Zahl; nun wird ein Teil der Rahmengröße herangezogen. Dies bedeutet, dass die Bewegungswerte aus alten Streams wahrscheinlich zu groß sind. Bei dieser Version werden alle Bewegungswerte ungleich null durch den Wert 0,2 ersetzt.

Maximale Anzahl der Datensätze für Plot. Geben Sie eine Methode für das Plotten umfangreicher Daten-Sets an. Sie können wahlweise eine maximal zulässige Größe für das Daten-Set angeben oder den Standardwert von 2.000 Datensätzen verwenden. Bei umfangreichen Daten-Sets steigt die Leistung, wenn Sie die Option Klasse oder Stichprobe aktivieren. Alternativ können Sie mit Alle Daten verwenden alle Datenpunkte gleichzeitig plotten lassen; dies kann sich jedoch beträchtlich auf die Leistung der Software auswirken.

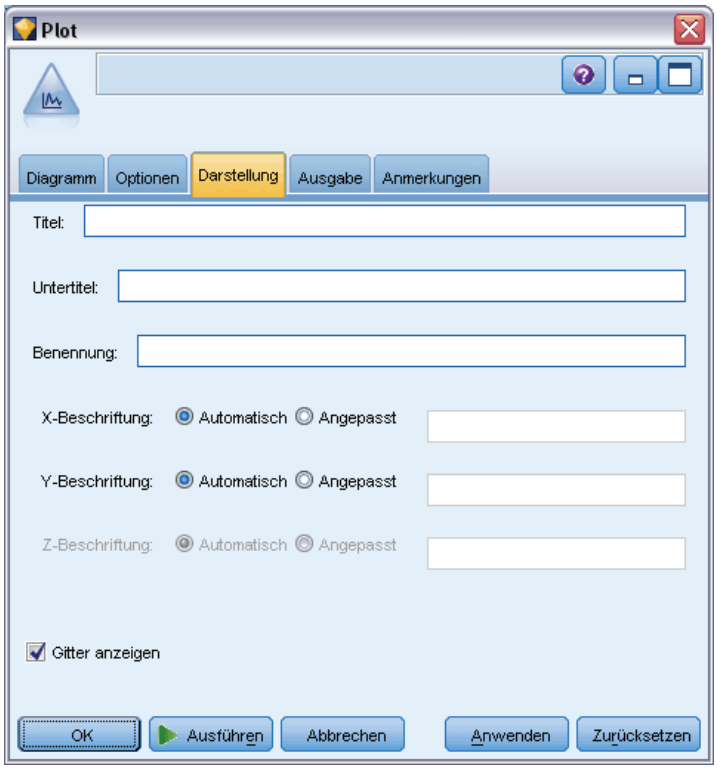
Hinweis: Beim X-Modus Überlagern oder Wie gelesen sind diese Optionen deaktiviert und es werden nur die ersten *n* Datensätze verwendet.

- **Klasse.** Hiermit aktivieren Sie die Klassierung, wenn das Daten-Set mehr Datensätze enthält als die angegebene Anzahl. Bei der Klassierung wird das Diagramm vor dem eigentlichen Plotten in feinmaschige Gitter aufgeteilt und es wird die Anzahl der Punkte gezählt, die in die einzelnen Gitterzellen fallen würden. Im endgültigen Diagramm wird je ein Punkt pro Zelle im Klassierschwerpunkt (Durchschnitt aller Punktpositionen in der Klasse) geplottet. Die Größe der geplotteten Symbole weist auf die Anzahl der Punkte im betreffenden Bereich hin (sofern Sie die Größe nicht als Überlagerung verwenden). Durch die Methode, den Schwerpunkt und die Größe zur Darstellung der Anzahl an Punkten heranzuziehen, eignet sich der klassierte Plot besonders gut für die Darstellung umfangreicher Daten-Sets, weil ein übermäßiges Plotten (unidentifizierbare Farbansammlungen) in dicht besetzten Bereichen vermieden wird und auch Symbolartefakte (künstliche Dichtemuster) verringert werden. Symbolartefakte treten auf, wenn bestimmte Symbole (insbesondere das Pluszeichen +) auf eine Weise kollidieren, bei der dichte Bereiche entstehen, die in den eigentlichen Rohdaten nicht vorhanden sind.
- **Beispiel.** Aus den Daten wird eine zufällige Stichprobe mit der im Textfeld eingegebenen Anzahl an Datensätzen zusammengestellt. Der Standardwert ist 2,000.

Plot – Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Abbildung 5-41
Einstellungen auf der Registerkarte "Darstellung" für einen Plotknoten



Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Benennung. Dient zur Eingabe des Texts, der zur Benennung des Diagramms verwendet werden soll.

X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *x*-Achsen-Beschriftung (horizontal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *y*-Achsen-Beschriftung (vertikal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

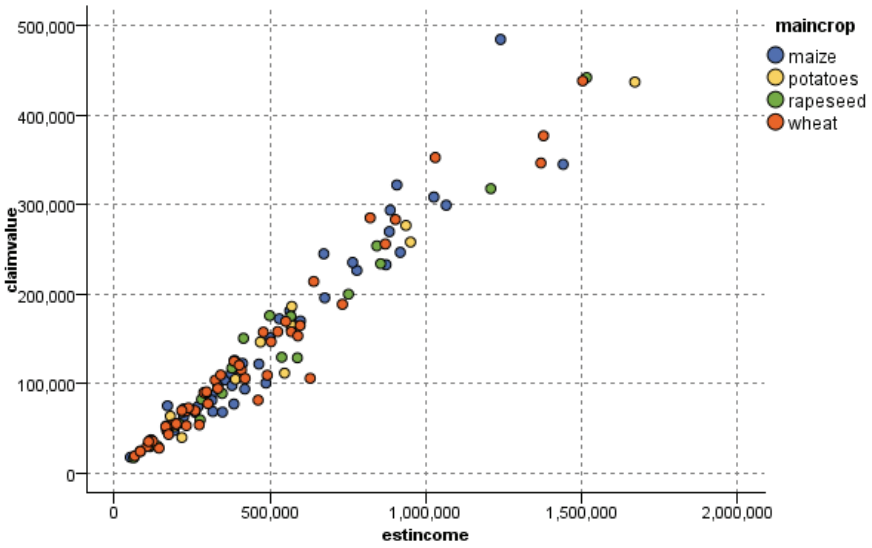
Z-Beschriftung. Nur bei 3-D-Diagrammen: Akzeptieren Sie entweder die automatisch generierte *z*-Achsen-Beschriftung oder wählen Sie Angepasst, um eine benutzerdefinierte Beschriftung anzugeben.

Gitter anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie ein Gitter hinter dem Plot oder dem Diagramm einblenden, das die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Das Gitter wird stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Verwendung eines Plotdiagramms

Plots und Multidiagramme sind im Grunde genommen Plots von *X* in Abhängigkeit von *Y*. Wenn Sie beispielsweise potenzielle Betrugsfälle in Bewerbungen um landwirtschaftliche Subventionen untersuchen (wie in *fraud.str* im Ordner *Demos* der IBM® SPSS® Modeler-Installation dargestellt), soll beispielsweise das auf der Bewerbung angegebene Einkommen in Abhängigkeit von dem Einkommen geplottet werden, das mithilfe eines neuronalen Netzes geschätzt wurde. Aus einer Überlagerung, z. B. dem Feldfruchttyp, geht hervor, ob eine Beziehung zwischen den Forderungen (Wert oder Anzahl) und der Art der Feldfrucht besteht.

Abbildung 5-42
Plot der Beziehung zwischen geschätztem Einkommen und Forderungswert mit Hauptfeldfruchttyp als Überlagerung



Plots, Multidiagramme und Evaluationsdiagramme sind zweidimensionale Darstellungen von *Y* gegen *X*. Die Arbeit mit diesen Diagrammen ist daher denkbar unkompliziert: Sie können ganz einfach Bereiche definieren, Elemente markieren und sogar Abschnitte einzeichnen. Außerdem können Sie Knoten für die durch diese Bereiche, Abschnitte bzw. Elemente dargestellten Daten generieren. Für weitere Informationen siehe Thema Untersuchen von Diagrammen auf S. 360.

Verteilungsknoten

Verteilungsdiagramme bzw. -tabellen zeigen das Auftreten symbolischer (nichtnumerischer) Werte, z. B. Hypothekenart oder Geschlecht, in einem Daten-Set. Verteilungsknoten eignen sich beispielsweise zum Aufzeigen von Unausgewogenheiten in den Daten, die mithilfe eines Balancierungsknotens vor dem Erstellen eines Modells ausgeglichen werden können. Über das Menü “Generieren” im Fenster eines Verteilungsdiagramms bzw. einer Verteilungstabelle können Sie automatisch einen Balancierungsknoten erzeugen lassen.

Sie können auch den Diagrammtafelknoten zur Erstellung von Diagrammen vom Typ “Balken für Häufigkeiten” verwenden. In diesem Knoten stehen jedoch mehr Optionen zur Auswahl. Für weitere Informationen siehe Thema Verfügbare integrierte -Grafiktafel-Visualisierungstypen auf S. 262.

Hinweis: Soll das Auftreten numerischer Werte aufgezeigt werden, verwenden Sie einen Histogrammknoten.

Verteilung – Registerkarte “Plot”

Abbildung 5-43
Einstellungen auf der Registerkarte “Plot” für Verteilungsknoten

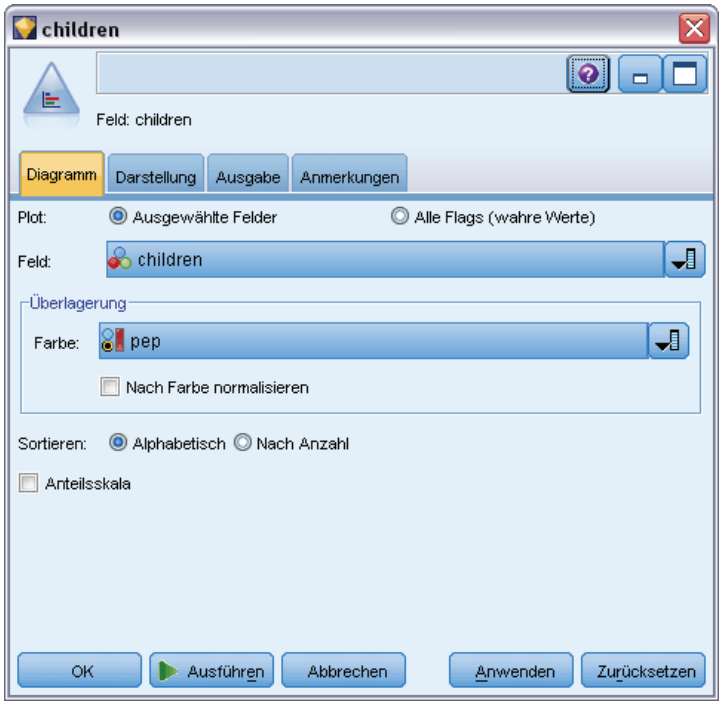


Diagramm. Wählen Sie den Typ der Verteilung aus. Mit Ausgewählte Felder lassen Sie die Verteilung für das ausgewählte Feld anzeigen. Mit Alle Flags (wahre Werte) rufen Sie die Verteilung der Wahr-Werte für die Flag-Felder im Daten-Set ab.

Feld. Wählen Sie ein nominales oder Flag-Feld aus, für das die Verteilung der Werte dargestellt werden soll. Die Liste enthält nur solche Felder, die nicht explizit als numerisch definiert wurden.

Überlagerung. Wählen Sie ein nominales oder Flag-Feld aus, das als Farbüberlagerung verwendet werden soll, um so die Verteilung der zugehörigen Werte innerhalb der einzelnen Werte für das angegebene Feld darzustellen. Mithilfe der Reaktionen auf eine Marketingkampagne (*pep*) als Überlagerung für die Anzahl der Kinder (*children*) können Sie beispielsweise die Ansprechbarkeit nach Familiengröße darstellen lassen. Für weitere Informationen siehe Thema Formatierungen, Überlagerungen, Fenster und Animation auf S. 246.

Nach Farbe normalisieren. Die Balken werden so skaliert, dass alle Balken die volle Breite des Diagramms einnehmen. Die Überlagerungswerte entsprechen einem Anteil jedes Balkens, sodass Vergleiche zwischen den Kategorien erleichtert werden.

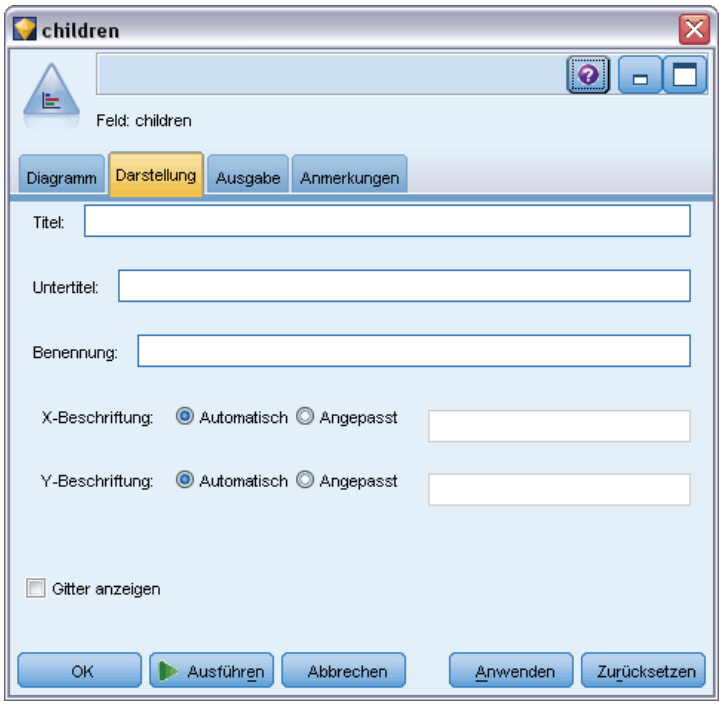
Sortieren. Wählen Sie die Methode aus, mit der die Werte im Verteilungsdiagramm dargestellt werden sollen. Mit der Option Alphabetisch werden die Werte in alphabetischer Reihenfolge angezeigt, mit Nach Anzahl dagegen absteigend nach der Anzahl der Vorkommen.

Anteilsskala. Die Verteilung der Werte wird so skaliert, dass der Wert mit der größten Anzahl die volle Breite des Plots einnimmt. Alle anderen Balken werden gemäß diesem Wert skaliert. Wenn Sie diese Option deaktivieren, werden die Balken gemäß der Gesamtanzahl der einzelnen Werte skaliert.

Verteilung – Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Abbildung 5-44
Registerkarte "Darstellung" - Einstellungen



Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Benennung. Dient zur Eingabe des Texts, der zur Benennung des Diagramms verwendet werden soll.

X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *x*-Achsen-Beschriftung (horizontal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *y*-Achsen-Beschriftung (vertikal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

Gitter anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie ein Gitter hinter dem Plot oder dem Diagramm einblenden, das die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Das Gitter wird stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Verwendung von Verteilungsknoten

Verteilungsknoten zeigen die Verteilung symbolischer Werte in einem Daten-Set. Diese Knoten werden häufig als Vorstufe für Bearbeitungsknoten eingesetzt, um die Daten zu untersuchen und eventuelle Unausgewogenheiten zu bereinigen. Wenn beispielsweise Instanzen mit Antwortenden ohne Kinder viel häufiger auftreten als andere Typen von Teilnehmern, können Sie diese Instanzen verringern, sodass in späteren Data Mining-Operationen eine nützlichere Regel aufgestellt werden kann. Mit einem Verteilungsknoten können Sie diese Unausgewogenheiten untersuchen und über die weitere Vorgehensweise entscheiden.

Der Verteilungsknoten ist dahingehend ungewöhnlich, dass er sowohl ein Diagramm als auch eine Tabelle zur Datenanalyse erstellt.

Abbildung 5-45
Verteilungsdiagramm zur Anzahl der Teilnehmer mit Kindern bzw. ohne Kinder, die auf eine Marketing-Kampagne reagiert haben

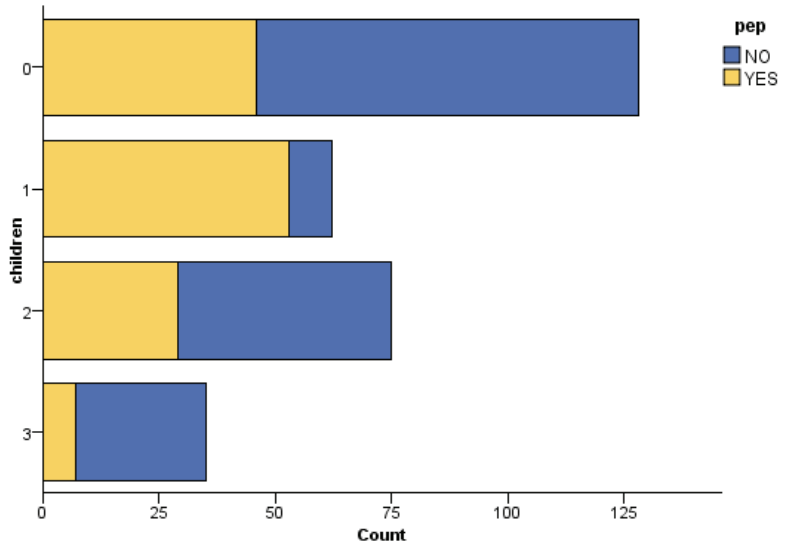
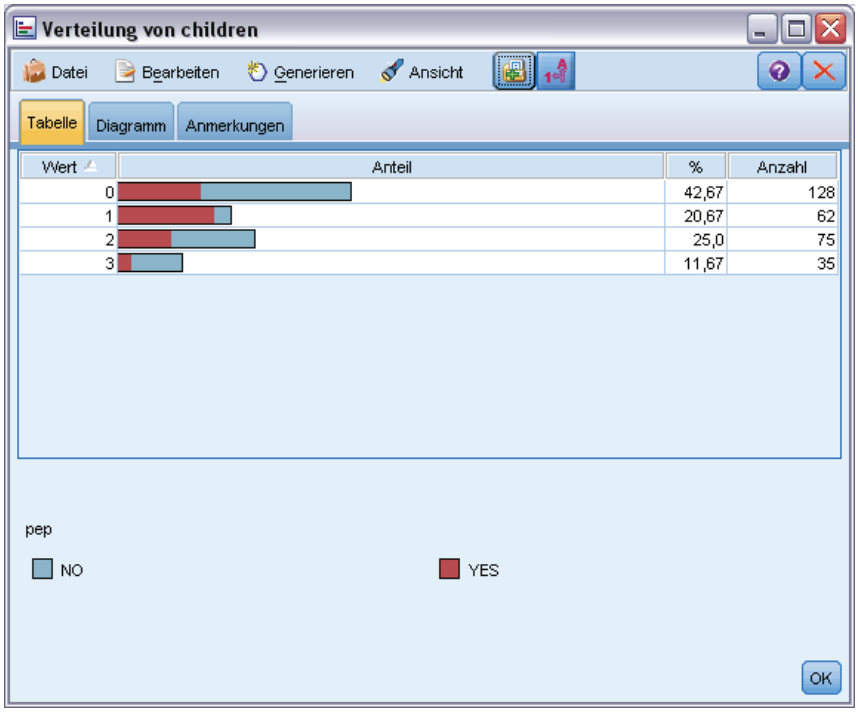


Abbildung 5-46
Verteilungstabelle zum Anteil der Teilnehmer mit Kindern bzw. ohne Kinder, die auf eine Marketing-Kampagne reagiert haben



Sobald Sie eine Verteilungstabelle und ein Verteilungsdiagramm erstellt und die Ergebnisse untersucht haben, können Sie mit den Optionen in den Menüs die Werte gruppieren, bestimmte Werte kopieren und eine Reihe von Knoten zur Datenvorbereitung erzeugen. Außerdem können Sie die Diagramm- und Tabelleninformationen zur Verwendung in anderen Anwendungen, wie beispielsweise MS Word oder MS PowerPoint, kopieren bzw. exportieren. Für weitere Informationen siehe Thema Drucken, Speichern, Kopieren und Exportieren von Diagrammen auf S. 395.

So können Sie Werte in einer Verteilungstabelle auswählen und kopieren:

- Klicken Sie mit der Maus, halten Sie die Maustaste gedrückt und wählen Sie eine Reihe von Werten durch Ziehen aus. Sie können auch mit dem Befehl Alles auswählen im Menü “Bearbeiten” alle Werte gleichzeitig auswählen.

- Wählen Sie im Menü “Bearbeiten” die Option Tabelle kopieren oder Tabelle kopieren (einschl. Feldnamen).

- Übernehmen Sie die Daten in die Zwischenablage oder fügen Sie sie in die gewünschte Anwendung ein.

  Hinweis: Die Balken werden nicht direkt kopiert. Stattdessen werden die Tabellenwerte kopiert. In der kopierten Tabelle werden also keine überlagerten Werte dargestellt.

So gruppieren Sie Werte aus einer Verteilungstabelle:

- Wählen Sie mehrere Werte zur Gruppierung mit Strg+Klicken aus.
- Wählen Sie im Menü “Bearbeiten” die Option Gruppieren.

Hinweis: Beim Gruppieren von Werten bzw. beim Aufheben der Gruppierung wird das Diagramm auf der Registerkarte “Diagramm” automatisch unter Berücksichtigung der Änderungen neu erstellt.

Weitere Möglichkeiten:

- Heben Sie die Gruppierung der Werte auf. Wählen Sie hierzu den Namen der Gruppe in der Verteilungsliste aus und wählen Sie im Menü “Bearbeiten” den Befehl Gruppierung aufheben.
- Bearbeiten Sie die Gruppen. Wählen Sie hierzu den Namen der Gruppe in der Verteilungsliste aus und wählen Sie im Menü “Bearbeiten” den Befehl Gruppe bearbeiten. Ein Dialogfeld wird geöffnet, in dem Sie die Werte in die Gruppe hinein und aus dieser hinaus verschieben können.

Abbildung 5-47
Dialogfeld “Gruppe bearbeiten”

Optionen im Menü “Generieren”

Mit den Optionen im Menü “Generieren” können Sie eine Teilgruppe mit Daten auswählen, ein Flag-Feld ableiten oder die Daten aus einem Diagramm bzw. einer Tabelle balancieren. Bei diesen Funktionen wird ein Datenvorbereitungsknoten erzeugt und in den Stream-Zeichenbereich platziert. Um den erzeugten Knoten nutzen zu können, verbinden Sie ihn mit einem vorhandenen Stream. Für weitere Informationen siehe Thema Generieren von Knoten aus Diagrammen auf S. 370.

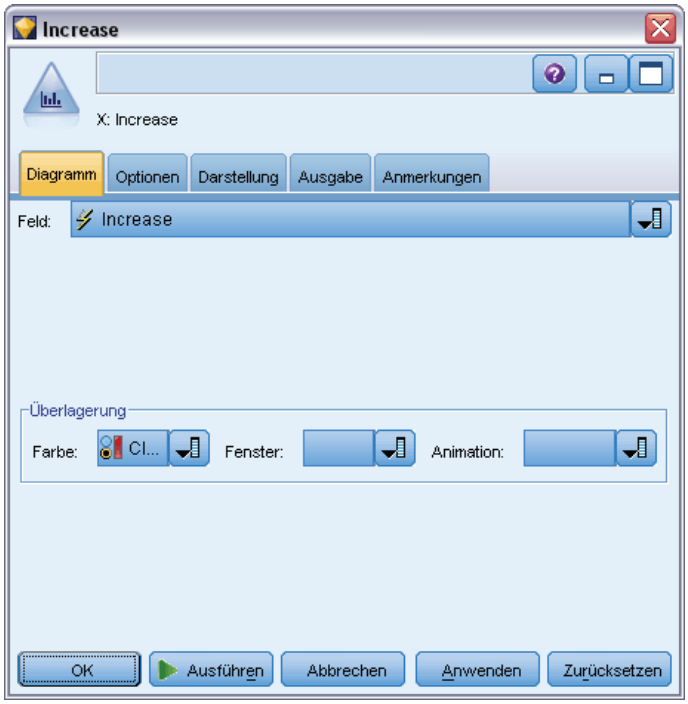
Histogrammknoten

Histogrammknoten zeigen das Auftreten bestimmter Werte in numerischen Feldern. Hiermit werden häufig die Daten vor der weiteren Bearbeitung und der Modellerstellung untersucht. Ähnlich wie Verteilungsknoten werden Histogrammknoten oft dazu herangezogen, Unausgewogenheiten in den Daten zu erkennen. Sie können Histogramme zwar auch mit dem Diagrammtafelknoten erstellen, in diesem Knoten stehen Ihnen jedoch mehr Optionen zur Auswahl. Für weitere Informationen siehe Thema Verfügbare integrierte -Grafiktafel-Visualisierungstypen auf S. 262.

Hinweis: Soll das Auftreten von Werten für symbolische Felder aufgezeigt werden, verwenden Sie einen Verteilungsknoten.

Histogramm – Registerkarte "Plot"

Abbildung 5-48
Einstellungen auf der Registerkarte "Plot" für Histogrammknoten

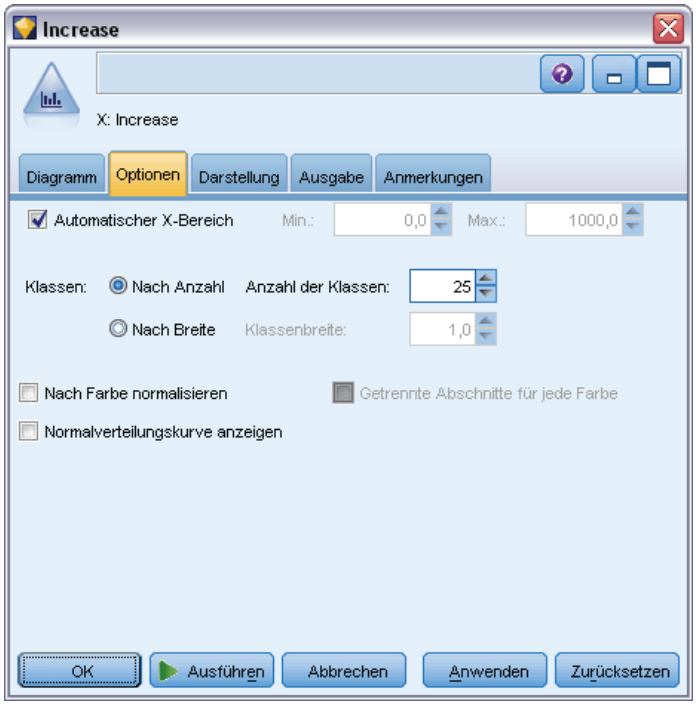


Feld. Wählen Sie ein numerisches Feld aus, für das die Verteilung der Werte dargestellt werden soll. Die Liste enthält nur solche Felder, die nicht explizit als symbolisch (kategorial) definiert wurden.

Überlagerung. Wählen Sie ein symbolisches Feld aus, mit dem die Kategorien der Werte für das angegebene Feld dargestellt werden sollen. Bei einem Überlagerungsfeld wird das Histogramm in ein Stapeldiagramm umgewandelt, bei dem die verschiedenen Kategorien des Überlagerungsfelds mithilfe von Farben gekennzeichnet sind. Bei Verwendung des Histogrammknotens gibt es drei Überlagerungstypen: Farbe, Fenster und Animation. Für weitere Informationen siehe Thema Formatierungen, Überlagerungen, Fenster und Animation auf S. 246.

Histogramm – Registerkarte "Optionen"

Abbildung 5-49
Einstellungen auf der Registerkarte "Optionen" für Histogrammknoten



Automatischer X-Bereich. Hiermit geben Sie an, dass der gesamte Wertebereich in den Daten entlang dieser Achse verwendet werden soll. Um nur eine explizite Untergruppe von Werten auf der Grundlage der angegebenen Werte für Min und Max zu verwenden, deaktivieren Sie diese Option. Geben Sie die gewünschten Werte ein oder stellen Sie sie mit den Pfeilen ein. Automatische Bereiche sind standardmäßig aktiviert, um so den raschen Aufbau der Diagramme zu gewährleisten.

Klassen. Wählen Sie entweder Nach Anzahl oder Nach Breite.

- Mit der Option Nach Anzahl lassen Sie eine feste Anzahl von Balken anzeigen, deren Breite vom angegebenen Bereich und der angegebenen Anzahl an Buckets abhängig ist. Geben Sie in der Option Anzahl der Klassen an, wie viele Klassen im Diagramm verwendet werden sollen. Mithilfe der Pfeile können Sie die Anzahl einstellen.
- Mit Nach Breite erstellen Sie ein Diagramm, dessen Balken eine feste Breite besitzen. Die Anzahl der Klassen ergibt sich aus der festgelegten Breite und dem Wertebereich. Geben Sie in der OptionKlassenbreite die Breite der Balken an.

Nach Farbe normalisieren. Alle Balken werden auf dieselbe Höhe gebracht. Überlagerte Werte werden dabei als Prozentsatz der Gesamtanzahl an Fällen in jedem Balken dargestellt.

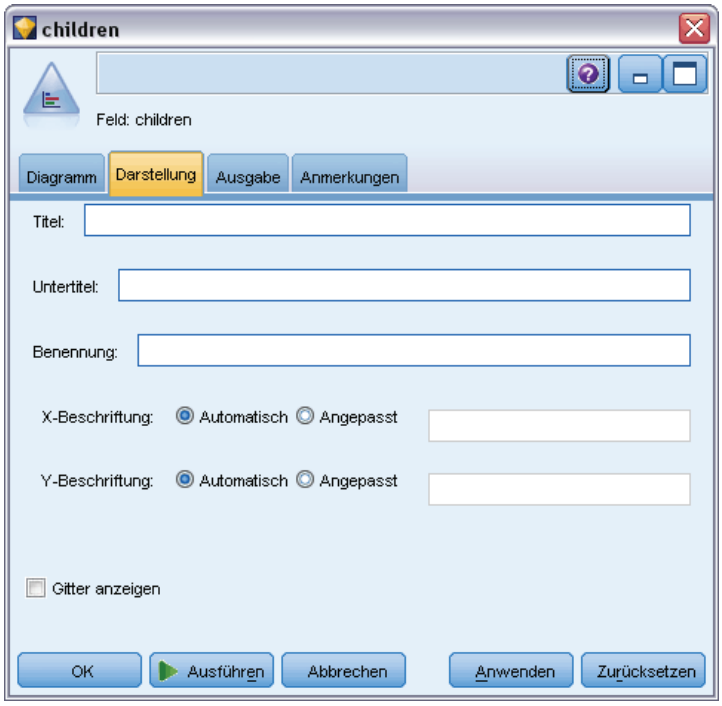
Normalverteilungskurve anzeigen. Wählen Sie diese Option aus, um eine Normalverteilungskurve in das Diagramm aufzunehmen, die Mittelwert und Varianz der Daten anzeigt.

Getrennte Abschnitte für jede Farbe. Jeder überlagerte Wert wird als getrennter Abschnitt im Diagramm dargestellt.

Histogramm – Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Abbildung 5-50
Einstellungen auf der Registerkarte "Darstellung" für die meisten Diagrammknoten



Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Benennung. Dient zur Eingabe des Texts, der zur Benennung des Diagramms verwendet werden soll.

X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *x*-Achsen-Beschriftung (horizontal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *y*-Achsen-Beschriftung (vertikal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

Gitter anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie ein Gitter hinter dem Plot oder dem Diagramm einblenden, das die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Das Gitter wird stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

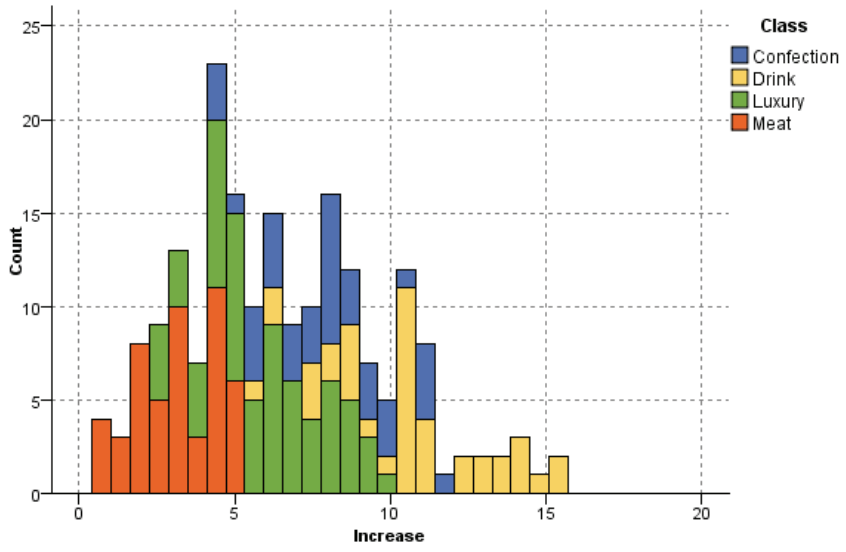
Histogramme

Histogramme zeigen die Verteilung der Werte in einem numerischen Feld, dessen Werte an der *x*-Achse dargestellt werden. Histogramme funktionieren ähnlich wie Sammlungsdiagramme. Bei Sammlungen wird die Verteilung der Werte für ein numerisches Feld *relativ zu den Werten eines anderen Felds* dargestellt, also nicht das Auftreten von Werten für ein einziges Feld.

Sobald Sie ein Diagramm erstellt haben, können Sie die Ergebnisse untersuchen und Abschnitte festlegen, um die Werte entlang der *x*-Achse aufzuspalten bzw. Regionen zu definieren. Außerdem können Sie Elemente innerhalb des Diagramms markieren. Für weitere Informationen siehe Thema Untersuchen von Diagrammen auf S. 360.

Mit Optionen im Menü “Generieren” können Sie Balancierungs- Auswahl- und Ableitungsknoten erstellen. Hierfür werden die Daten im Diagramm bzw. genauer die Daten innerhalb bestimmter Abschnitte, Bereiche oder markierter Elemente verwendet. Dieser Diagrammtyp wird häufig als Vorbereitung auf Bearbeitungsknoten eingesetzt, um die Daten zu untersuchen und etwaige Unausgewogenheiten mithilfe eines Balancierungsknotens, der aus dem Diagrammknoten heraus erzeugt wird, auszugleichen. Darüber hinaus können Sie einen Flag-Ableitungsknoten erzeugen und so ein Feld hinzufügen, aus dem hervorgeht, in welchen Abschnitt die einzelnen Datensätze fallen, oder auch einen Auswahlknoten, mit dem Sie alle Datensätze in einem bestimmten Set oder Wertebereich auswählen. Diese Funktionen sorgen dafür, dass eine bestimmte Untergruppe an Daten zur näheren Untersuchung im Mittelpunkt verbleibt. Für weitere Informationen siehe Thema Generieren von Knoten aus Diagrammen auf S. 370.

Abbildung 5-51
Histogramm mit der Verteilung gesteigerter Käufe nach Kategorie aufgrund einer Werbeaktion

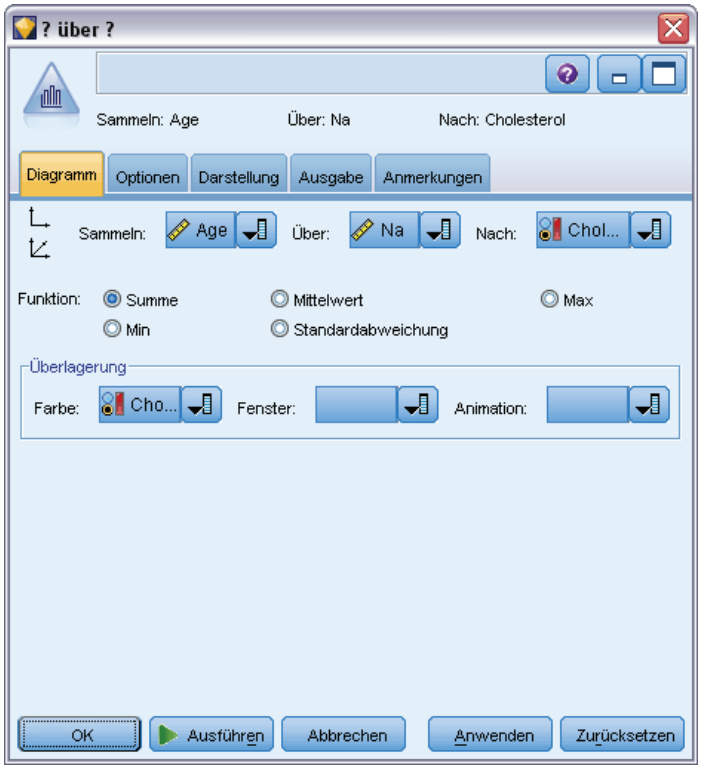


Sammlungsknoten

Sammlungen sind nahezu mit Histogrammen identisch, mit dem Unterschied, dass bei Sammlungen die Verteilung der Werte für ein numerisches Feld relativ zu den Werten eines anderen Felds dargestellt wird, also nicht das Auftreten von Werten für ein einziges Feld. Eine Sammlung eignet sich besonders für die Darstellung einer Variablen oder eines Felds, dessen Werte sich mit der Zeit verändern. Mithilfe eines 3-D-Diagramms können Sie außerdem eine symbolische Achse anlegen, auf der die Verteilungen nach Kategorie aufgetragen sind. Zweidimensionale Sammlungen werden als gestapelte Balkendiagramme (ggf. mit Überlagerungen) angezeigt. Für weitere Informationen siehe Thema Formatierungen, Überlagerungen, Fenster und Animation auf S. 246.

Sammlung – Registerkarte "Plot"

Abbildung 5-52
Einstellungen auf der Registerkarte "Plot" für Sammlungsknoten



Sammeln. Wählen Sie ein Feld aus, dessen Werte gesammelt und über den Wertebereich für das unter Über angegebene Feld dargestellt werden sollen. Die Liste enthält nur solche Felder, die nicht als symbolisch definiert sind.

Über. Wählen Sie ein Feld aus, dessen Werte für die Darstellung des unter Sammeln angegebenen Felds herangezogen werden sollen.

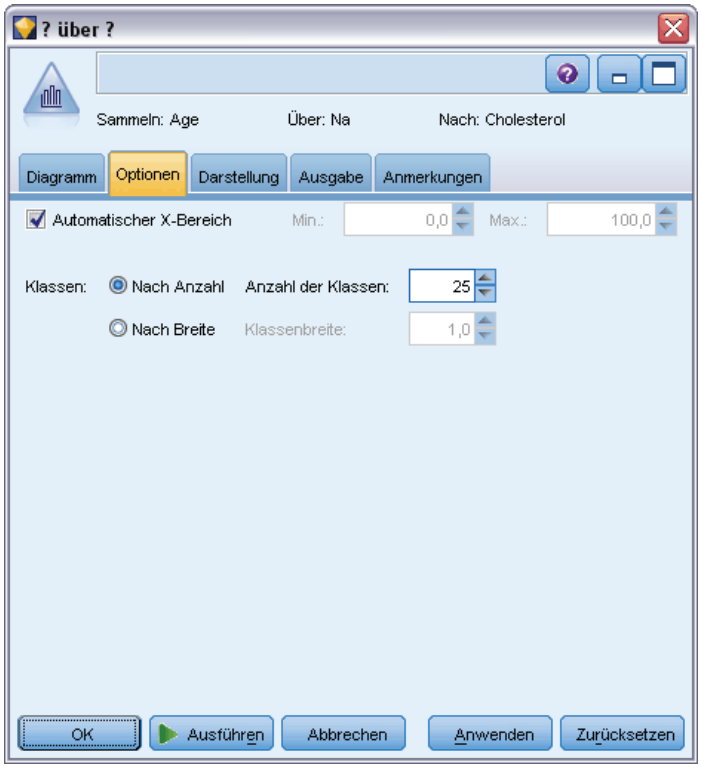
Nach. Mit dieser Option (beim Erstellen eines 3-D-Diagramms aktiviert) können Sie ein nominales oder Flag-Feld auswählen, mit dem das Sammlungsfeld nach Kategorien dargestellt werden soll.

Funktion. Legen Sie fest, was die einzelnen Balken im Sammlungsdiagramm enthalten sollen. Die folgenden Optionen stehen zur Auswahl: Summe, Mittelwert, Max, Min und Standardabweichung.

Überlagerung. Wählen Sie ein symbolisches Feld aus, mit dem die Kategorien der Werte für das ausgewählte Feld dargestellt werden sollen. Wenn Sie ein Überlagerungsfeld auswählen, wird die Sammlung umgewandelt und es entstehen mehrere Balken in verschiedenen Farben für die einzelnen Kategorien. Für diesen Knoten gibt es drei Überlagerungstypen: Farbe, Fenster und Animation. Für weitere Informationen siehe Thema Formatierungen, Überlagerungen, Fenster und Animation auf S. 246.

Sammlung – Registerkarte "Optionen"

Abbildung 5-53
Einstellungen auf der Registerkarte "Optionen" für Sammlungsknoten



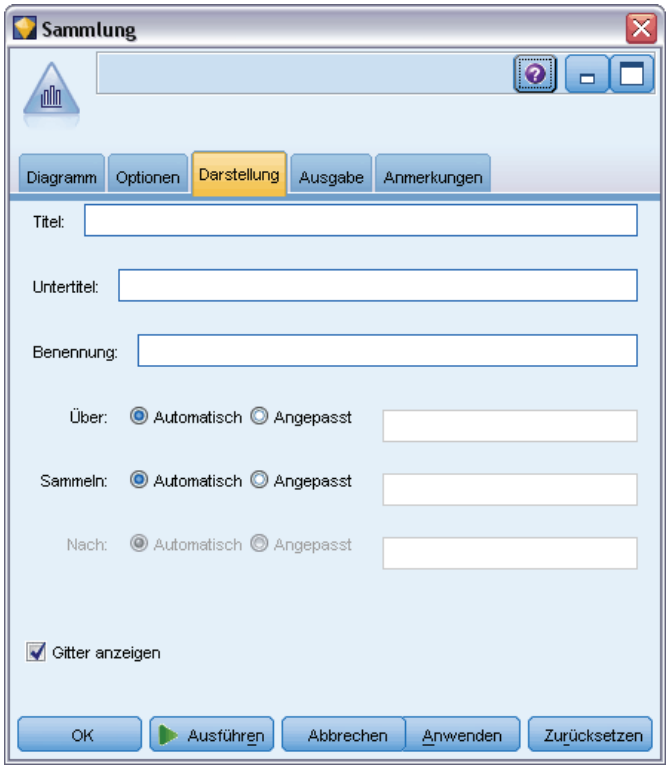
Automatischer X-Bereich. Hiermit geben Sie an, dass der gesamte Wertebereich in den Daten entlang dieser Achse verwendet werden soll. Um nur eine explizite Untergruppe von Werten auf der Grundlage der angegebenen Werte für Min und Max zu verwenden, deaktivieren Sie diese Option. Geben Sie die gewünschten Werte ein oder stellen Sie sie mit den Pfeilen ein. Automatische Bereiche sind standardmäßig aktiviert, um so den raschen Aufbau der Diagramme zu gewährleisten.

Klassen. Wählen Sie entweder Nach Anzahl oder Nach Breite.

- Mit der Option Nach Anzahl lassen Sie eine feste Anzahl von Balken anzeigen, deren Breite vom angegebenen Bereich und der angegebenen Anzahl an Buckets abhängig ist. Geben Sie in der Option Anzahl der Klassen an, wie viele Klassen im Diagramm verwendet werden sollen. Mithilfe der Pfeile können Sie die Anzahl einstellen.
- Mit Nach Breite erstellen Sie ein Diagramm, dessen Balken eine feste Breite besitzen. Die Anzahl der Klassen ergibt sich aus der festgelegten Breite und dem Wertebereich. Geben Sie in der OptionKlassenbreite die Breite der Balken an.

Sammlung – Registerkarte "Darstellung"

Abbildung 5-54
Einstellungen auf der Registerkarte "Darstellung" für Sammlungsknoten



Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Benennung. Dient zur Eingabe des Texts, der zur Benennung des Diagramms verwendet werden soll.

Beschriftung "Über". Akzeptieren Sie entweder die automatisch generierte Beschriftung oder wählen Sie Angepasst, um eine benutzerdefinierte Beschriftung anzugeben.

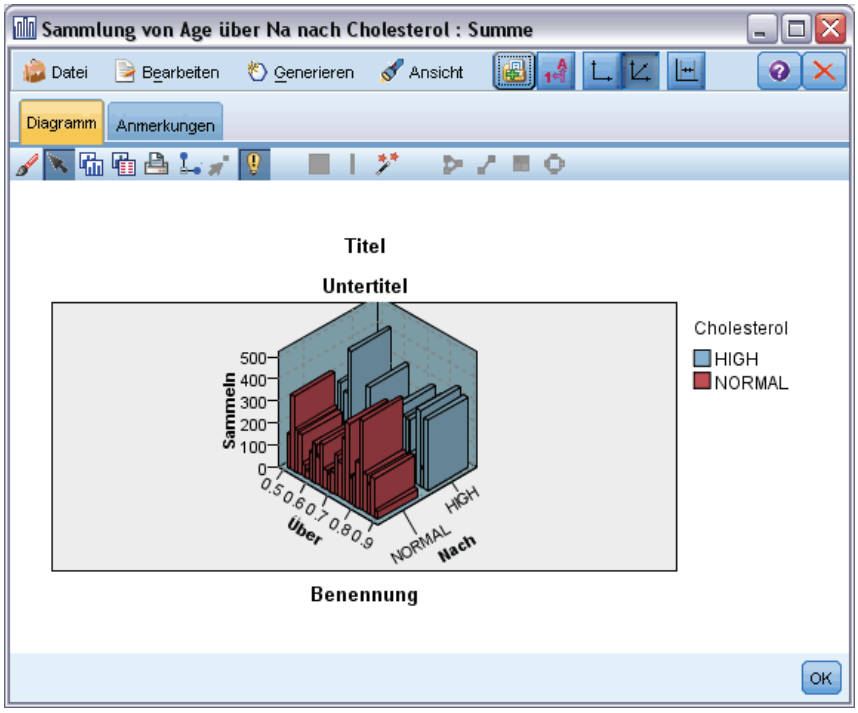
Beschriftung "Sammeln". Akzeptieren Sie entweder die automatisch generierte Beschriftung oder wählen Sie Angepasst, um eine benutzerdefinierte Beschriftung anzugeben.

Beschriftung "Nach". Akzeptieren Sie entweder die automatisch generierte Beschriftung oder wählen Sie Angepasst, um eine benutzerdefinierte Beschriftung anzugeben.

Gitter anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie ein Gitter hinter dem Plot oder dem Diagramm einblenden, das die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Das Gitter wird stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Das folgende Beispiel zeigt, wo die Darstellungsoptionen bei einer 3-D-Version des Diagramms platziert sind.

Abbildung 5-55
Position der Diagramm-Darstellungsoptionen bei einem 3-D-Sammlungsdiagramm



Verwendung eines Sammlungsdiagramms

Bei Sammlungen wird die Verteilung der Werte für ein numerisches Feld *relativ zu den Werten eines anderen Felds* dargestellt, also nicht das Auftreten von Werten für ein einziges Feld. Histogramme funktionieren ähnlich wie Sammlungsdiagramme. Histogramme zeigen die Verteilung der Werte in einem numerischen Feld, dessen Werte an der *x*-Achse dargestellt werden.

Sobald Sie ein Diagramm erstellt haben, können Sie die Ergebnisse untersuchen und Abschnitte festlegen, um die Werte entlang der *x*-Achse aufzuspalten bzw. Regionen zu definieren. Außerdem können Sie Elemente innerhalb des Diagramms markieren. Für weitere Informationen siehe Thema Untersuchen von Diagrammen auf S. 360.

Mit Optionen im Menü "Generieren" können Sie Balancierungs- Auswahl- und Ableitungsknoten erstellen. Hierfür werden die Daten im Diagramm bzw. genauer die Daten innerhalb bestimmter Abschnitte, Bereiche oder markierter Elemente verwendet. Dieser

Diagrammtyp wird häufig als Vorbereitung auf Bearbeitungsknoten eingesetzt, um die Daten zu untersuchen und etwaige Unausgewogenheiten mithilfe eines Balancierungsknotens, der aus dem Diagrammknoten heraus erzeugt wird, auszugleichen. Darüber hinaus können Sie einen Flag-Ableitungsknoten erzeugen und so ein Feld hinzufügen, aus dem hervorgeht, in welchen Abschnitt die einzelnen Datensätze fallen, oder auch einen Auswahlknoten, mit dem Sie alle Datensätze in einem bestimmten Set oder Wertebereich auswählen. Diese Funktionen sorgen dafür, dass eine bestimmte Untergruppe an Daten zur näheren Untersuchung im Mittelpunkt verbleibt. Für weitere Informationen siehe Thema Generieren von Knoten aus Diagrammen auf S. 370.

Abbildung 5-56
3-D-Sammlungsdiagramm für die Summe von "Verh_Na/K" über "Alter" für hohe und normale Cholesterolspiegel.

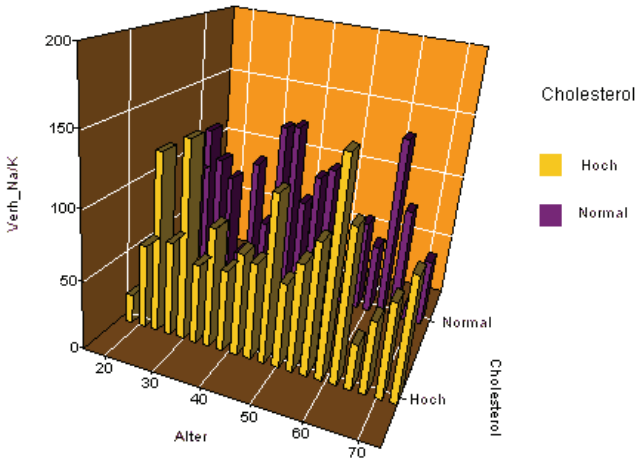
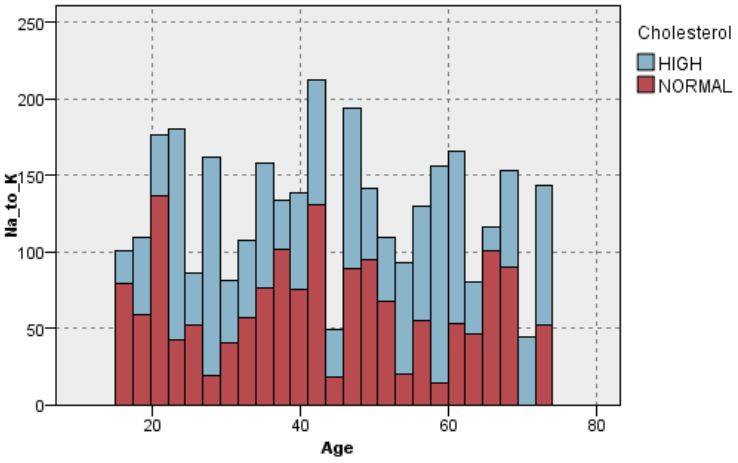


Abbildung 5-57
Sammlungsdiagramm ohne Anzeige der z-Achse, jedoch mit "Cholesterol" als Farbüberlagerung

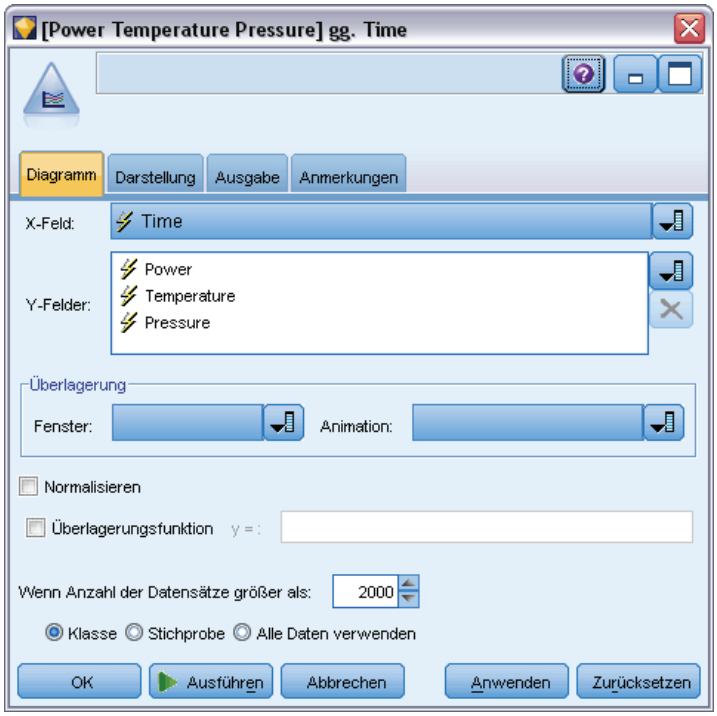


Multidiagrammknoten

Ein Multidiagramm ist eine besondere Art eines Plots, bei dem mehrere *Y*-Felder über einem einzelnen *X*-Feld dargestellt werden. Die *Y*-Felder werden als farbige Linien geplottet, die jeweils einem Plotknoten mit dem Stil Linie und dem X-Modus Sortieren entsprechen. Multidiagramme eignen sich für Zeitsequenzdaten, bei denen die Fluktuation mehrerer Variablen im Lauf der Zeit untersucht werden soll.

Multidiagramm – Registerkarte "Plot"

Abbildung 5-58
Einstellungen auf der Registerkarte "Plot" für Multidiagrammknoten



X-Feld. Wählen Sie ein Feld in der Liste aus, das auf der horizontalen *x*-Achse dargestellt werden soll.

Y-Felder. Wählen Sie mindestens ein Feld in der Liste aus, das über den Bereich der *X*-Feldwerte dargestellt werden soll. Mit der Feldauswahl-Schaltfläche können Sie mehrere Felder auswählen. Mit der Schaltfläche "Löschen" können Sie Felder wieder aus der Liste entfernen.

Überlagerung. Die Kategorien für die Datenwerte können auf unterschiedliche Weise dargestellt werden. Lassen Sie beispielsweise mehrere Plots für die einzelnen Werte in den Daten mithilfe einer Animationsüberlagerung darstellen. Dies ist nützlich für Sets mit mehr als 10 Kategorien. Bei Sets mit mehr als 15 Kategorien kann die Leistung beeinträchtigt werden. Für weitere Informationen siehe Thema Formatierungen, Überlagerungen, Fenster und Animation auf S. 246.

Normalisieren. Hiermit lassen Sie alle *Y*-Werte auf den Bereich 0–1 zur Darstellung im Diagramm skalieren. Durch Normalisieren können Sie die Beziehung zwischen Linien untersuchen, die im Diagramm ansonsten aufgrund von Unterschieden im Wertebereich für die einzelnen Reihen verdeckt sind. Das Normalisieren wird bei der Darstellung mehrerer Linien im selben Diagramm und für den Vergleich von Plots in nebeneinander angeordneten Teilfenstern empfohlen. (Eine Normalisierung ist nicht erforderlich, wenn alle Datenwerte in einen ähnlichen Bereich fallen.)

Abbildung 5-59
Standard-Multidiagramm mit der Kraftwerksfluktuation im Lauf der Zeit (Hinweis: Ohne Normalisierung ist der Plot für den Druck nicht sichtbar)

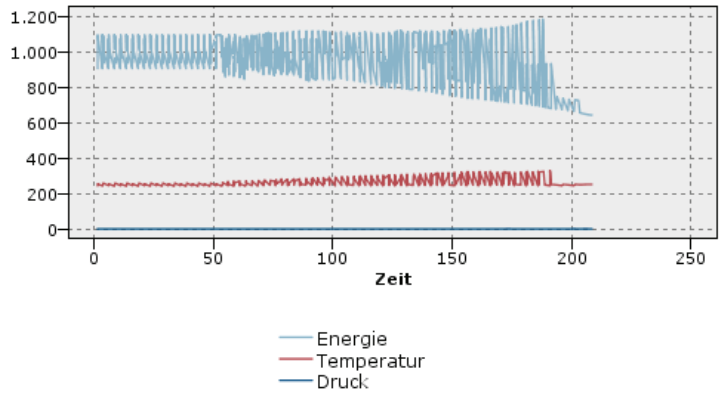
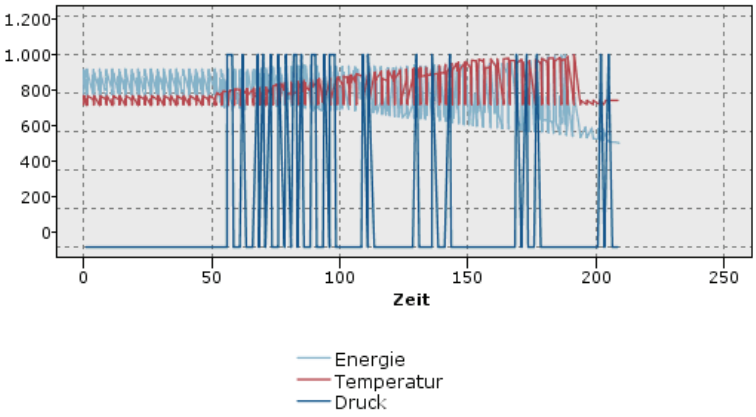


Abbildung 5-60
Normalisiertes Multidiagramm mit einem Plot für den Druck



Überlagerungsfunktion. Hiermit geben Sie eine bekannte Funktion an, die mit tatsächlichen Werten verglichen werden soll. Um beispielsweise die Istwerte mit den vorhergesagten Werten zu vergleichen, plotten Sie die Funktion y = x als Überlagerung. Geben Sie eine Funktion für y = im Textfeld an. Die Standardfunktion lautet y = x; Sie können jedoch auch andere Funktionen festlegen, z. B. quadratische Funktionen oder beliebige Ausdrücke im Hinblick auf x.

Hinweis: Überlagerungsfunktionen sind für Fenster- und Animationsdiagramme nicht verfügbar.

Wenn Anzahl der Datensätze größer als. Geben Sie eine Methode für das Plotten umfangreicher Daten-Sets an. Sie können wahlweise eine maximal zulässige Größe für das Daten-Set angeben oder den Standardwert von 2.000 Punkten verwenden. Bei umfangreichen Daten-Sets steigt die Leistung, wenn Sie die Option Klasse oder Stichprobe aktivieren. Alternativ können Sie mit Alle Daten verwenden alle Datenpunkte gleichzeitig plotten lassen; dies kann sich jedoch beträchtlich auf die Leistung der Software auswirken.

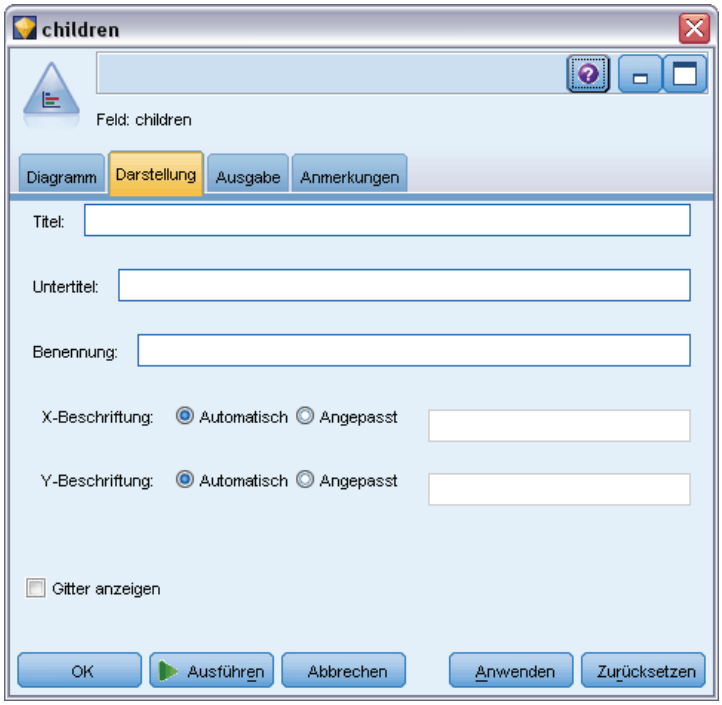
Hinweis: Beim X-Modus Überlagern oder Wie gelesen sind diese Optionen deaktiviert und es werden nur die ersten *n* Datensätze verwendet.

- **Klasse.** Hiermit aktivieren Sie die Klassierung, wenn das Daten-Set mehr Datensätze enthält als die angegebene Anzahl. Beim Klassieren wird das Diagramm vor dem eigentlichen Plotten in feinmaschige Gitter aufgeteilt und es wird die Anzahl der Verbindungen gezählt, die in die einzelnen Gitterzellen fallen würden. Im endgültigen Diagramm wird je eine Verbindung pro Zelle im Klassierschwerpunkt (Durchschnitt aller Verbindungspunkte in der Klasse) verwendet.
- **Beispiel.** Es wird eine zufällige Stichprobe mit der angegebenen Anzahl von Datensätzen aus den Daten gebildet.

Multiplot – Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Abbildung 5-61
Einstellungen auf der Registerkarte "Darstellung" für die meisten Diagrammknoten



Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Benennung. Dient zur Eingabe des Texts, der zur Benennung des Diagramms verwendet werden soll.

X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *x*-Achsen-Beschriftung (horizontal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

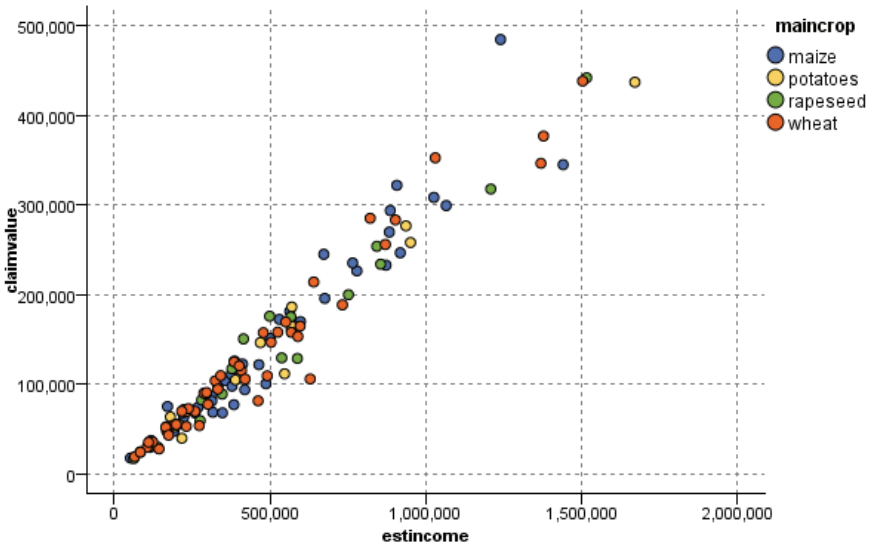
Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *y*-Achsen-Beschriftung (vertikal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

Gitter anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie ein Gitter hinter dem Plot oder dem Diagramm einblenden, das die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Das Gitter wird stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Verwendung eines Multidiagramms

Plots und Multidiagramme sind im Grunde genommen Plots von *X* in Abhängigkeit von *Y*. Wenn Sie beispielsweise potenzielle Betrugsfälle in Bewerbungen um landwirtschaftliche Subventionen untersuchen (wie in *fraud.str* im Ordner *Demos* der IBM® SPSS® Modeler-Installation dargestellt), soll beispielsweise das auf der Bewerbung angegebene Einkommen in Abhängigkeit von dem Einkommen geplottet werden, das mithilfe eines neuronalen Netzes geschätzt wurde. Aus einer Überlagerung, z. B. dem Feldfruchttyp, geht hervor, ob eine Beziehung zwischen den Forderungen (Wert oder Anzahl) und der Art der Feldfrucht besteht.

Abbildung 5-62
Plot der Beziehung zwischen geschätztem Einkommen und Forderungswert mit Hauptfeldfruchttyp als Überlagerung



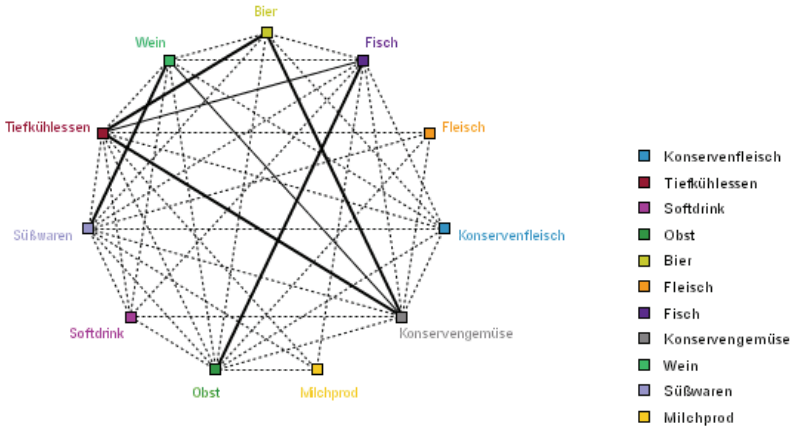
Plots, Multidiagramme und Evaluationsdiagramme sind zweidimensionale Darstellungen von *Y* gegen *X*. Die Arbeit mit diesen Diagrammen ist daher denkbar unkompliziert: Sie können ganz einfach Bereiche definieren, Elemente markieren und sogar Abschnitte einzeichnen. Außerdem

können Sie Knoten für die durch diese Bereiche, Abschnitte bzw. Elemente dargestellten Daten generieren. Für weitere Informationen siehe Thema Untersuchen von Diagrammen auf S. 360.

Netzdiagrammknoten

Netzdiagrammknoten zeigen die Stärke der Beziehung zwischen den Werten aus mindestens zwei symbolischen Feldern. Die Verbindungen werden mithilfe verschiedener Linientypen im Diagramm dargestellt, aus denen die Stärke der jeweiligen Verbindung hervorgeht. Mit Netzdiagrammknoten können Sie beispielsweise die Beziehung zwischen dem Kauf verschiedener Artikel auf einer e-Commerce-Website oder in einem traditionellen Einzelhandelsgeschäft untersuchen.

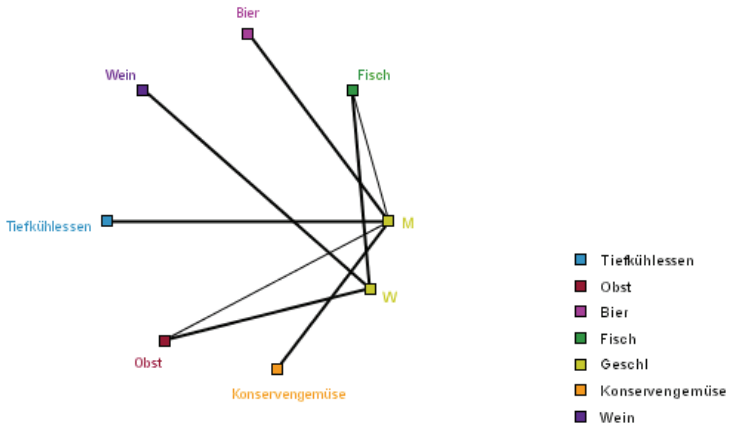
Abbildung 5-63
Netzdiagramm mit Beziehungen zwischen dem Kauf von Lebensmitteln



Gerichtete Netzdiagramme

Gerichtete Netzdiagrammknoten zeigen wie die Netzdiagrammknoten die Stärke der Beziehungen zwischen symbolischen Feldern. In gerichteten Netzdiagrammen sind jedoch nur die Verbindungen von mindestens einem Ausgangsfeld zu einem einzelnen Zielfeld ersichtlich. Die Verbindungen sind unidirektional, verlaufen also nur als “Einbahnstraßen”.

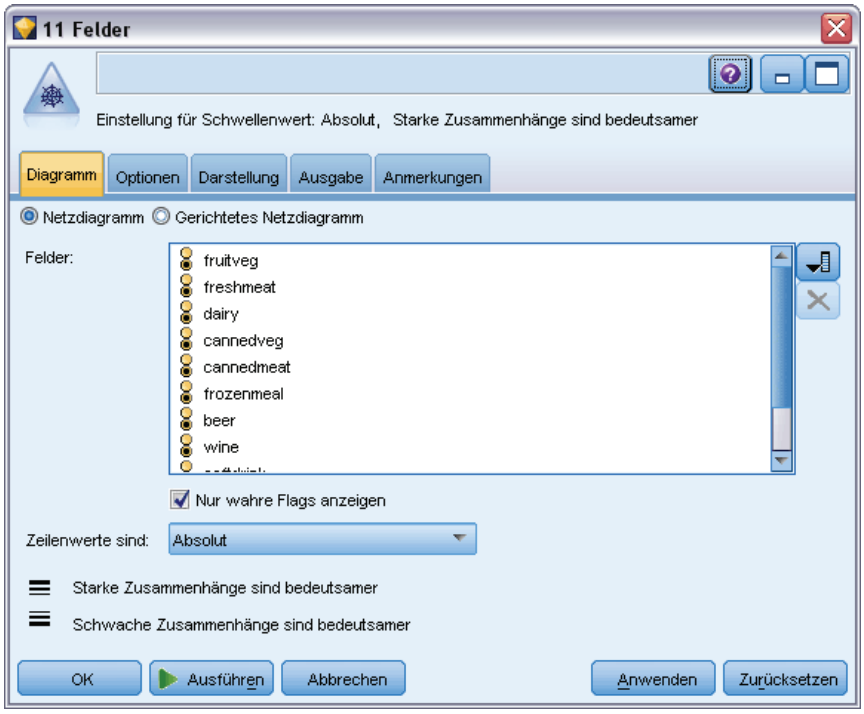
Abbildung 5-64
Gerichtetes Netzdiagramm mit der Beziehung zwischen dem Kauf von Lebensmitteln und dem Geschlecht



Wie bei Netzdiagrammknoten werden die Verbindungen mithilfe verschiedener Linientypen im Diagramm dargestellt, aus denen die Stärke der jeweiligen Verbindung hervorgeht. Mit einem gerichteten Netzdiagrammknoten können Sie beispielsweise die Beziehung zwischen dem Geschlecht des Käufers und der Neigung zum Kauf bestimmter Artikel untersuchen.

Netzdiagramm – Registerkarte "Plot"

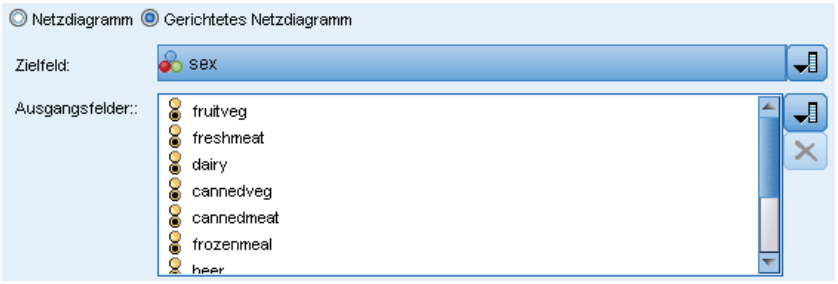
Abbildung 5-65
Einstellungen auf der Registerkarte "Plot" für Netzdiagrammknoten



Netzdiagramm. Hiermit erstellen Sie ein Netzdiagramm, das die Stärke der Beziehungen zwischen allen angegebenen Feldern verdeutlicht.

Gerichtetes Netzdiagramm. Ein gerichtetes Netzdiagramm wird erstellt, aus dem die Stärke der Beziehungen zwischen mehreren Feldern und den Werten eines einzigen Felds (z. B. Geschlecht oder Religion) hervorgeht. Wenn diese Option ausgewählt ist, wird ein "Zielfeld" aktiviert und das nachfolgende Steuerelement für die Felder wird zur näheren Verdeutlichung in Ausgangsfelder umbenannt.

Abbildung 5-66
Optionen bei gerichteten Netzdiagrammen



Zielfeld (nur bei gerichteten Netzdiagrammen). Wählen Sie ein Flag- oder nominales Feld für ein zielgerichtetes Netzdiagramm aus. Die Liste enthält nur solche Felder, die nicht explizit als numerisch definiert sind.

Felder/Ausgangsfelder. Wählen Sie die gewünschten Felder für die Erstellung des gerichteten Netzdiagramms aus. Die Liste enthält nur solche Felder, die nicht explizit als numerisch definiert sind. Mit der Feldauswahl-Schaltfläche können Sie mehrere Felder auswählen. Alternativ können Sie die Felder nach Typ auswählen.

Hinweis: Bei gerichteten Netzdiagrammen dient dieses Steuerelement zur Auswahl der Ausgangsfelder.

Nur wahre Flags anzeigen. Es werden nur wahre Flags für ein Flag-Feld angezeigt. Diese Option vereinfacht die Netzdiagramm-Anzeige und wird häufig bei Daten verwendet, bei denen das Auftreten positiver Werte von besonderer Bedeutung ist.

Zeilenwerte sind. Wählen Sie einen Schwellenwerttyp aus der Dropdown-Liste aus.

- Mit Absolut werden die Schwellenwerte auf der Grundlage der Anzahl an Datensätzen festgelegt, in denen die einzelnen Wertepaare vorkommen.
- Mit Prozent insgesamt rufen Sie die absolute Anzahl der Fälle ab, die im Zusammenhang als Anteil am Gesamtauftreten der einzelnen Wertepaare im Netzdiagramm dargestellt werden.
- Aus den Feldern Prozentsätze vom kleineren Feld/Wert und Prozentsätze vom größeren Feld/Wert geht hervor, welches Feld bzw. welcher Wert für die Evaluation der Prozentsätze herangezogen werden soll. Beispiel: 100 Datensätze besitzen den Wert *MedY* für das Feld *Medikament*, nur 10 Datensätze dagegen den Wert *NIEDRIG* im Feld *BP*. Wenn sieben Datensätze sowohl den Wert *MedY* als auch *NIEDRIG* aufweisen, beträgt der Prozentsatz entsprechend 70 % oder 7 %, abhängig davon, welches Feld referenziert wird, also kleiner (*BP*) oder größer (*Medikament*).

Hinweis: Bei gerichteten Netzdiagrammen sind die dritte und vierte oben genannte Option nicht verfügbar. Stattdessen können Sie die Optionen Prozentsätze vom Feld/Wert “Bis” und Prozentsätze vom Feld/Wert “Von” auswählen.

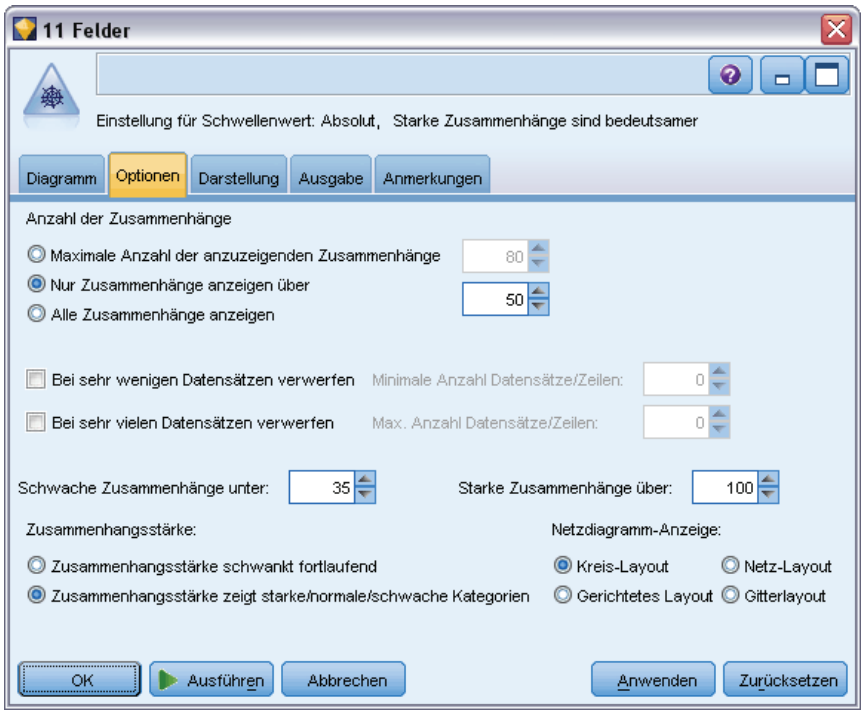
Starke Zusammenhänge sind bedeutsamer. Diese Option ist standardmäßig aktiviert und ist die Standarddarstellung der Zusammenhänge zwischen Feldern.

Schwache Zusammenhänge sind bedeutsamer. Hiermit kehren Sie die Bedeutung der als fett gedruckte Linien dargestellten Zusammenhänge um. Diese Option wird häufig im Rahmen der Betrugserkennung oder bei der Untersuchung von Ausreißern herangezogen.

Netzdiagramm – Registerkarte “Optionen”

Bei Netzdiagrammknoten enthält die Registerkarte “Optionen” eine Reihe weiterer Optionen, mit denen Sie das Ausgabediagramm anpassen können.

Abbildung 5-67
Einstellungen auf der Registerkarte “Optionen” für Netzdiagrammknoten



Anzahl der Zusammenhänge. Mit den nachstehenden Optionen wird die Anzahl der im Ausgabediagramm dargestellten Zusammenhänge festgelegt. Ein Teil dieser Optionen, z. B. Schwache Zusammenhänge unter oder Starke Zusammenhänge über, stehen auch im Ausgabediagrammfenster zur Verfügung. Des Weiteren können Sie die Anzahl der angezeigten Zusammenhänge mit einem Schieberegler im fertigen Diagramm einstellen.

- **Maximale Anzahl der anzuzeigenden Zusammenhänge.** Geben Sie die maximale Anzahl der Zusammenhänge ein, die im Ausgabediagramm dargestellt werden sollen. Mithilfe der Pfeile können Sie den Wert einstellen.

- **Nur Zusammenhänge anzeigen über.** Geben Sie den Mindestwert an, ab dem eine Verbindung im Netzdiagramm dargestellt werden soll. Mithilfe der Pfeile können Sie den Wert einstellen.
- **Alle Zusammenhänge anzeigen.** Alle Zusammenhänge werden angezeigt, unabhängig von den Mindest- und Höchstwerten. Bei dieser Option steigt ggf. die Verarbeitungszeit an, wenn eine große Anzahl an Feldern vorliegt.

Bei sehr wenigen Datensätzen verwerfen. Zusammenhänge, die durch zu wenige Datensätze gestützt sind, werden ignoriert. Um den Schwellenwert für diese Option festzulegen, geben Sie den gewünschten Wert in das Feld Minimale Anzahl Datensätze/Zeilen ein.

Bei sehr vielen Datensätzen verwerfen. Stark gestützte Verbindungen werden ignoriert. Geben Sie den gewünschten Wert in das Feld Max. Anzahl Datensätze/Zeilen ein.

Schwache Zusammenhänge unter. Bestimmen Sie einen Schwellenwert für schwache Verbindungen (gepunktete Linien) und normale Verbindungen (normale Linien). Alle Verbindungen unterhalb dieses Werts gelten als schwach.

Starke Zusammenhänge über. Bestimmen Sie einen Schwellenwert für starke Verbindungen (dicke Linien) und normale Verbindungen (normale Linien). Alle Verbindungen oberhalb dieses Werts gelten als stark.

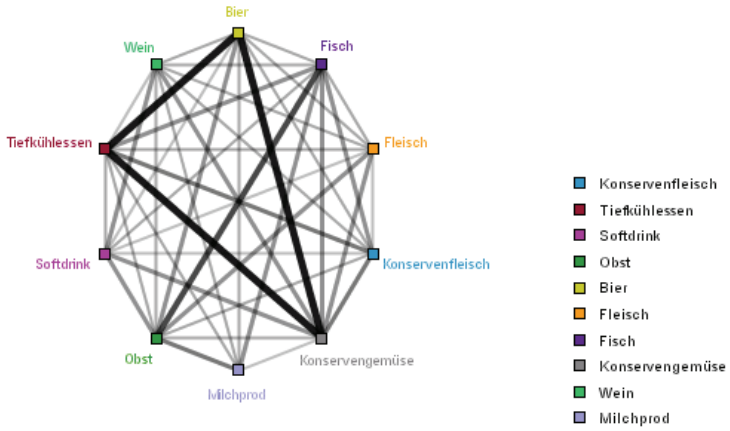
Zusammenhangsstärke. Legen Sie Optionen zur Steuerung der Stärke von Zusammenhängen fest:

- **Zusammenhangsstärke schwankt fortlaufend.** Es wird ein Bereich von Zusammenhangsstärken dargestellt, der die Schwankungen bei der Verbindungsstärke auf der Grundlage der tatsächlichen Datenwerte wiedergibt.
- **Zusammenhangsstärke zeigt starke/normale/schwache Kategorien.** Es werden drei Verbindungsstärken dargestellt (stark, normal und schwach). Die Trennwerte für diese Kategorien können wahlweise oben bestimmt werden oder auch im fertigen Diagramm.

Netzdiagramm-Anzeige. Wählen Sie einen Typ für die Netzdiagramm-Anzeige aus:

- **Kreis-Layout.** Die standardmäßige Netzdiagramm-Anzeige wird verwendet.
- **Netz-Layout.** Die stärksten Zusammenhänge werden mithilfe eines Algorithmus gruppiert. Auf diese Weise werden starke Zusammenhänge durch räumliche Differenzierung und durch gewichtete Linien hervorgehoben.
- **Gerichtetes Layout.** Wählen Sie diese Option, um eine gerichtete Webanzeige zu erstellen, die die Auswahl Zielfeld als Schwerpunkt für die Richtung verwendet.
- **Gitterlayout.** Wählen Sie diese Option aus, um eine Webanzeige zu erstellen, die in einem Gittermuster mit regelmäßigen Abständen ausgelegt ist.

Abbildung 5-68
Netz-Diagramm mit starken Verbindungen von "Tiefkühlware" und "Gemüse in Dosen" zu anderen Lebensmitteln



Netzdiagramm – Registerkarte "Darstellung"

Abbildung 5-69
Einstellungen auf der Registerkarte "Darstellung" für einen Netzdiagrammknoten

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

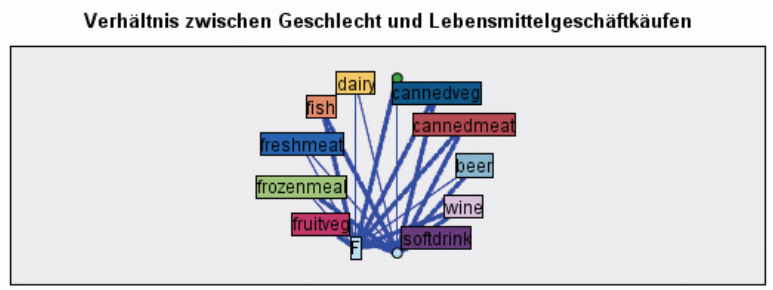
Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Benennung. Dient zur Eingabe des Texts, der zur Benennung des Diagramms verwendet werden soll.

Legende anzeigen. Dient zur Angabe, ob die Legende angezeigt werden soll oder nicht. Bei Plots mit zahlreichen Feldern wird die Darstellung ggf. verbessert, wenn Sie die Legende ausblenden.

Beschriftungen als Knoten anzeigen. Gibt an, dass die Beschriftungen nicht seitlich aufgeführt werden sollen, sondern direkt in jedem Knoten. Bei Plots mit einer geringen Anzahl an Feldern kann so die Übersichtlichkeit des Diagramms verbessert werden.

Abbildung 5-70
Netzdiagramm mit Beschriftungen als Knoten



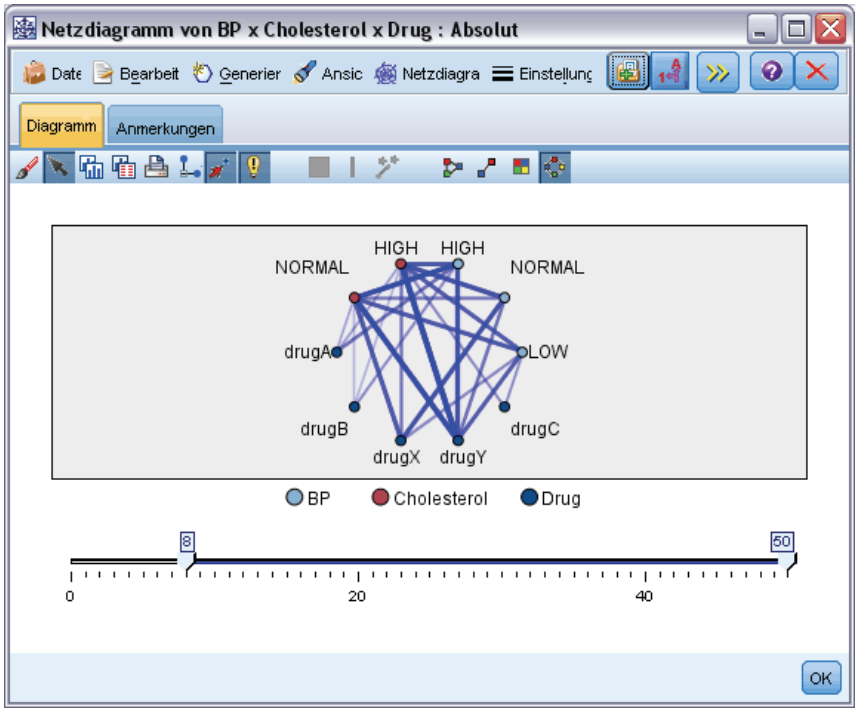
Verwendung eines Netzdiagramms

Netzdiagrammknoten zeigen die Stärke der Beziehung zwischen den Werten aus mindestens zwei symbolischen Feldern. Die Verbindungen werden in Form von verschiedenen Linientypen, mit denen die größer werdende Stärke der Verbindungen verdeutlicht wird, in einem Diagramm dargestellt. Mit einem Netzdiagrammknoten können Sie beispielsweise die Beziehung zwischen Cholesterolspiegel, Blutdruck und dem eingenommenen Medikament bei der Behandlung des Patienten untersuchen.

- Starke Verbindungen sind mit einer dicken Linie gekennzeichnet. Dies bedeutet, dass die beiden Werte eng zusammenhängen und näher untersucht werden sollten.
- Mittelstarke Verbindungen sind als normal dicke Linien dargestellt.
- Schwache Verbindungen sind mit einer gepunkteten Linie gekennzeichnet.
- Befindet sich keine Linie zwischen zwei Werten, bedeutet dies entweder, dass die betreffenden Werte niemals gemeinsam in einem einzigen Datensatz auftreten oder dass diese Kombination in einer Anzahl an Datensätzen vorliegt, die unter dem im Dialogfeld für den Netzdiagrammknoten festgelegten Schwellenwert liegt.

Sobald Sie einen Netzdiagrammknoten erstellt haben, stehen verschiedene Optionen zur Auswahl, mit denen Sie die Darstellung des Diagramms anpassen und Knoten für die weitere Analyse erzeugen können.

Abbildung 5-71
Netzdiagramm mit einer Reihe starker Verbindungen, z. B. normaler Blutdruck und MedX oder hoher Cholesterolspiegel und MedY



Bei Netzdiagrammknoten und gerichteten Netzdiagrammknoten stehen die folgenden Möglichkeiten zur Auswahl:

- Ändern Sie das Layout der Netzdiagramm-Anzeige.
- Blenden Sie verschiedene Punkte aus, um so die Darstellung zu vereinfachen.
- Ändern Sie die Schwellenwerte für die Linienstile.
- Heben Sie Linien zwischen Werten hervor und kennzeichnen Sie so eine “ausgewählte” Beziehung.
- Erzeugen Sie einen Auswahlknoten für einen oder mehrere “ausgewählte” Datensätze oder auch einen Flag-Ableitungsknoten, der mit mindestens einer Beziehung im Netzdiagramm assoziiert ist.

So passen Sie die Punkte an:

- Punkte **verschieben**: Klicken Sie mit der Maus auf einen Punkt und ziehen Sie diesen an die gewünschte Position. Das Netzdiagramm wird neu gezeichnet, um so die neue Position wiederzugeben.
- Punkte **ausblenden**: Klicken Sie mit der rechten Maustaste auf einen Punkt im Netzdiagramm und wählen Sie im Kontextmenü die Option Ausblenden oder Ausblenden und neu zeichnen. Bei der Option Ausblenden lassen Sie lediglich den ausgewählten Punkt und die zugehörigen Linien ausblenden. Bei der Option Ausblenden und neu zeichnen wird das Netzdiagramm

neu gezeichnet, sodass die vorgenommenen Änderungen ersichtlich werden. Alle manuell vorgenommenen Verschiebungen werden rückgängig gemacht.

- Alle ausgeblendeten Punkte **anzeigen**: Wählen Sie im Diagrammfenster im Menü "Netzdiagramm" den Befehl Alle anzeigen oder Alle anzeigen und neu zeichnen. Mit der Option Alle anzeigen und neu zeichnen lassen Sie das Netzdiagramm neu zeichnen, sodass auch alle bislang ausgeblendeten Punkte und deren Verbindungen wieder sichtbar werden.

So können Sie Linien auswählen oder "hervorheben":

Ausgewählte Linien werden in roter Farbe hervorgehoben.

▶ Klicken Sie zum Auswählen einer einzelnen Linie bei gedrückter linker Maustaste auf die Linie.

▶ Um mehrere Linien auszuwählen, führen Sie eine der folgenden Aktionen aus:

- Ziehen Sie mithilfe des Cursors einen Kreis um die Punkte auf, deren Linien Sie auswählen möchten.
- Halten Sie die Strg-Taste gedrückt und klicken Sie mit der linken Maustaste auf die einzelnen Linien, die Sie auswählen möchten.

Sie können die Auswahl aller Linien aufheben, indem Sie in den Diagrammhintergrund klicken oder Auswahl aufheben aus dem Web-Menü im Diagrammfenster wählen.

So lassen Sie das Netzdiagramm mithilfe eines anderen Layouts anzeigen:

▶ Wählen Sie im Menü "Web" Kreis-Layout, Netz-Layout, Gerichtetes Layout oder Gitterlayout aus, um das Layout des Diagramms zu ändern.

So schalten Sie den Links-Schieberegler ein bzw. aus.

▶ Wählen Sie im Menü "Ansicht" die Option Links-Schieberegler.

So können Sie Datensätze für eine einzelne Beziehung auswählen oder mit einem Flag versehen:

▶ Klicken Sie mit der rechten Maustaste auf die Linie für die relevante Beziehung.

▶ Wählen Sie im Kontextmenü die Option Auswahlknoten für Zusammenhang generieren oder Ableitungsknoten für Zusammenhang generieren.

In den Stream-Zeichenbereich wird automatisch ein Auswahlknoten oder Ableitungsknoten mit den richtigen Optionen und Bedingungen aufgenommen.

- Mit dem Auswahlknoten werden alle Datensätze in der betreffenden Beziehung ausgewählt.
- Der Ableitungsknoten erzeugt ein Flag, aus dem hervorgeht, ob die ausgewählte Beziehung für Datensätze im gesamten Daten-Set gilt. Der Name des Flag-Felds besteht aus den beiden Werten in der Beziehung, getrennt durch einen Unterstrich, z. B. *NIEDRIG_MedC* oder *MedC_NIEDRIG*.

So können Sie Datensätze für eine Gruppe von Beziehungen auswählen oder mit einem Flag versehen:

▶ Wählen Sie die Linie(n) für die relevanten Beziehungen in der Netzdiagramm-Anzeige aus.

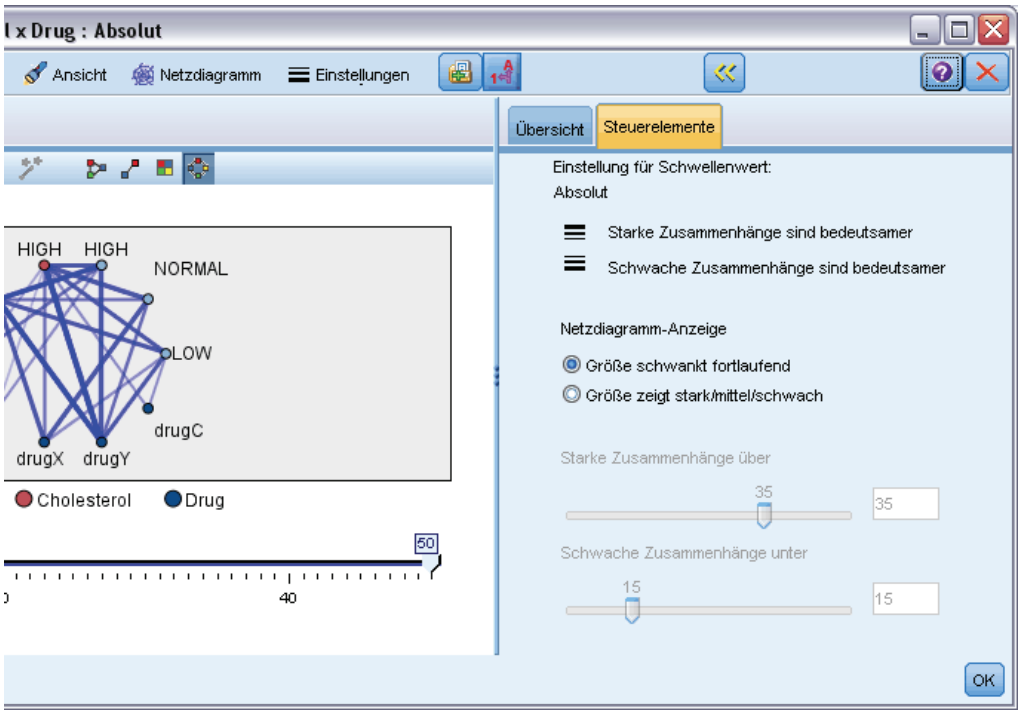
- Wählen Sie im Diagrammfenster im Menü “Generieren” den Befehl Auswahlknoten (“UND”), Auswahlknoten (“ODER”), Ableitungsknoten (“UND”) oder Ableitungsknoten (“ODER”).
  - Bei den “ODER”-Knoten werden die Bedingungen getrennt voneinander betrachtet. Der Knoten gilt also für alle Datensätze, bei denen mindestens eine der ausgewählten Beziehungen vorliegt.
  - Bei den “UND”-Knoten werden die Bedingungen gemeinsam betrachtet. Der Knoten gilt also für alle Datensätze, bei denen alle ausgewählten Beziehungen vorliegen. Falls ausgewählte Beziehungen sich gegenseitig ausschließen, tritt ein Fehler auf.

Sobald die Auswahl abgeschlossen ist, wird automatisch ein Auswahlknoten oder Ableitungsknoten mit den richtigen Optionen und Bedingungen in den Stream-Zeichenbereich aufgenommen.

Anpassen der Netzdiagramm-Schwellenwerte

Sobald Sie ein Netzdiagramm erstellt haben, können Sie die Schwellenwerte für die Linienstile mit dem Schieberegler einstellen und so die minimale noch sichtbare Linie ändern. Des Weiteren können zusätzliche Optionen für die Schwellenwerte abgerufen werden. Klicken Sie hierzu auf den gelben Doppelpfeil in der Symbolleiste. Das Netzdiagrammfenster wird erweitert. Klicken Sie anschließend auf die Registerkarte Steuerelemente und wählen Sie die gewünschten zusätzlichen Optionen.

Abbildung 5-72
Erweitertes Fenster mit Darstellungs- und Schwellenwert-Optionen



Einstellung für Schwellenwert. Dies ist der Typ des Schwellenwerts, den Sie beim Erstellen im Dialogfeld des Netzdiagrammknotens ausgewählt haben.

Starke Zusammenhänge sind bedeutsamer. Diese Option ist standardmäßig aktiviert und ist die Standarddarstellung der Zusammenhänge zwischen Feldern.

Schwache Zusammenhänge sind bedeutsamer. Hiermit kehren Sie die Bedeutung der als fett gedruckte Linien dargestellten Zusammenhänge um. Diese Option wird häufig im Rahmen der Betrugserkennung oder bei der Untersuchung von Ausreißern herangezogen.

Netzdiagramm-Anzeige. Legen Sie Optionen zur Steuerung der Stärke von Zusammenhängen im Ausgabediagramm fest:

- **Zusammenhangsstärke schwankt fortlaufend.** Es wird ein Bereich von Zusammenhangsstärken dargestellt, der die Schwankungen bei der Verbindungsstärke auf der Grundlage der tatsächlichen Datenwerte wiedergibt.
- **Größe zeigt stark/mittel/schwach.** Es werden drei Verbindungsstärken dargestellt (stark, normal und schwach). Die Trennwerte für diese Kategorien können wahlweise oben bestimmt werden oder auch im fertigen Diagramm.

Starke Zusammenhänge über. Bestimmen Sie einen Schwellenwert für starke Verbindungen (dicke Linien) und normale Verbindungen (normale Linien). Alle Verbindungen oberhalb dieses Werts gelten als stark. Stellen Sie den Wert mit dem Schieberegler ein oder geben Sie einen Wert in das Feld ein.

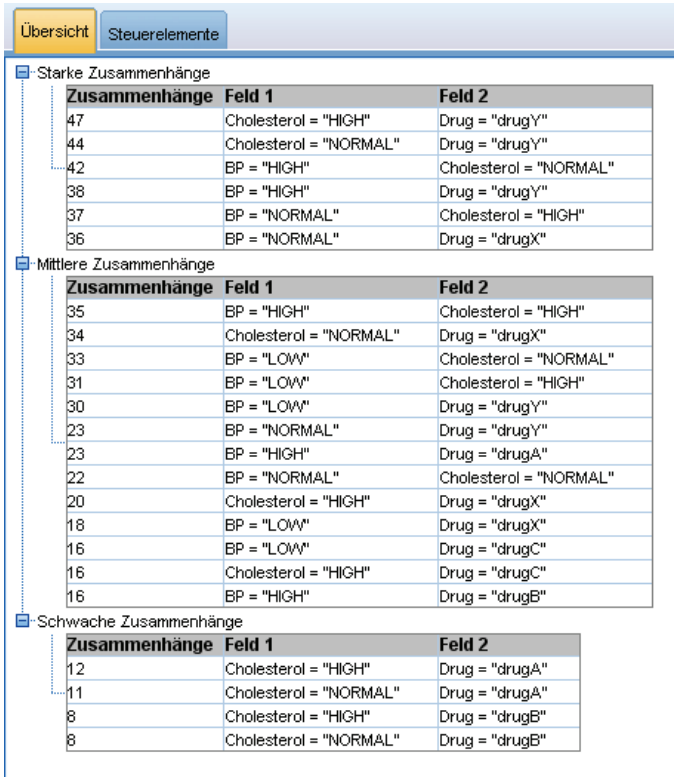
Schwache Zusammenhänge unter. Bestimmen Sie einen Schwellenwert für schwache Verbindungen (gepunktete Linien) und normale Verbindungen (normale Linien). Alle Verbindungen unterhalb dieses Werts gelten als schwach. Stellen Sie den Wert mit dem Schieberegler ein oder geben Sie einen Wert in das Feld ein.

Wenn Sie die Schwellenwerte für ein Netzdiagramm angepasst haben, können Sie die Netzdiagramm-Anzeige mit den neuen Schwellenwerten aktualisieren (neu zeichnen). Verwenden Sie hierzu das Menü in der Symbolleiste des Netzdiagramms. Sobald Sie die richtigen Einstellungen gefunden haben, die zu den aussagekräftigsten Mustern führen, können Sie die ursprünglichen Einstellungen im Netzdiagrammknoten (auch als "übergeordneter Netzdiagrammknoten" bezeichnet) aktualisieren. Wählen Sie hierzu im Diagrammfenster im Menü "Netzdiagramm" den Befehl Übergeordneten Knoten aktualisieren.

Erstellen einer Netzdiagramm-Übersicht

Sie können eine Netzdiagramm-Übersicht anlegen, in der die starken, mittleren und schwachen Linien aufgeführt werden. Klicken Sie hierzu auf den gelben Doppelpfeil in der Symbolleiste. Das Netzdiagrammfenster wird erweitert. Klicken Sie anschließend auf die Registerkarte Übersicht. Hier werden Tabellen für die einzelnen Arten der Zusammenhänge aufgeführt. Mit den Umschalttasten können Sie die Tabellen erweitern und reduzieren.

Abbildung 5-73
Netzdiagramm-Übersicht mit einer Liste der Verbindungen zwischen Blutdruck, Cholesterol und Medikamententyp



Übersicht | Steuerelemente

Starke Zusammenhänge

Zusammenhänge	Feld 1	Feld 2
47	Cholesterol = "HIGH"	Drug = "drugY"
44	Cholesterol = "NORMAL"	Drug = "drugY"
42	BP = "HIGH"	Cholesterol = "NORMAL"
38	BP = "HIGH"	Drug = "drugY"
37	BP = "NORMAL"	Cholesterol = "HIGH"
36	BP = "NORMAL"	Drug = "drugX"

Mittlere Zusammenhänge

Zusammenhänge	Feld 1	Feld 2
35	BP = "HIGH"	Cholesterol = "HIGH"
34	Cholesterol = "NORMAL"	Drug = "drugX"
33	BP = "LOW"	Cholesterol = "NORMAL"
31	BP = "LOW"	Cholesterol = "HIGH"
30	BP = "LOW"	Drug = "drugY"
23	BP = "NORMAL"	Drug = "drugY"
23	BP = "HIGH"	Drug = "drugA"
22	BP = "NORMAL"	Cholesterol = "NORMAL"
20	Cholesterol = "HIGH"	Drug = "drugX"
18	BP = "LOW"	Drug = "drugX"
16	BP = "LOW"	Drug = "drugC"
16	Cholesterol = "HIGH"	Drug = "drugC"
16	BP = "HIGH"	Drug = "drugB"

Schwache Zusammenhänge

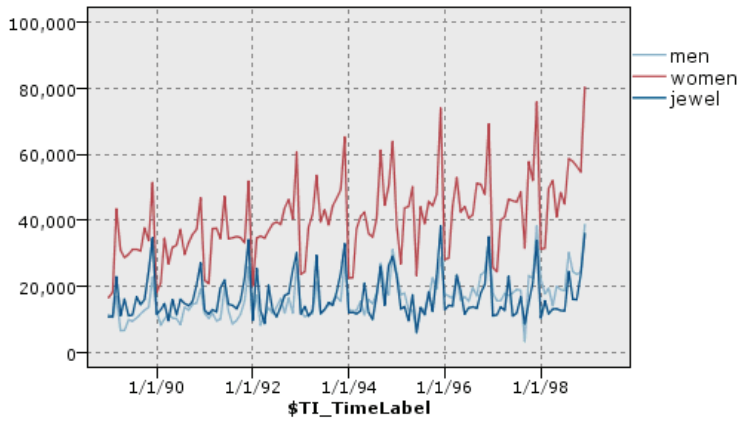
Zusammenhänge	Feld 1	Feld 2
12	Cholesterol = "HIGH"	Drug = "drugA"
11	Cholesterol = "NORMAL"	Drug = "drugA"
8	Cholesterol = "HIGH"	Drug = "drugB"
8	Cholesterol = "NORMAL"	Drug = "drugB"

Um die Übersicht auszudrucken, wählen Sie Folgendes aus dem Menü im Netzdiagrammfenster:
Datei > Übersicht drucken

Zeitdiagrammknoten

Mit Zeitdiagrammknoten können Sie die Darstellung einer oder mehrerer Zeitreihen über einen bestimmten Zeitraum anzeigen. Die geplotteten Reihen müssen numerische Werte enthalten. Außerdem wird vorausgesetzt, dass sie in einem Zeitbereich mit einheitlichen Abschnitten auftreten. Normalerweise wird vor einem Zeitdiagrammknoten ein Zeitintervallknoten verwendet, um ein *TimeLabel*-Feld zu erstellen, das standardmäßig zur Beschriftung der *x*-Achse im Diagramm verwendet wird. Für weitere Informationen siehe Thema Zeitintervallknoten in Kapitel 4 auf S. 219.

Abbildung 5-74
Darstellung der Verkäufe bei Herren- und Damenbekleidung und Schmuck im Laufe der Zeit



Erstellen von Interventionen und Ereignissen

Sie können Ereignis- und Interventionsfelder aus dem Zeitdiagramm erstellen, indem Sie einen Ableitungsknoten (Flag oder nominal) aus den Kontextmenüs generieren. Sie können beispielsweise ein Ereignisfeld für den Fall eines Bahnstreiks erstellen, der den Ableitungsstatus “Wahr” aufweist, wenn das Ereignis eingetreten ist, und ansonsten den Ableitungsstatus “Falsch”. Bei einem Interventionsfeld, beispielsweise für eine Preiserhöhung, könnten Sie eine Ableitungsanzahl verwenden, um das Datum der Erhöhung anzugeben; dabei wird der Wert “0” für den alten und “1” für den neuen Preis verwendet. Für weitere Informationen siehe Thema Ableitungsknoten in Kapitel 4 auf S. 167.

Zeit – Registerkarte "Plot"

Abbildung 5-75
Einstellungen auf der Registerkarte "Plot" für Zeitdiagrammknoten

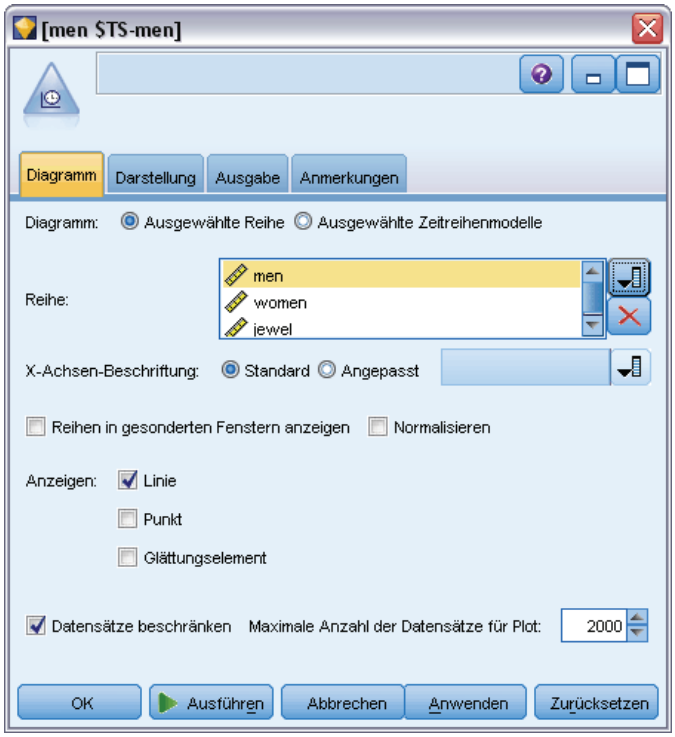


Diagramm. Bietet eine Auswahl für das Plotten der Zeitreihendaten.

- **Ausgewählte Reihe.** Plottet Werte für ausgewählte Zeitreihen. Wenn Sie diese Option beim Plotten von Konfidenzintervallen auswählen, müssen Sie das Kontrollkästchen Normalisieren deaktivieren.
- **Ausgewählte Zeitreihenmodelle**. In Verbindung mit einem Zeitreihenmodell plottet diese Option alle verwandten Felder (tatsächliche und vorhergesagte Werte sowie die Konfidenzintervalle) für mindestens eine ausgewählte Zeitreihe. Mit dieser Option werden einige anderen Optionen im Dialogfeld deaktiviert. Diese Option eignet sich besonders für das Plotten von Konfidenzintervallen.

Datenreihen. Wählen Sie mindestens ein Feld mit Zeitreihendaten für den Plot aus. Die Daten müssen numerisch sein.

X-Achsen-Beschriftung. Wählen Sie entweder die Standardbeschriftung oder ein einzelnes Feld aus, das als Beschriftung für die *x*-Achse in Plots dienen soll. Bei Auswahl von "Standard" verwendet das System das aus einem aufwärts gelegenen Zeitintervallknoten erstellte TimeLabel-Feld bzw. aufeinanderfolgende ganze Zahlen, wenn kein Zeitintervallknoten vorhanden ist. Für weitere Informationen siehe Thema Zeitintervallknoten in Kapitel 4 auf S. 219.

Reihe in gesonderten Fenstern anzeigen. Gibt an, ob jede Reihe in einem gesonderten Fenster angezeigt werden soll. Wenn Sie nicht verschiedene Fenster verwenden möchten, werden alternativ alle Zeitreihen im selben Diagramm dargestellt und es stehen keine Glättungselemente

zur Verfügung. Wenn alle Zeitreihen im selben Diagramm dargestellt werden, wird jede Reihe in einer anderen Farbe angezeigt.

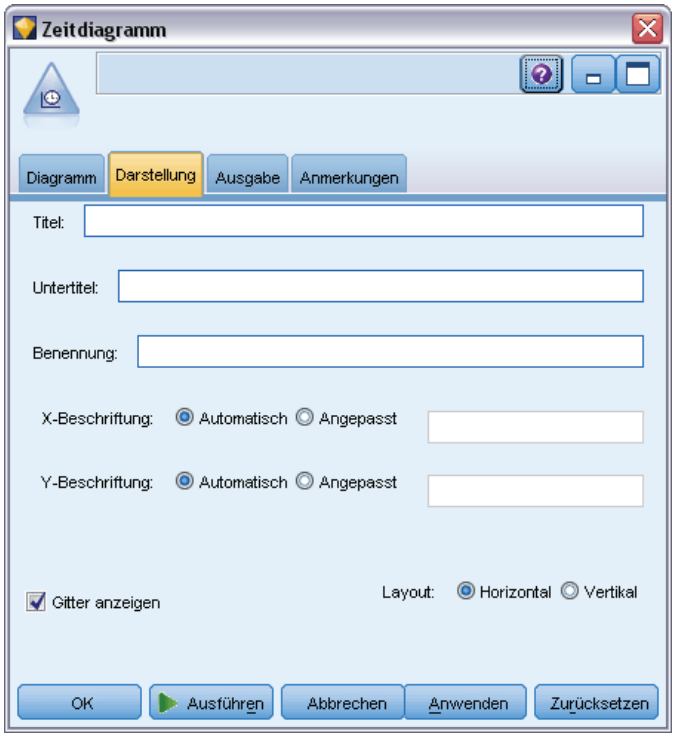
Normalisieren. Hiermit lassen Sie alle *Y*-Werte auf den Bereich 0–1 zur Darstellung im Diagramm skalieren. Durch Normalisieren können Sie die Beziehung zwischen Linien untersuchen, die im Diagramm ansonsten aufgrund von Unterschieden im Wertebereich für die einzelnen Reihen verdeckt sind. Das Normalisieren wird bei der Darstellung mehrerer Linien im selben Diagramm und für den Vergleich von Plots in nebeneinander angeordneten Teilfenstern empfohlen. (Eine Normalisierung ist nicht erforderlich, wenn alle Datenwerte in einen ähnlichen Bereich fallen.)

Anzeigen. Wählen Sie mindestens ein Element aus, das in Ihrem Plot angezeigt werden soll. Sie können aus Linien, Punkten und Glättungselementen (LOESS-Smoother) wählen. Glättungselemente sind nur verfügbar, wenn die Reihen in gesonderten Fenstern angezeigt werden. Standardmäßig ist das Linienelement ausgewählt. Denken Sie daran, mindestens ein Plotelement auszuwählen, bevor Sie den Grafikknoten ausführen; andernfalls gibt das System eine Fehlermeldung aus, die angibt, dass Sie keine Elemente für den Plot ausgewählt haben.

Datensätze beschränken. Wählen Sie diese Option, wenn Sie die Anzahl der zu plottenden Datensätze begrenzen möchten. Geben Sie unter der Option Maximale Anzahl der Datensätze für Plot die Anzahl der zu plottenden Datensätze ein (Lesebeginn ist der Anfang der Datendatei). Standardmäßig ist dieser Wert auf 2.000 gesetzt. Wenn Sie die letzten *n* Datensätze in Ihrer Datendatei plotten möchten, können Sie vor diesen Knoten einen Sortierknoten schalten, um die Datensätze in zeitlich absteigender Reihenfolge zu ordnen.

Zeitdiagramm – Registerkarte "Darstellung"

Abbildung 5-76
Einstellungen auf der Registerkarte "Darstellung" für einen Zeitdiagrammknoten



Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Benennung. Dient zur Eingabe des Texts, der zur Benennung des Diagramms verwendet werden soll.

X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *x*-Achsen-Beschriftung (horizontal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *y*-Achsen-Beschriftung (vertikal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

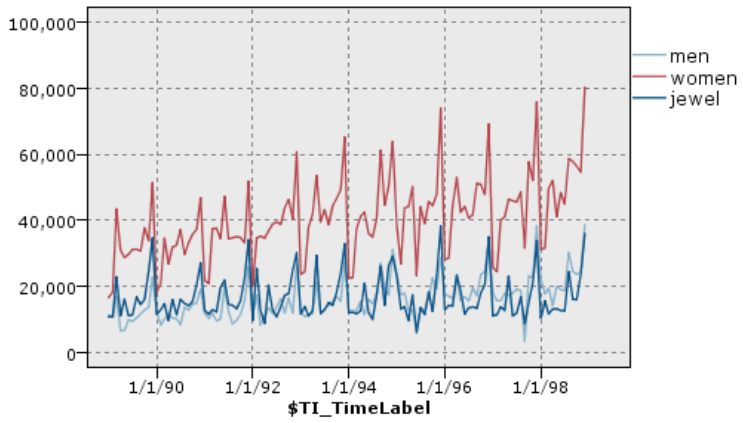
Gitter anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie ein Gitter hinter dem Plot oder dem Diagramm einblenden, das die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Das Gitter wird stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Layout. Nur bei Zeitdiagrammen können Sie angeben, ob die Zeitwerte entlang einer horizontalen oder einer vertikalen Achse dargestellt werden sollen.

Verwendung eines Zeitdiagramms

Sobald Sie ein Zeitdiagramm erstellt haben, stehen verschiedene Optionen zur Auswahl, mit denen Sie die Darstellung des Diagramms anpassen und Knoten für die weitere Analyse erzeugen können. Für weitere Informationen siehe Thema Untersuchen von Diagrammen auf S. 360.

Abbildung 5-77
Darstellung der Verkäufe bei Herren- und Damenbekleidung und Schmuck im Laufe der Zeit



Sobald Sie ein Zeitdiagramm erstellt, Abschnitte definiert und die Ergebnisse untersucht haben, können Sie mit den Optionen im Menü “Generieren” und im Kontextmenü verschiedene Auswahl- und Ableitungsknoten erstellen. Für weitere Informationen siehe Thema Generieren von Knoten aus Diagrammen auf S. 370.

Evaluationsknoten

Der Evaluationsknoten eröffnet eine unkomplizierte Möglichkeit, Vorhersagemodelle auszuwerten und miteinander zu vergleichen, um so das am besten geeignete Modell für die Anwendung zu ermitteln. Evaluationsdiagramme zeigen die Leistung der Modelle beim Vorhersagen bestimmter Ergebnisse. Hierzu werden Datensätze auf der Grundlage des vorhergesagten Werts und der Konfidenz der Vorhersage sortiert. Die Datensätze werden dabei in gleich große Gruppen (**Quantile**) aufgeteilt; anschließend wird der Wert des Geschäftskriteriums für jedes Quantil geplottet, vom höchsten Wert bis zum niedrigsten Wert. Mehrere Modelle werden als separate Linien im Plot dargestellt.

Zur Handhabung der Ergebnisse wird ein bestimmter Wert oder Wertebereich als **Treffer** definiert. Ein Treffer weist in der Regel auf einen gewissen Erfolg hin (z. B. auf einen Verkauf an einen Kunden) oder auf ein relevantes Ereignis (z. B. auf eine bestimmte medizinische Diagnose). Auf der Registerkarte “Optionen” des Dialogfelds können Sie Trefferkriterien definieren oder Sie können die standardmäßigen Trefferkriterien verwenden:

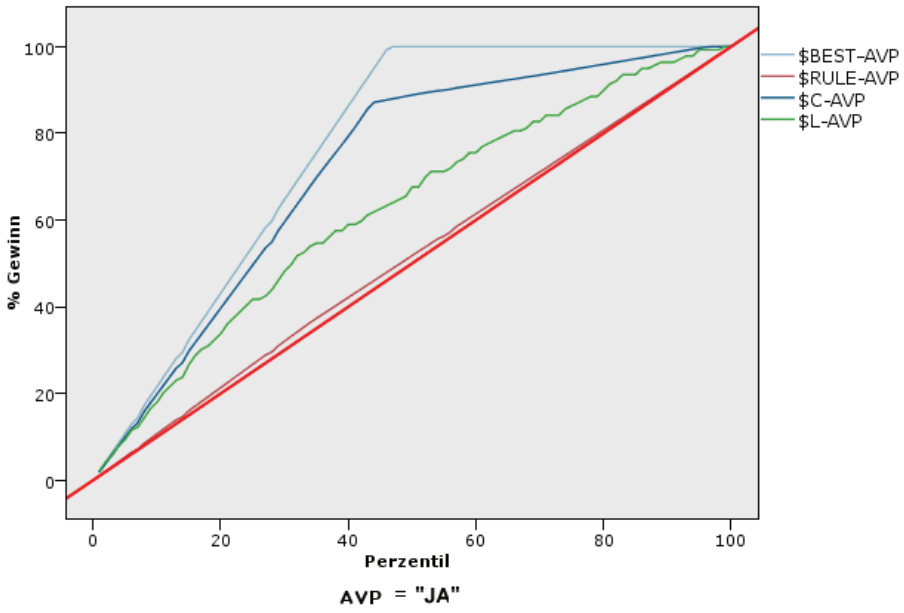
- **Flag**-Ausgabefelder sind unkompliziert; ein Treffer steht für *wahre* Werte.
- Bei **Nominal**-Ausgabefeldern definiert der erste Wert im Set einen Treffer.
- Bei **Stetig**-Ausgabefeldern entspricht ein Treffer einem Wert, der größer ist als der Mittelpunkt des Bereichs für das betreffende Feld.

Es stehen fünf Typen von Evaluationsdiagrammen zur Auswahl, bei denen der Schwerpunkt jeweils auf einem anderen Auswertungskriterium liegt.

Gewinndiagramme

Gewinne sind definiert als der Anteil an allen Treffern, der in den einzelnen Quantilen vorliegt. Die Gewinne werden wie folgt berechnet: (Anzahl der Treffer im Quantil / Gesamtanzahl der Treffer) × 100 %.

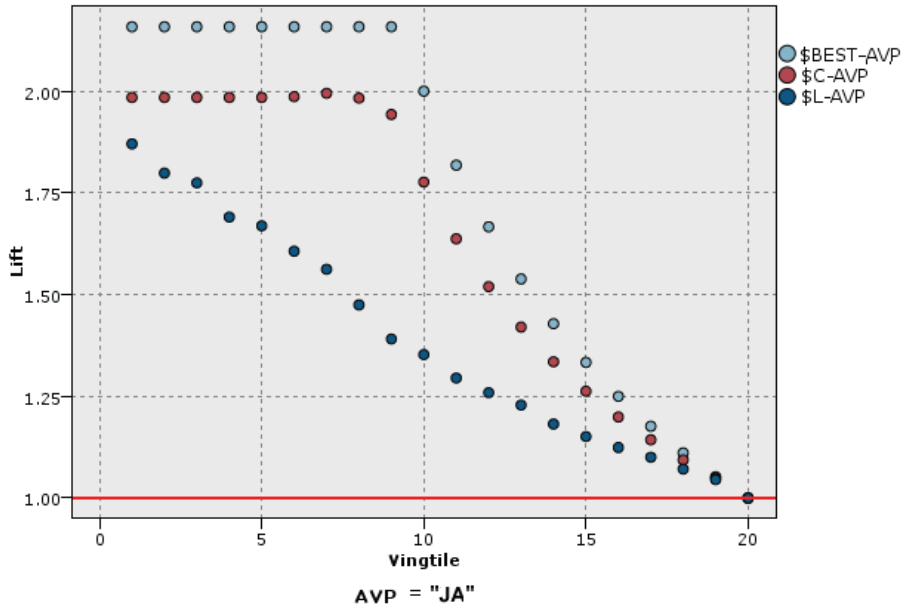
Abbildung 5-78
Gewinndiagramm (kumulativ) mit Basis, bester Linie und Geschäftsregel



Lift Charts

Beim Lift wird der Prozentsatz der Datensätze in jedem Quantil, die als Treffer gelten, mit dem Gesamtprozentsatz der Treffer in den Trainingsdaten verglichen. Die Berechnung läuft wie folgt ab: (Treffer im Quantil / Datensätze im Quantil) / (Gesamtanzahl der Treffer / Gesamtanzahl der Datensätze).

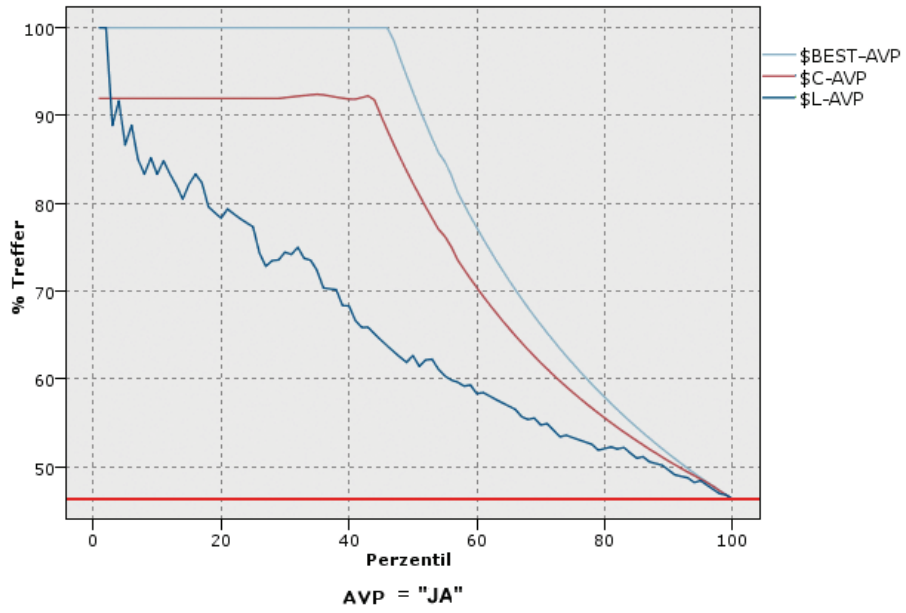
Abbildung 5-79
Lift Chart (kumulativ) mit Punkten und bester Linie



Trefferdiagramme

Treffer bezeichnen einfach den Prozentsatz der Datensätze im Quantil, die als Treffer gelten. Die Treffer werden wie folgt berechnet: (Treffer im Quantil/Datensätze im Quantil) × 100 %.

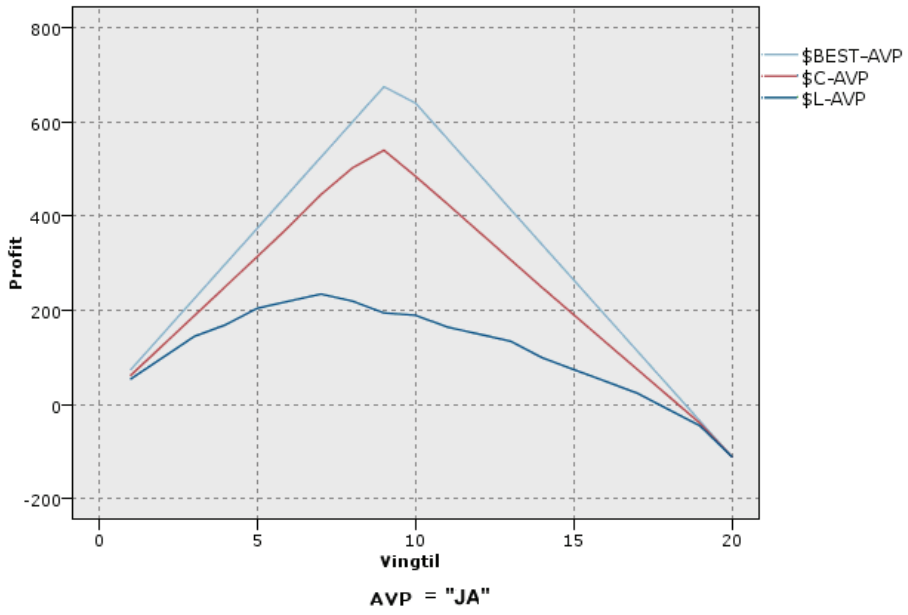
Abbildung 5-80
Trefferdiagramm (kumulativ) mit bester Linie



Profitdiagramme

Der Profit entspricht dem **Umsatz** für jeden Datensatz abzüglich der **Kosten** für den betreffenden Datensatz. Die Profite für ein Quantil entsprechen einfach der Summe der Profite für alle Datensätze im Quantil. Umsätze gelten definitionsgemäß nur für Treffer, Kosten dagegen für alle Datensätze. Die Profite und Kosten können fest sein oder auch durch Felder in den Daten definiert werden. Die Profite werden wie folgt berechnet: (Summe des Umsatzes für die Datensätze im Quantil − Summe der Kosten für die Datensätze im Quantil).

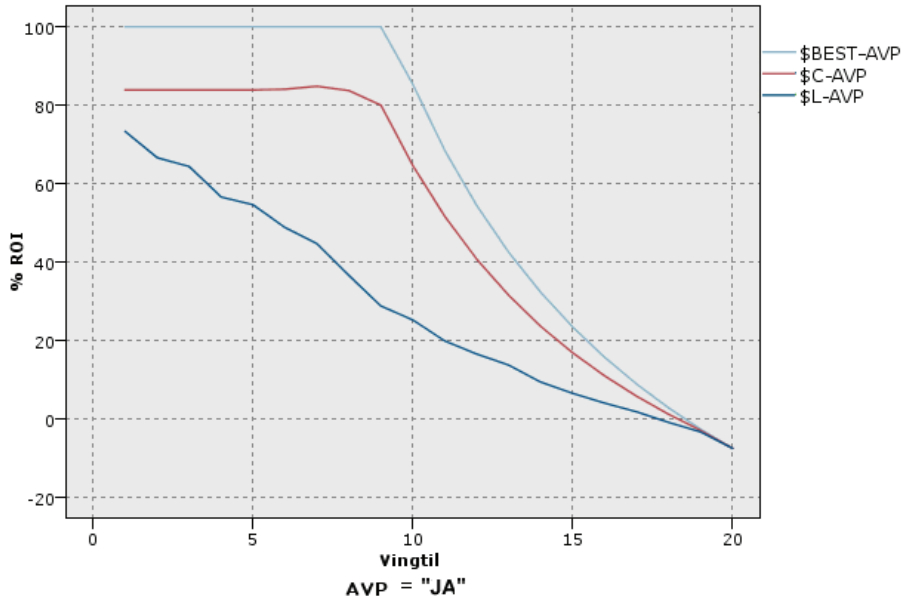
Abbildung 5-81
Profitdiagramm (kumulativ) mit bester Linie



ROI-Diagramme

Der ROI (die Kapitalrendite) weist gewisse Ähnlichkeiten mit dem Profit auf; auch hier wird eine Definition für Umsätze und Kosten herangezogen. Beim ROI werden die Profite mit den Kosten für das Quantil verglichen. Der ROI wird wie folgt berechnet: (Profite für das Quantil/Kosten für das Quantil) × 100 %.

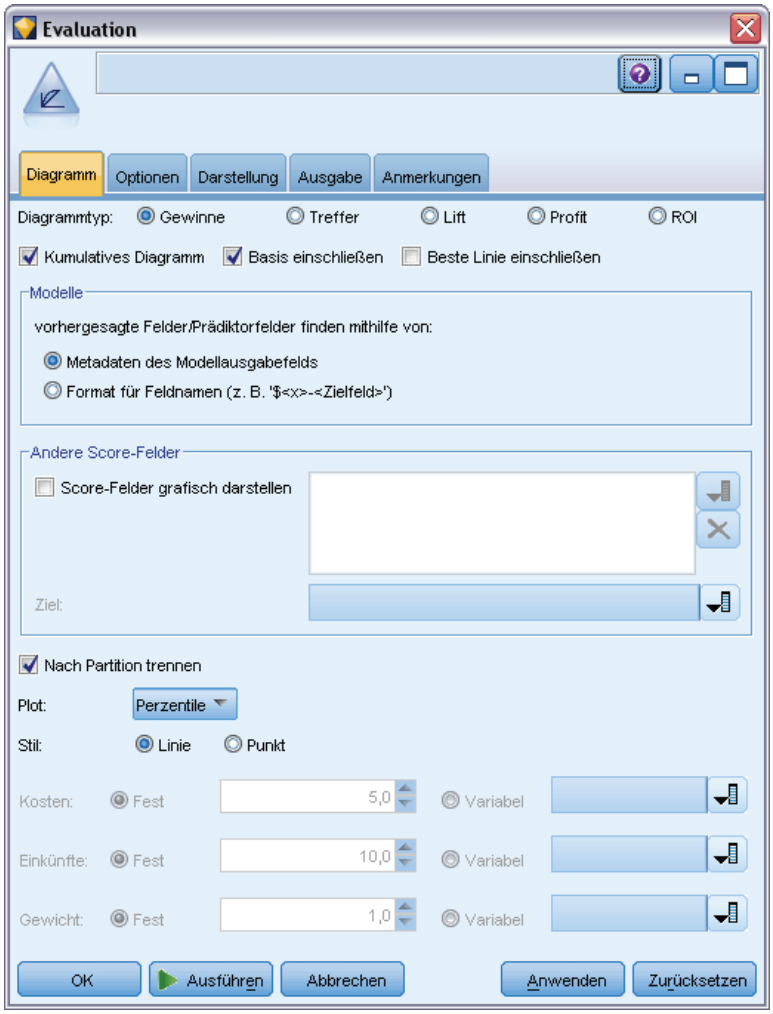
Abbildung 5-82
ROI-Diagramm (kumulativ) mit bester Linie



Auch Evaluationsdiagramme können kumulativ sein, sodass jeder Punkt dem Wert für das entsprechende Quantil zuzüglich aller höheren Quantile entspricht. Kumulative Diagramme geben die Gesamtleistung von Modellen in der Regel besser wieder; nicht kumulative Diagramme weisen dagegen häufig auf bestimmte Problembereiche in den Modellen hin.

Evaluation – Registerkarte "Plot"

Abbildung 5-83
Einstellungen auf der Registerkarte "Plot" für Evaluationsknoten



Diagrammtyp. Wählen Sie einen der folgenden Typen aus: Gewinne, Treffer, Lift, Profit oder ROI (Kapitalrendite).

Kumulatives Diagramm. Es wird ein kumulatives Diagramm erstellt. Die Werte in kumulativen Diagrammen werden für jedes Quantil zuzüglich aller höheren Quantile geplottet.

Basis einschließen. In das Diagramm wird eine Basis aufgenommen, die auf eine völlig zufällige Verteilung der Treffer hinweist, sodass die Konfidenz irrelevant wird. (Bei Profit- und ROI-Diagrammen ist die Option Basis einschließen nicht verfügbar.)

Beste Linie einschließen. In das Diagramm wird eine beste Linie aufgenommen, die auf völlige Konfidenz hinweist (Treffer in 100 % aller Fälle).

Vorhergesagte Felder/Prädiktorfelder finden mithilfe von. Wählen Sie entweder Metadaten des Modellausgabefelds, um anhand der zugehörigen Metadaten nach den vorhergesagten Feldern im Diagramm zu suchen, oder wählen Sie Format für Feldnamen , um nach diesen Feldern anhand ihres Namens zu suchen.

Score-Felder grafisch darstellen. Aktivieren Sie dieses Kontrollkästchen, um das Auswahlwerkzeug für Score-Felder zu aktivieren. Wählen Sie anschließend mindestens ein Bereichs-Score-Feld bzw. stetiges Score-Feld aus, also Felder, bei denen es sich nicht um Vorhersagemodelle im strengen Wortsinn handelt, die jedoch möglicherweise bei der Rangordnung der Datensätze hinsichtlich ihrer Trefferneigung nützlich sein könnten. Der Evaluationsknoten kann alle Kombinationen von einem oder mehreren Score-Feldern mit einem oder mehreren Vorhersagemodellen vergleichen. Ein typisches Beispiel kann der Vergleich mehrerer RFM-Felder mit dem besten vorhandenen Vorhersagemodell sein.

Ziel. Wählen Sie mithilfe der Feldauswahl das Zielfeld aus. Sie können jedes beliebige instanziierte Flag- oder nominales Feld mit mindestens zwei Werten auswählen.

Hinweis: Dieses Zielfeld gilt nur für das Scoren von Feldern (Vorhersagemodelle definieren ihre eigenen Ziele) und wird ignoriert, wenn auf der Registerkarte “Optionen” ein Trefferkriterium festgelegt wurde.

Nach Partition aufteilen. Wenn Datensätze mithilfe eines Partitionsfelds in Trainings-, Test- und Validierungsstichproben aufgeteilt werden, lassen Sie mit dieser Option ein separates Evaluationsdiagramm für die einzelnen Partitionen anzeigen. Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 auf S. 208.

Hinweis: Wenn Sie eine Partition aufteilen, werden Datensätze mit Nullwerten im Partitionsfeld von der Auswertung ausgeschlossen. Dieses Problem tritt nicht auf, wenn Sie einen Partitionsknoten verwenden, weil diese Knoten keine Nullwerte erzeugen.

Diagramm. Wählen Sie die Größe der Quantile, die im Diagramm geplottet werden sollen, in der Dropdown-Liste aus. Die folgenden Optionen stehen zur Auswahl: Quartile, Quintile, Dezile, Vingtile, Perzentile und 1000-tile.

Stil. Wählen Sie die Option Linie oder Punkt.

Profit- und ROI-Diagramme. Bei Profit- und ROI-Diagrammen stehen weitere Steuerelemente zur Verfügung, mit denen Sie die Kosten, den Umsatz und die Gewichte festlegen können.

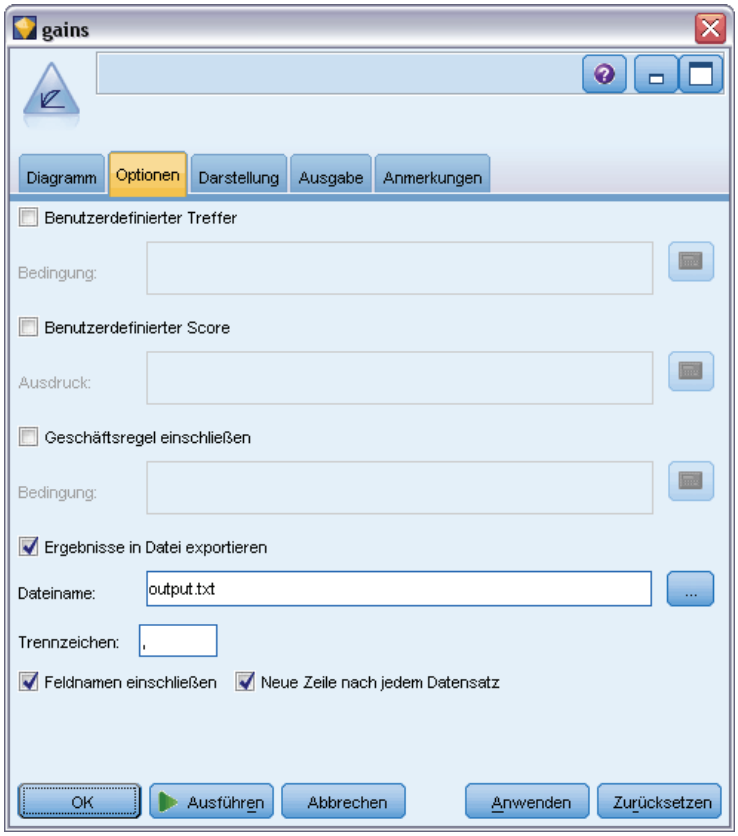
- **Kosten.** Geben Sie die Kosten für die einzelnen Datensätze an. Wählen Sie die Option Fest oder Variabel für den Umsatz. Bei festen Kosten geben Sie den Wert der Kosten ein. Bei variablen Kosten klicken Sie auf die Feldauswahl-Schaltfläche und bestimmen Sie ein Feld als Kostenfeld.
- **Umsatz.** Geben Sie den Umsatz für die einzelnen Datensätze ein, die als Treffer gelten. Wählen Sie die Option Fest oder Variabel für den Umsatz. Bei einem festen Umsatz geben Sie den Wert des Umsatzes ein. Bei einem variablen Umsatz klicken Sie auf die Feldauswahl-Schaltfläche und bestimmen Sie ein Feld als Umsatzfeld.
- **Gewicht.** Wenn die Datensätze in den Daten für mehrere Einheiten stehen, können Sie die Ergebnisse mithilfe der Häufigkeitsgewichtungen anpassen. Geben Sie die Gewichtung für die einzelnen Datensätze im Feld Fest oder Variabel an. Bei einer festen Gewichtung geben Sie den Wert für das Gewicht an (die Anzahl der Einheiten pro Datensatz). Bei variablen

Gewichtungen klicken Sie auf die Feldauswahl—Schaltfläche und bestimmen Sie ein Feld als Gewichtsfeld.

Evaluation – Registerkarte "Optionen"

Auf der Registerkarte "Optionen" für Evaluationsdiagramme können Sie die Treffer, die Scoring-Kriterien und die Geschäftsregeln für das Diagramm auf flexible Weise festlegen. Des Weiteren stehen Optionen zur Verfügung, mit denen Sie die Ergebnisse der Modellauswertung exportieren.

Abbildung 5-84
Einstellungen auf der Registerkarte "Optionen" für Evaluationsknoten



Benutzerdefinierter Treffer. Geben Sie eine benutzerdefinierte Bedingung für einen Treffer an. Diese Option ist von Nutzen, wenn das relevante Ereignis definiert werden soll (also nicht aus dem Typ des Zielfelds und der Reihenfolge der Werte abgeleitet).

- **Bedingung.** Wenn Sie oben die Option Benutzerdefinierter Treffer wählen, muss ein CLEM-Ausdruck für eine Trefferbedingung festgelegt werden. Beispiel: @TARGET = "YES" ist eine gültige Bedingung, aus der hervorgeht, dass der Wert *Yes* im Zielfeld als Treffer bei der Auswertung gezählt wird. Die angegebene Bedingung wird für alle Zielfelder herangezogen. Geben Sie die gewünschte Bedingung in das Feld ein oder erzeugen Sie einen Bedingungsausdruck mit Expression Builder. Falls die Daten instanziiert wurden, können Sie die Werte direkt aus Expression Builder einfügen.

Benutzerdefinierter Score. Geben Sie eine Bedingung ein, mit der die Fälle gescort werden, bevor sie Quantilen zugeordnet werden. Der Standard-Score wird aus dem vorhergesagten Wert und der Konfidenz berechnet. Im Feld "Ausdruck" können Sie einen benutzerdefinierten Scoring-Ausdruck erstellen.

- **Ausdruck.** Geben Sie einen CLEM-Ausdruck für das Scoring an. Wenn beispielsweise die numerische Ausgabe im Bereich 0–1 so geordnet wird, dass niedrige Werte besser eingestuft werden als hohe Werte, können Sie einen Treffer als `@TARGET < 0.5` definieren und den zugehörigen Score als `1 • @PREDICTED`. Der Score-Ausdruck muss in einem numerischen Wert resultieren. Geben Sie die gewünschte Bedingung in das Feld ein oder erzeugen Sie einen Bedingungsausdruck mit Expression Builder.

Geschäftsregel einschließen. Geben Sie eine Regelbedingung gemäß den relevanten Kriterien an. Beispielsweise können Sie eine Regel für alle Fälle anzeigen, in denen gilt: `mortgage = "Y" and income >= 33000`. Geschäftsregeln werden im Diagramm gezeichnet und im Schlüssel als *Regel* gekennzeichnet.

- **Bedingung.** Geben Sie einen CLEM-Ausdruck zur Definition einer Geschäftsregel im Ausgabediagramm an. Geben Sie den gewünschten Bedingungsausdruck in das Feld ein oder erzeugen Sie einen Bedingungsausdruck mit Expression Builder. Falls die Daten instanziiert wurden, können Sie die Werte direkt aus Expression Builder einfügen.

Ergebnisse in Datei exportieren. Die Ergebnisse der Modellauswertung werden in eine Textdatei mit Trennzeichen exportiert. Sie können diese Datei einlesen und spezielle Analysen für die berechneten Werte vornehmen. Legen Sie die folgenden Optionen für den Export fest:

- **Dateiname.** Geben Sie den Dateinamen für die Ausgabedatei ein. Mit der Auslassungsschaltfläche (...) wechseln Sie zum gewünschten Ordner.
- **Trennzeichen.** Geben Sie das Zeichen ein (z. B. Komma oder Leerschritt), das als Feldtrennzeichen verwendet werden soll.

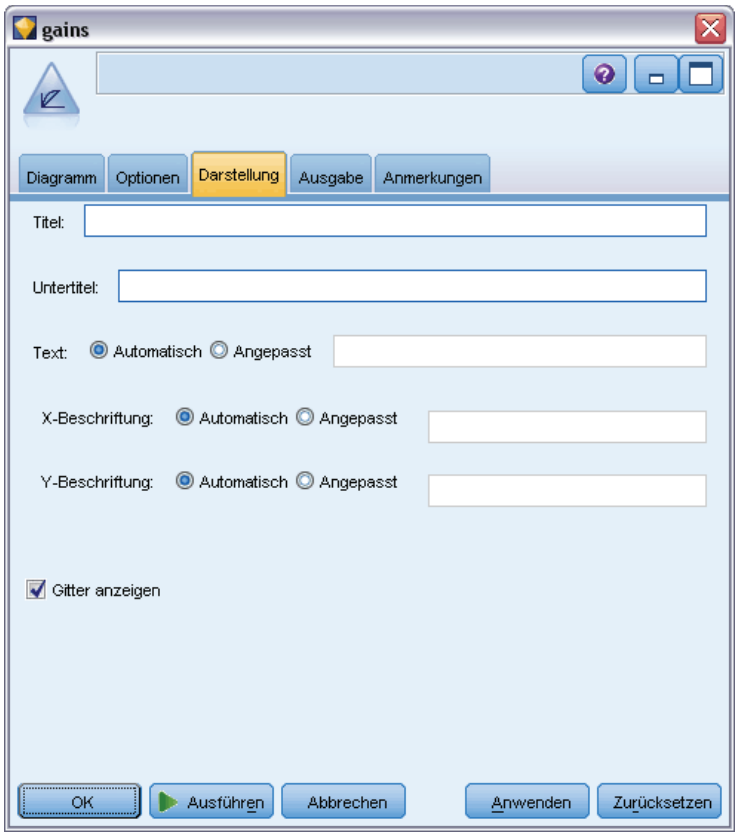
Feldnamen einschließen. Die Feldnamen werden in die erste Zeile der Ausgabedatei eingetragen.

Neue Zeile nach jedem Datensatz. Jeder Datensatz beginnt in einer neuen Zeile.

Evaluation – Registerkarte "Darstellung"

Vor der Diagrammerstellung können Sie Darstellungsoptionen angeben.

Abbildung 5-85
Einstellungen auf der Registerkarte "Darstellung" für Evaluationsknoten



Titel. Dient zur Eingabe des Texts, der als Titel des Diagramms verwendet werden soll.

Untertitel. Dient zur Eingabe des Texts, der als Untertitel des Diagramms verwendet werden soll.

Text. Akzeptieren Sie entweder die automatisch generierte Beschriftung oder wählen Sie Angepasst, um eine benutzerdefinierte Beschriftung anzugeben.

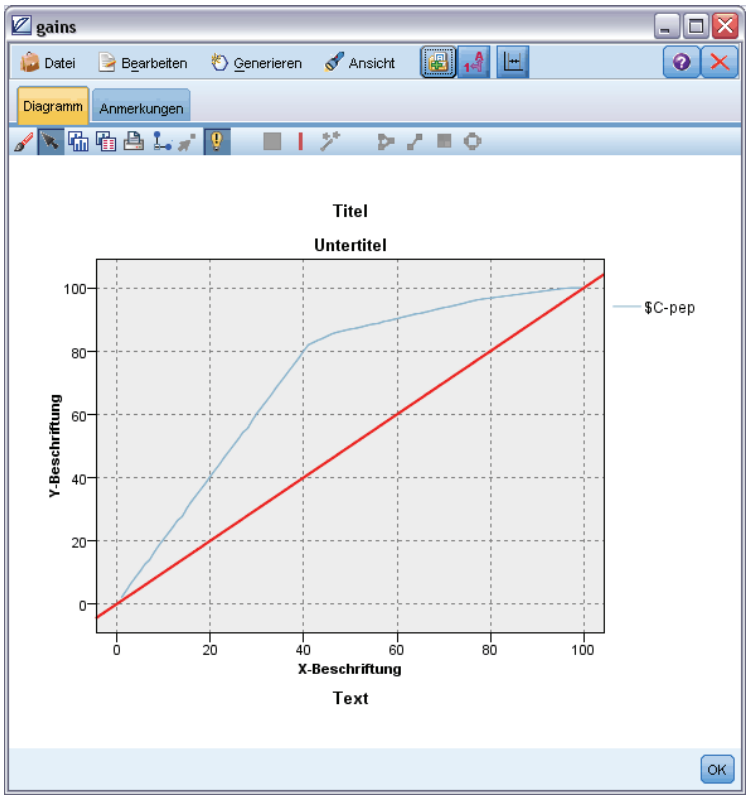
X-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *x*-Achsen-Beschriftung (horizontal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

Y-Beschriftung. Akzeptieren Sie entweder die automatisch generierte *y*-Achsen-Beschriftung (vertikal) oder wählen Sie Angepasst, um eine Beschriftung anzugeben.

Gitter anzeigen. Diese Option ist standardmäßig aktiviert. Hiermit lassen Sie ein Gitter hinter dem Plot oder dem Diagramm einblenden, das die Bestimmung der Bereichs- und Bandabschnittpunkte erleichtert. Das Gitter wird stets in weißer Farbe angezeigt; bei einem weißen Diagrammhintergrund erfolgt die Anzeige in Grau.

Das folgende Beispiel zeigt, wo sich die Darstellungsoptionen im Diagramm befinden.

Abbildung 5-86
Position der Diagramm-Darstellungsoptionen bei einem Evaluationsdiagramm



Lesen der Ergebnisse einer Modellauswertung

Die Interpretation eines Evaluationsdiagramms ist zu einem gewissen Grad abhängig vom jeweiligen Diagrammtyp; einige Merkmale sind jedoch allen Evaluationsdiagrammen gemeinsam. Bei kumulativen Diagrammen weisen höhere Linien auf bessere Modelle hin, insbesondere auf der linken Seite des Diagramms. Werden mehrere Modelle miteinander verglichen, schneiden sich die Linien häufig, sodass ein Modell in einem Teil des Diagramms höher ist und ein anderes Modell in einem anderen Diagrammteil. In diesem Fall sollten Sie den erforderlichen Teil der Stichprobe berücksichtigen (mit dem ein Punkt auf der *x*-Achse definiert wird), wenn Sie sich für ein bestimmtes Modell entscheiden.

Die meisten nicht kumulativen Diagramme sind einander sehr ähnlich. Bei guten Modellen sind nicht kumulative Diagramme auf der linken Seite des Diagramms hoch und auf der rechten Seite des Diagramms niedrig. (Zeigt ein nicht kumulatives Diagramm ein Sägezahn-Muster, können Sie das Diagramm glätten, indem Sie die Anzahl der zu plottenden Quantile verringern und das Diagramm neu zeichnen lassen.) Ein Abfall auf der linken Seite des Diagramms oder eine Spitze auf der rechten Seite weist auf Bereiche hin, in denen das Modell nur wenig aussagekräftig ist. Eine gerade Linie über das ganze Diagramm entsteht, wenn ein Modell im Grunde genommen keinerlei Informationen liefert.

Gewinndiagramme. Kumulative Gewinndiagramme beginnen stets bei 0 %, verlaufen von links nach rechts und enden bei 100 %. Bei einem guten Modell steigt die Gewinntabelle steil in Richtung 100 % an und flacht dann ab. Bei einem Modell ohne Informationsgehalt verläuft eine diagonale Linie von links unten nach rechts oben. (Dies ist im Diagramm sichtbar, wenn Sie die Option Basis einschließen aktiviert haben.)

Lift Charts. Kumulative Lift Charts beginnen in der Regel bei einem Wert über 1,0 und fallen von links nach rechts allmählich ab. Die rechte Kante des Diagramms entspricht dem gesamten Daten-Set; das Verhältnis der Treffer in den kumulativen Quantilen zu den Treffern in den Daten beträgt 1,0. Bei einem guten Modell sollte der Lift auf der linken Seite deutlich über 1,0 beginnen, von links nach rechts auf einem hohen Niveau verbleiben und dann auf der rechten Seite des Diagramms abrupt auf 1,0 fallen. Bei einem Modell ohne Informationsgehalt liegt die Linie im gesamten Diagramm bei einem Wert um 1,0. (Falls die Option Basis einschließen aktiviert ist, wird im Diagramm eine horizontale Linie bei 1,0 als Referenz eingeblendet.)

Trefferdiagramme. Kumulative Trefferdiagramme besitzen große Ähnlichkeit mit Lift Charts, mit Ausnahme der Skalierung. Trefferdiagramme beginnen in der Regel bei einem Wert nahe 100 % und fallen dann allmählich auf die Gesamttrefferrate (Gesamtanzahl der Treffer / Gesamtanzahl der Datensätze) auf der rechten Seite des Diagramms ab. Bei einem guten Modell beginnt die Linie auf der linken Seite genau oder nahe bei 100 %, von links nach rechts auf einem hohen Niveau verbleiben und dann auf der rechten Seite des Diagramms abrupt auf die Gesamttrefferrate fallen. Bei einem Modell ohne Informationsgehalt liegt die Linie im gesamten Diagramm bei einem Wert um die Gesamttrefferrate. (Falls die Option Basis einschließen aktiviert ist, wird im Diagramm eine horizontale Linie bei der Gesamttrefferrate als Referenz eingeblendet.)

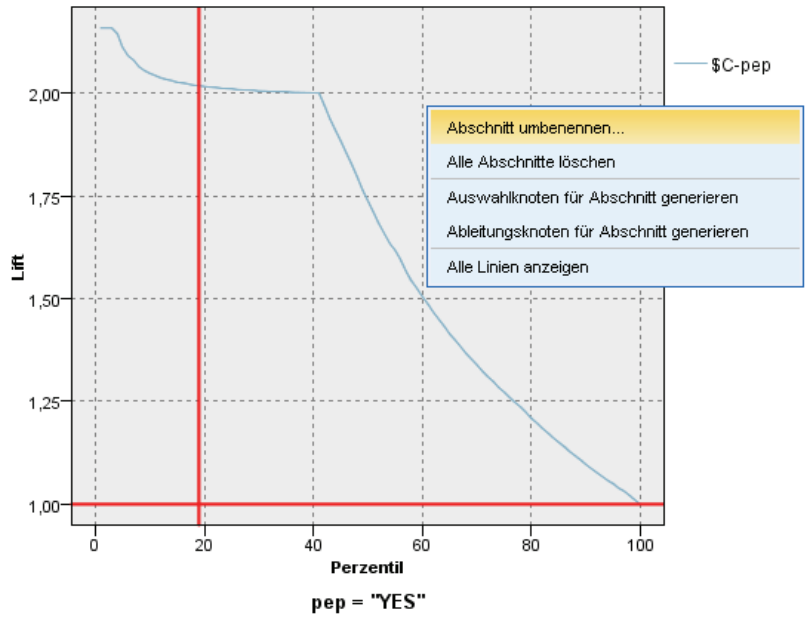
Profitdiagramme. Kumulative Profitdiagramme zeigen die Summe der Profite, wenn Sie die Größe der ausgewählten Stichprobe (von links nach rechts) erhöhen. Profitdiagramme beginnen in der Regel in der Nähe von 0, steigen dann von links nach rechts stetig bis zu einer Spitze oder einem hohen Niveau in der Mitte an und fallen dann zur rechten Kante des Diagramms hin ab. Bei einem guten Modell zeigen die Profite eine klar ausgeprägte Spitze im Mittelteil des Diagramms. Bei einem Modell ohne Informationsgehalt verläuft die Linie relativ gerade; die Linie kann ansteigen, abfallen oder auf demselben Niveau verbleiben, abhängig von der vorliegenden Kosten-Umsatz-Struktur.

ROI-Diagramme. Kumulative ROI-Diagramme (Kapitalrendite) verlaufen in der Regel ähnlich wie Trefferdiagramme und Lift Charts, mit Ausnahme der Skalierung. ROI-Diagramme beginnen in der Regel bei einem Wert oberhalb von 0 % und fallen dann allmählich auf den Gesamt-ROI für das gesamte Daten-Set ab; dieser Wert kann durchaus auch negativ sein. Bei einem guten Modell sollte die Linie auf der linken Seite deutlich über 0 % beginnen, von links nach rechts auf einem hohen Niveau verbleiben und dann auf der rechten Seite des Diagramms relativ abrupt auf den Gesamt-ROI abfallen. Bei einem Modell ohne Informationsgehalt liegt die Linie im gesamten Diagramm beim Gesamt-ROI.

Verwendung eines Evaluationsdiagramms

Evaluationsdiagramme können auf ähnliche Weise mithilfe der Maus untersucht werden wie Histogramme oder Sammlungsdiagramme. Die *x*-Achse bezeichnet die Modell-Scores in den angegebenen Quantilen (z. B. Vingtile oder Dezile).

Abbildung 5-87
Arbeiten mit einem Evaluationsdiagramm

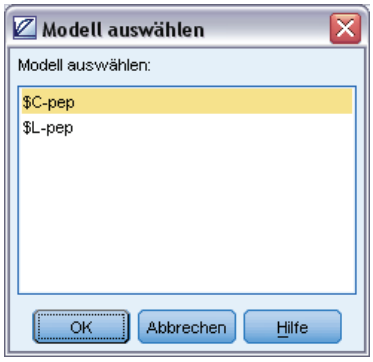


Sie können die *x*-Achse wie bei Histogrammen in Abschnitte aufteilen. Greifen Sie über das Aufteilungssymbol auf die Optionen zu, mit denen die Achse automatisch in gleich große Abschnitte aufgeteilt wird. Für weitere Informationen siehe Thema Untersuchen von Diagrammen auf S. 360.Sollen die Grenzen der Abschnitte manuell bearbeitet werden, wählen Sie im Menü “Bearbeiten” den Befehl Diagrammabschnitte.

Sobald Sie ein Evaluationsdiagramm erstellt, Abschnitte definiert und die Ergebnisse untersucht haben, können Sie mit den Optionen im Menü “Generieren” und im Kontextmenü automatisch verschiedene Knoten auf der Grundlage der Auswahl im Diagramm erstellen. Für weitere Informationen siehe Thema Generieren von Knoten aus Diagrammen auf S. 370.

Wenn Sie Knoten aus einem Evaluationsdiagramm heraus erstellen, werden Sie aufgefordert, ein einzelnes Modell aus den verfügbaren Modellen im Diagramm auszuwählen.

Abbildung 5-88
Auswählen eines Modells zum Erstellen von Knoten



Wählen Sie ein Modell aus und klicken Sie auf OK. Der neue Knoten wird im Stream-Zeichenbereich erzeugt.

Untersuchen von Diagrammen

Während Sie im Bearbeitungsmodus Layout und Erscheinungsbild des Diagramms bearbeiten können, können Sie im Interaktionsmodus eine analytische Untersuchung der im Diagramm dargestellten Daten und Werte vornehmen. Das Hauptziel der Untersuchung besteht in der Analyse der Daten und der anschließenden Identifizierung von Werten mithilfe von Abschnitten, Bereichen und Markierungen zum Generieren von Auswahl-, Ableitungs- oder Balancierungsknoten. Um diesen Modus auszuwählen, wählen Sie in den Menüs die Optionsfolge Ansicht > Interaktionsmodus (oder klicken Sie auf das entsprechende Symbol in der Symbolleiste).

Bei einigen Diagrammen können alle Untersuchungstools verwendet werden, bei anderen dagegen ist nur ein einziges verfügbar. Der Interaktionsmodus umfasst folgende Aktionen:

- Definieren und Bearbeiten von Abschnitten, die zur Aufteilung der Werte entlang einer *x*-Skalenachse verwendet werden. Für weitere Informationen siehe Thema Verwendung von Abschnitten auf S. 361.
- Definieren und Bearbeiten von Bereichen, die zur Identifizierung einer Gruppe von Werten innerhalb der rechteckigen Fläche verwendet werden. Für weitere Informationen siehe Thema Verwenden von Bereichen auf S. 365.
- Markieren von Elementen (und Aufheben von Markierungen) zur manuellen Auswahl von Werten, die zum Generieren eines Auswahl- oder Ableitungsknotens verwendet werden könnten. Für weitere Informationen siehe Thema Verwenden markierter Elemente auf S. 368.
- Generieren von Knoten mithilfe der durch Abschnitte, Bereiche, markierte Elemente und Netzlinks identifizierten Werte zur Verwendung im Stream. Für weitere Informationen siehe Thema Generieren von Knoten aus Diagrammen auf S. 370.

Abbildung 5-89
Diagramm mit Menü zum Generieren

Verwendung von Abschnitten

In jedem Diagramm mit einem metrischen Feld auf der *x*-Achse können Sie vertikale Abschnittslinien zeichnen, um den Wertebereich auf der *x*-Achse aufzuteilen. Bei aus mehreren Fenstern bestehenden Diagrammen wird eine Abschnittslinie, die in einem Fenster gezogen wird, auch in den anderen Fenstern angezeigt.

Nicht bei allen Diagrammen sind Abschnitte zulässig. Zu den Diagrammen, bei denen Abschnitte zulässig sind, gehören: Histogramme, Balkendiagramme und Verteilungen, Plots (Linien-, Streu-, Zeitdiagramme usw.), Sammlungen und Evaluationsdiagramme. Bei Diagrammen mit Fenstereinteilung werden die Abschnitte in allen Fenstern angezeigt. In SPLOMs wird außerdem manchmal eine horizontale Abschnittslinie angezeigt, da die Achse, auf der der Feld-/Variablenabschnitt gezeichnet wurde, vertauscht wurde.

Abbildung 5-90
Diagramm mit drei Abschnitten

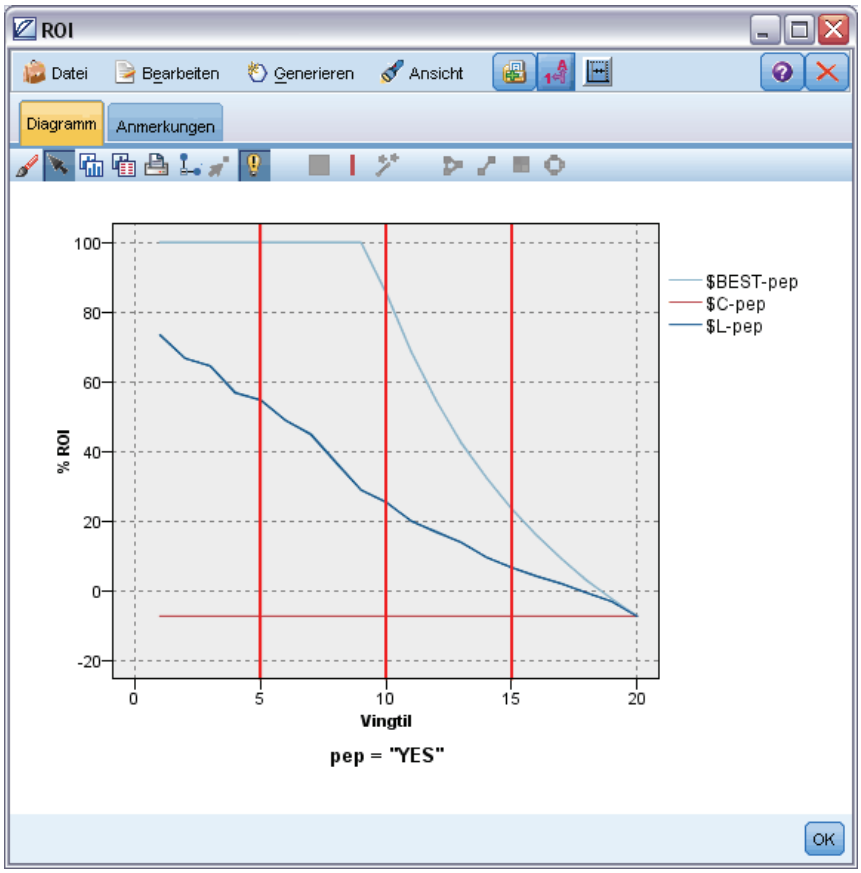
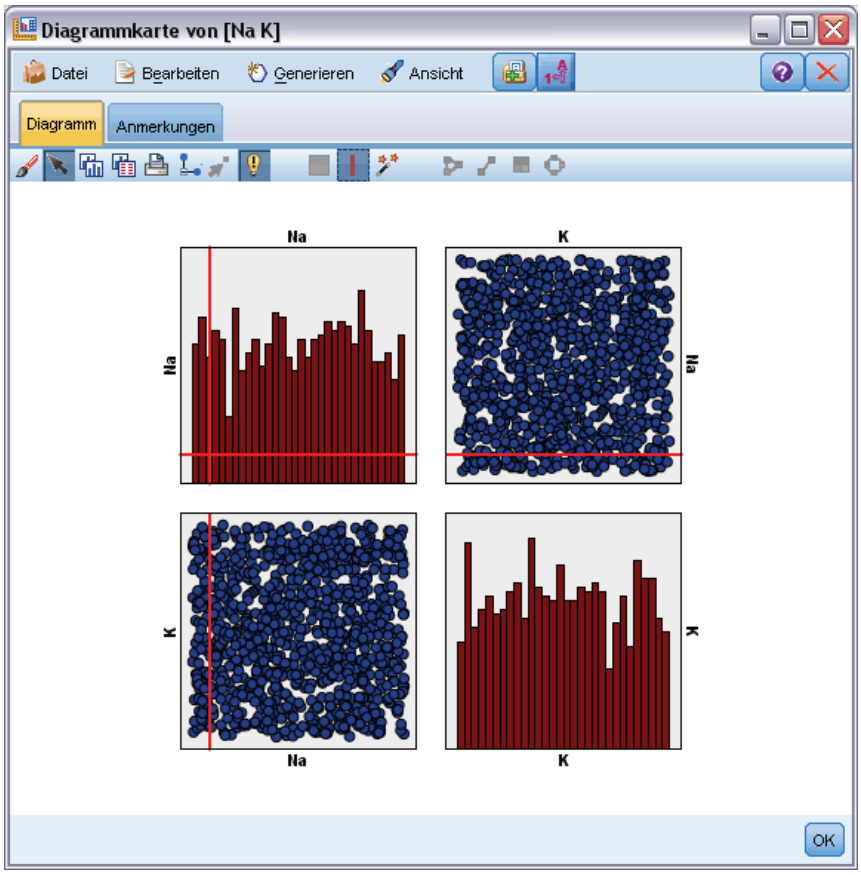


Abbildung 5-91
SPLOM mit Abschnitten



Definieren von Abschnitten

Diagramme ohne Abschnitte werden durch Einfügen einer Abschnittslinie in zwei Abschnitte aufgeteilt. Der Wert der Abschnittslinie stellt den Ausgangspunkt (auch als Untergrenze bezeichnet) des zweiten Abschnitts dar, wenn das Diagramm von links nach rechts gelesen wird. Bei Diagrammen mit zwei Abschnitten wird durch Einfügen einer Abschnittslinie einer der Abschnitte geteilt, wodurch sich drei Abschnitte ergeben. Standardmäßig erhalten die Abschnitte die Bezeichnung *AbschnittN*, wobei *N* für die Anzahl der Abschnitte von links nach rechts auf der *x*-Achse steht.

Nach der Festlegung eines Abschnitts können Sie ihn durch Ziehen und Ablegen auf der *x*-Achse neu positionieren. Weitere Schnellverfahren können für Aufgaben wie Umbenennen, Löschen oder Generieren von Knoten für den betreffenden Abschnitt durch Rechtsklick innerhalb des Abschnitts angezeigt werden.

So definieren Sie Abschnitte:

- Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge Ansicht > Interaktionsmodus.
- Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche “Abschnitt zeichnen”.

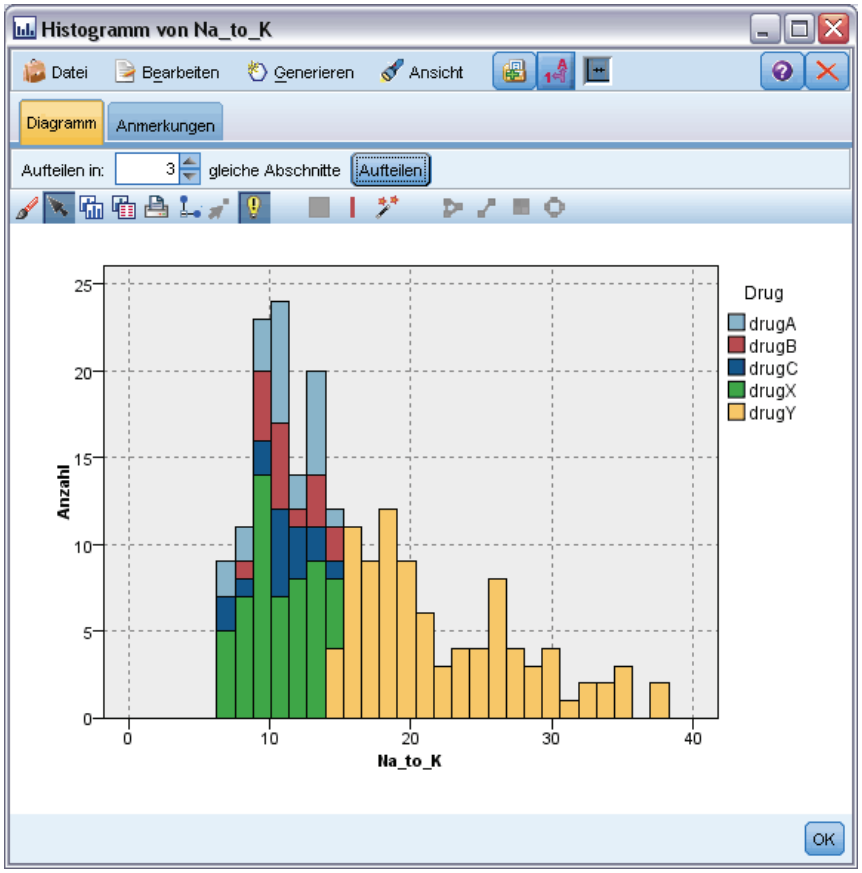
Abbildung 5-92
Symbolleistenschaltfläche "Abschnitt zeichnen"

- Klicken Sie bei Diagrammen, bei denen Abschnitte zulässig sind, auf den Wertepunkt der *x*-Achse, an dem eine Abschnittslinie definiert werden soll.

Hinweis: Alternativ können Sie auf die Symbolleistenschaltfläche Diagramm in Abschnitte teilen klicken, die Anzahl der gewünschten gleich großen Abschnitte eingeben und auf Aufteilen klicken.

Abbildung 5-93
Aufteilungssymbol, mit dem die Symbolleiste um Optionen zum Aufteilen in Abschnitte erweitert wird

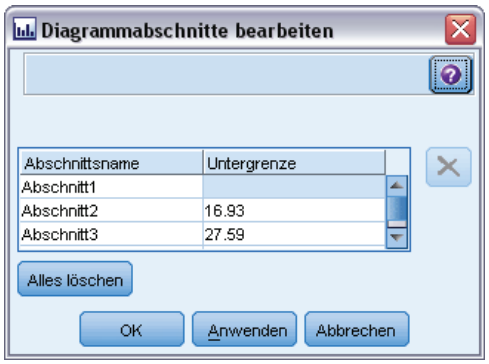
Abbildung 5-94
Symbolleiste zum Erstellen gleich großer Abschnitte mit aktivierten Abschnitten



Bearbeiten, Umbenennen und Löschen von Abschnitten

Die Eigenschaften bestehender Abschnitte können im Dialogfeld "Diagrammabschnitte bearbeiten" oder über die Kontextmenüs im Diagramm selbst bearbeitet werden.

Abbildung 5-95
Dialogfeld "Diagrammabschnitte bearbeiten"



So bearbeiten Sie Abschnitte:

- Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge Ansicht > Interaktionsmodus.
- Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche "Abschnitt zeichnen".
- Wählen Sie in den Menüs die Optionsfolge Bearbeiten > Diagrammabschnitte. Das Dialogfeld "Diagrammabschnitte bearbeiten" wird geöffnet.
- Wenn das Diagramm mehrere Felder enthält (beispielsweise bei SPLOM-Diagrammen), können Sie das gewünschte Feld in der Dropdown-Liste auswählen.
- Sie können einen neuen Abschnitt hinzufügen, indem Sie einen Namen und eine Untergrenze eingeben. Drücken Sie die Eingabetaste, um eine neue Zeile zu beginnen.
- Sie können die Grenze eines Abschnitts durch Anpassung des Werts für Untergrenze bearbeiten.
- Abschnitte können durch Eingabe eines neuen Abschnittsnamens umbenannt werden.
- Sie können Abschnitte löschen, indem Sie die Linie in der Tabelle auswählen und auf die Schaltfläche "Löschen" klicken.
- Klicken Sie auf OK, um die Änderungen zu übernehmen und das Dialogfeld zu schließen.

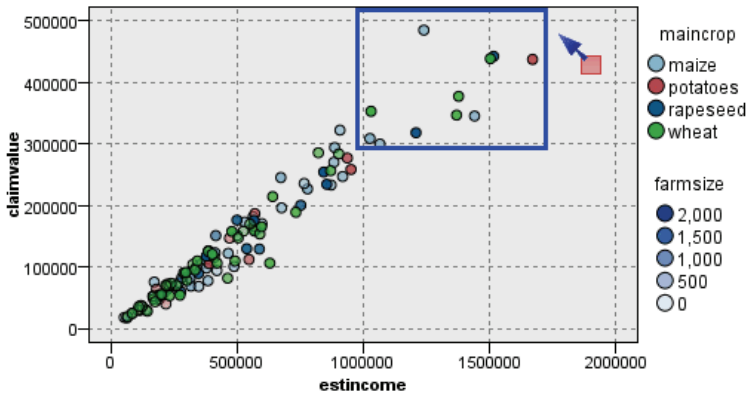
  Hinweis: Alternativ können Sie Abschnitte direkt im Diagramm löschen und umbenennen, indem Sie mit der rechten Maustaste auf die Linie des Abschnitts klicken und die gewünschte Option aus den Kontextmenüs auswählen.

Verwenden von Bereichen

In Diagrammen mit zwei Skalenachsen (oder Bereichsachsen) können Sie Bereiche zeichnen, um Werte innerhalb einer von Ihnen gezeichneten rechteckigen Fläche, dem so genannten Bereich, zu gruppieren. Ein **Bereich** ist ein Teil des Diagramms, der durch einen bestimmten Mindest- und Höchstwert für *X* und *Y* beschrieben wird. Bei aus mehreren Fenstern bestehenden Diagrammen wird ein Bereich, der in einem Fenster gezeichnet wird, auch in den anderen Fenstern angezeigt.

Nicht bei allen Diagrammen sind Bereiche zulässig. Zu den Diagrammen, bei denen Bereiche zulässig sind, gehören: Plots (Linien-, Streu-, Blasen-, Zeitdiagramme usw.), SPLOM und Sammlungen. Diese Bereiche werden im (X,Y)-Raum gezeichnet und können daher nicht in 1D-Plots, 3D-Plots und animierten Plots definiert werden. Bei Diagrammen mit Fenstereinteilung werden die Bereiche in allen Fenstern angezeigt. Bei einer Streudiagramm-Matrix (SPLOM) wird ein Bereich in den zugehörigen oberen Plots angezeigt, nicht jedoch in den diagonalen Plots, da diese nur ein einziges metrisches Feld zeigen.

Abbildung 5-96
Definieren eines Bereichs mit hohen Forderungswerten



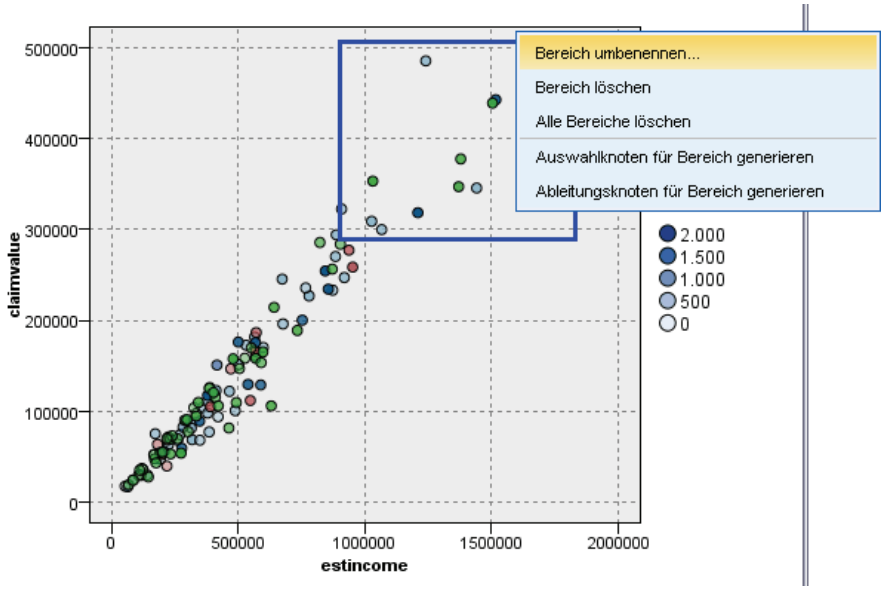
Definieren von Bereichen

Beim Definieren von Bereichen wird immer eine Gruppierung von Werten erstellt. Standardmäßig erhält jeder neue Bereich die Bezeichnung *Bereich<N>*, wobei *N* für die Anzahl der bereits erstellten Bereiche steht.

Nachdem Sie einen Bereich definiert haben, können Sie mit der rechten Maustaste auf die Bereichslinie klicken, um einige grundlegende Schnellverfahren anzuzeigen. Durch Rechtsklick innerhalb des Abschnitts (nicht auf die Linie) können zahlreiche weitere Schnellverfahren für Aufgaben wie Umbenennen, Löschen oder Generieren von Auswahl- und Ableitungsknoten für den betreffenden Bereich angezeigt werden.

Sie können Untergruppen von Datensätzen auf der Grundlage dessen auswählen, ob diese Datensätze in einem bestimmten Bereich oder in einem von mehreren Bereichen liegen. Des Weiteren können Sie Bereichsinformationen für einen Datensatz aufnehmen, indem Sie einen Ableitungsknoten erstellen, um Datensätze mit einem Flag zu versehen, basierend darauf, ob sie in einem bestimmten Bereich liegen. Für weitere Informationen siehe Thema Generieren von Knoten aus Diagrammen auf S. 370.

Abbildung 5-97
Untersuchen des Bereichs mit hohen Forderungswerten



So definieren Sie Bereiche:

- Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge Ansicht > Interaktionsmodus.
- Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche “Bereich zeichnen”.

Abbildung 5-98
Symbolleistenschaltfläche “Bereich zeichnen”

- Bei Diagrammen, bei denen Bereiche zulässig sind, können Sie den rechteckigen Bereich durch Klicken und Ziehen mit der Maus zeichnen.

Bearbeiten, Umbenennen und Löschen von Bereichen

Die Eigenschaften bestehender Bereiche können im Dialogfeld “Diagrammbereiche bearbeiten” oder über die Kontextmenüs im Diagramm selbst bearbeitet werden.

Abbildung 5-99
Festlegen von Eigenschaften für die definierten Bereiche

So bearbeiten Sie Bereiche:

- Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge Ansicht > Interaktionsmodus.
- Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche “Bereich zeichnen”.
- Wählen Sie in den Menüs die Optionsfolge Bearbeiten > Diagrammbereiche. Das Dialogfeld “Diagrammbereiche bearbeiten” wird geöffnet.
- Wenn das Diagramm mehrere Felder enthält (z. B. bei SPLOM-Diagrammen), müssen Sie das Feld für den Bereich in den Spalten *Feld A* und *Feld B* festlegen.
- Ein neuer Bereich auf einer neuen Linie kann durch Eingabe eines Namens, (ggf.) Auswahl von Feldnamen und Festlegen der Ober- und Untergrenzen für jedes Feld hinzugefügt werden. Drücken Sie die Eingabetaste, um eine neue Zeile zu beginnen.
- Mit den Werten Min und Max für *A* und *B* können Sie bestehende Bereichsgrenzen bearbeiten.
- Zum Umbenennen von Bereichen wählen Sie den Namen des Bereichs in der Tabelle aus.
- Sie können Bereiche löschen, indem Sie die Linie in der Tabelle auswählen und auf die Schaltfläche “Löschen” klicken.
- Klicken Sie auf OK, um die Änderungen zu übernehmen und das Dialogfeld zu schließen.

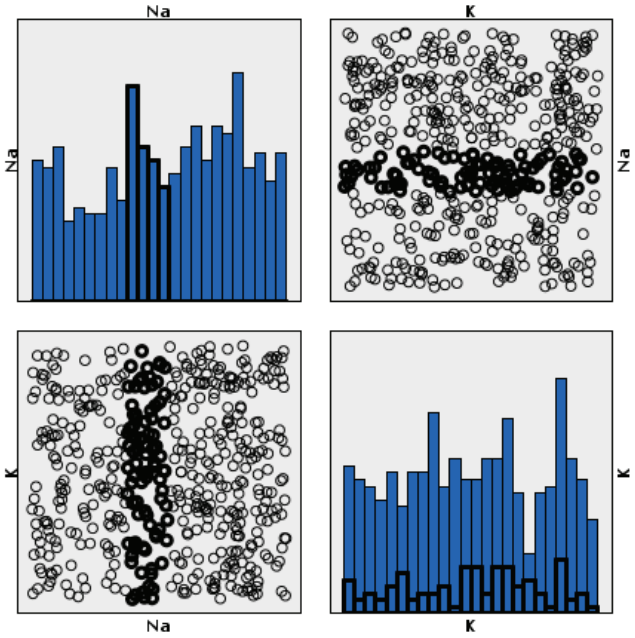
  Hinweis: Alternativ können Sie Bereiche direkt im Diagramm löschen und umbenennen, indem Sie mit der rechten Maustaste auf die Linie des Bereichs klicken und die gewünschte Option aus den Kontextmenüs auswählen.

Verwenden markierter Elemente

In jedem Diagramm können Elemente, wie Balken, Segmente und Punkte, markiert werden. Linien, Flächen und Oberflächen können nur in Zeitdiagrammen, Multidiagrammen und Evaluationsdiagrammen markiert werden, da die Linien sich in diesen Fällen auf Felder beziehen. Bei der Markierung eines Elements heben sie im Grunde alle Daten hervor, für die dieses Element steht. Bei Diagrammen, bei denen derselbe Fall an mehreren Stellen dargestellt wird (wie bei

SPLOM) ist Markieren dasselbe wie Einfärben. Sie können Elemente in Diagrammen und sogar innerhalb von Abschnitten und Bereichen markieren. Wenn Sie ein Element markieren und anschließend wieder in den Bearbeitungsmodus wechseln, bleibt die Markierung weiterhin sichtbar.

Abbildung 5-100
Markieren von Elementen in SPLOMs



Die Markierung von Elementen wird durch Klicken auf das jeweilige Element im Diagramm eingefügt und aufgehoben. Wenn Sie auf ein Element klicken, um es zu markieren, wird das Element mit einem dicken farbigen Rahmen angezeigt. Wenn Sie erneut auf das Element klicken, verschwindet der Rahmen und die Markierung des Elements ist aufgehoben. Um mehrere Elemente zu markieren, können Sie entweder beim Klicken auf die Elemente die Strg-Taste gedrückt halten oder die Maus mit der Zauberstab-Funktion über jedes der zu markierenden Elemente ziehen. Beachten Sie: Wenn Sie auf eine weitere Fläche oder ein weiteres Element klicken, ohne die Strg-Taste gedrückt zu halten, wird bei allen bisher ausgewählten Elementen die Markierung aufgehoben.

Aus den markierten Elementen im Diagramm können Auswahl- und Ableitungsknoten generiert werden. Für weitere Informationen siehe Thema Generieren von Knoten aus Diagrammen auf S. 370.

So markieren Sie Elemente:

- Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge Ansicht > Interaktionsmodus.
- Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche “Elemente markieren”.
- Klicken Sie auf das gewünschte Element oder klicken Sie und ziehen Sie mithilfe der Maus einen Rahmen um mehrere Elemente auf.

Generieren von Knoten aus Diagrammen

Eine der leistungsstärksten Funktionen von IBM® SPSS® Modeler-Diagrammen ist die Möglichkeit, Knoten aus einem Diagramm oder einer Auswahl innerhalb des Diagramms zu generieren. So können Sie beispielsweise aus einem Zeitdiagramm Ableitungs- und Auswahlknoten auf der Grundlage einer Datenauswahl bzw. eines Datenbereichs generieren, wodurch eine Untergruppe der Daten erstellt wird. Diese leistungsstarke Funktion kann beispielsweise zur Ermittlung und zum Ausschluss von Ausreißern verwendet werden.

Immer, wenn ein Abschnitt gezeichnet werden kann, kann auch ein Ableitungsknoten erstellt werden. Bei Diagrammen mit zwei Skalenachsen können Sie Ableitungs- bzw. Auswahlknoten aus den in Ihrem Diagramm gezeichneten Bereichen generieren. Bei Diagrammen mit markierten Elementen können Sie Ableitungsknoten, Auswahlknoten und in einigen Fällen Filterknoten aus diesen Elementen generieren. Die Generierung von Balancierungsknoten ist für alle Diagramme aktiviert, die eine Verteilung von Häufigkeiten (Anzahlwerten) anzeigen.

Abbildung 5-101
Diagramm mit Menü zum Generieren

Beim Generieren von Knoten wird der neue Knoten jeweils direkt im Stream-Zeichenbereich platziert, sodass Sie ihn mit einem bestehenden Stream verbinden können. Aus Diagrammen können die folgenden Knotentypen generiert werden: Auswahl-, Ableitungs-, Balancierungs-, Filter- und Umkodierungsknoten.

Auswahlknoten

Auswahlknoten können generiert werden, um für die Verarbeitung im weiteren Stream-Verlauf einen Test auf Einschluss der Datensätze innerhalb eines Bereichs und Ausschluss aller Datensätze außerhalb des Bereichs (oder umgekehrt) durchzuführen.

- **Bei Abschnitten**. Sie können einen Auswahlknoten generieren, der die Datensätze innerhalb des betreffenden Abschnitts ein- bzw. ausschließt. Auswahlknoten nur für Abschnitte ist nur über Kontextmenüs verfügbar, da Sie auswählen müssen, welcher Abschnitt im Auswahlknoten verwendet werden soll.
- **Bei Bereichen**. Sie können einen Auswahlknoten generieren, der die Datensätze innerhalb des betreffenden Bereichs ein- bzw. ausschließt.
- **Bei markierten Elementen**. Sie können Auswahlknoten generieren, um die Datensätze zu erfassen, die den markierten Elementen bzw. Netzdiagrammzusammenhängen entsprechen.

Ableitungsknoten

Ableitungsknoten können aus Bereichen, Abschnitten und markierten Elementen generiert werden. Alle Diagramme können Ableitungsknoten erstellen. Bei Evaluationsdiagrammen wird ein Dialogfeld zur Auswahl des Modells angezeigt. Bei Netzdiagrammen sind Ableitungsknoten (“UND”) und Ableitungsknoten (“ODER”) möglich.

- **Bei Abschnitten**. Sie können einen Ableitungsknoten generieren, der für jedes auf der Achse markierte Intervall eine Kategorie erstellt. Verwenden Sie hierzu die im Dialogfeld “Abschnitte bearbeiten” aufgeführten Abschnittsnamen als Kategorienamen.
- **Bei Bereichen**. Sie können einen Ableitungsknoten (Ableitungstyp: Flag) erstellen, der ein Flag-Feld mit der Bezeichnung *Bereich (Flag)* erstellt, bei dem die Flags auf *T* (Datensätze in einem Bereich) bzw. *F* (Datensätze außerhalb aller Bereiche) gesetzt sind. Außerdem können Sie einen Ableitungsknoten (Ableitungstyp: Set) generieren, der ein Set mit einem Wert für jede Region mit einem neuen Feld *Bereich* für jeden Datensatz erstellt. Dabei wird der Name des Bereichs, in den die Datensätze fallen, als Wert verwendet. Datensätze, die außerhalb aller Bereiche liegen, erhalten den Namen des Standardbereichs. Wertenamen werden im Bearbeitungsdialogfeld für Bereiche als Bereichsnamen aufgelistet.
- **Bei markierten Elementen**. Sie können einen Ableitungsknoten generieren, der ein Flag berechnet, das für alle markierten Elemente *Wahr* und für alle anderen Datensätze *Falsch* ist.

Balancierungsknoten

Balancierungsknoten können generiert werden, um Ungleichgewichte in den Daten zu korrigieren, beispielsweise durch Verringerung des Auftretens häufiger Werte (Menüoption Balancierungsknoten (verringern)) oder durch Erhöhen des Auftretens seltener Werte (Menüoption Balancierungsknoten (erhöhen)). Die Generierung von Balancierungsknoten wird für alle Diagramme aktiviert, die eine Häufigkeitsverteilung anzeigen, wie Histogramm, Punkt, Sammlung, Balken für Häufigkeiten, Kreis für Häufigkeiten und Multidiagramm.

Filterknoten

Filterknoten können generiert werden, um Felder anhand der im Diagramm markierten Linien bzw. Knoten umzubenennen oder zu filtern. Bei Evaluationsdiagrammen generiert die Linie für die beste Anpassung keinen Filterknoten.

Umkodierungsknoten

Umkodierungsknoten können zum Umkodieren von Werten generiert werden. Diese Option wird für Verteilungsdiagramme verwendet. Sie können einen Umkodierungsknoten für **Gruppen** generieren, um bestimmte Werte eines angezeigten Felds in Abhängigkeit davon neu zu kodieren, ob sie in einer Gruppe enthalten sind (Gruppen können mithilfe von Strg+Klicken auf der Registerkarte "Tabellen" ausgewählt werden). Außerdem können Sie einen Umkodierungsknoten für **Werte** generieren, um Daten in ein bestehendes Set mit mehreren Werten neu zu kodieren. Ein Beispiel hierfür ist die Umkodierung von Daten in ein Standardset von Werten, um Finanzdaten von mehreren Unternehmen zu Analysezwecken zusammenzuführen.

Hinweis: Sind die Werte bereits vordefiniert, können Sie sie als Textdatei (Einfachdatei) in SPSS Modeler einlesen und dann mithilfe einer Verteilung alle Werte anzeigen lassen. Anschließend können Sie einen Umkodierungsknoten (Werte) direkt aus dem Diagramm heraus erzeugen. Dadurch werden alle Zielwerte in die Spalte *Neue Werte* (Dropdown-Liste) im Umkodierungsknoten eingefügt.

Generieren von Knoten aus Diagrammen

Mithilfe des Menüs "Generieren" im Diagrammausgabefenster können Knoten generiert werden. Der generierte Knoten wird in den Stream-Zeichenbereich platziert. Um den Knoten nutzen zu können, verbinden Sie ihn mit einem vorhandenen Stream.

So erzeugen Sie einen Knoten aus einem Diagramm:

- Vergewissern Sie sich, dass Sie sich im Interaktionsmodus befinden. Wählen Sie in den Menüs die Optionsfolge Ansicht > Interaktionsmodus.
- Klicken Sie in der Symbolleiste des Interaktionsmodus auf die Schaltfläche "Bereich".
- Legen Sie die Abschnitte, Bereiche und markierten Elemente fest, die Sie zur Generierung des gewünschten Knotens benötigen.
- Wählen Sie im Menü "Generieren" den gewünschten Knotentyp aus. Nur die zulässigen Knoten sind aktiviert.

  Hinweis: Alternativ können Sie Knoten direkt im Diagramm generieren, indem Sie mit der rechten Maustaste klicken und die gewünschte Generierungsoption aus den Kontextmenüs auswählen.

Bearbeiten von Visualisierungen

Während Sie im Sondierungsmodus die durch die Visualisierung dargestellten Daten und Werte erforschen, gestattet Ihnen der Bearbeitungsmodus, das Layout und Aussehen der Visualisierung zu ändern. Sie können beispielsweise die Schriftarten und Farben so anpassen, dass sie den in Ihrem Unternehmen geltenden Stilrichtlinien entsprechen. Um diesen Modus auszuwählen,

wählen Sie in den Menüs die Optionsfolge Ansicht > Bearbeitungsmodus (oder klicken Sie auf das entsprechende Symbol in der Symbolleiste).

Im Bearbeitungsmodus stehen mehrere Symbolleisten zur Verfügung, mit denen sich die verschiedenen Aspekte des Visualisierungslayouts beeinflussen lassen. Wenn Sie feststellen, dass Sie nicht alle davon verwenden, können Sie die nicht benötigten Symbolleisten ausblenden, um den Platz in dem Dialogfeld zu vergrößern, in dem das Diagramm angezeigt wird. Um die Symbolleisten auszuwählen bzw. ihre Auswahl aufzuheben, klicken Sie im Menü "Ansicht" auf den Namen der entsprechenden Symbolleiste.

Hinweis: Um Ihre Visualisierungen mit weiteren Details zu versehen, können Sie Titel, Fußnoten und Achsenbeschriftungen zuweisen. Für weitere Informationen siehe Thema Hinzufügen von Titeln und Fußnoten auf S. 390.

Zur Bearbeitung einer Visualisierung im **Bearbeitungsmodus** sind mehrere Optionen verfügbar. Sie verfügen über folgende Möglichkeiten:

- Text bearbeiten und formatieren.
- Ändern Sie die Füllfarbe, die Transparenz und das Muster von Rahmen und Grafikelementen.
- Ändern der Farbe und Striche für Ränder und Linien.
- Drehen und Verändern der Form und des Seitenverhältnisses von Punktelementen.
- Änderung der Größe grafischer Elemente (wie Balken und Punkte).
- Anpassen des Raums um Elemente durch Verwendung von Rändern und Textabstand.
- Geben Sie Formate für Zahlen an.
- Änderung der Achsen- und Skaleneinstellungen.
- Sortieren, Ausschließen und Verkleinern von Kategorien auf einer Kategorienachse.
- Legen Sie die Orientierung von Feldern.
- Weisen Sie einem Koordinatensystem Transformationen zu.
- Ändern Sie Statistiken, Grafikelementtypen und Kollisionsmodifikatoren.
- Änderung der Position der Legende.
- Weisen Sie Visualisierungs-Stilvorlagen hinzu.

Die folgenden Punkte beschreiben, wie diese verschiedenen Aufgaben durchgeführt werden. Es empfiehlt sich, darüber hinaus auch die allgemeinen Regeln für die Bearbeitung von Diagrammen zu lesen.

So wechseln Sie in den Bearbeitungsmodus:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Ansicht > Bearbeitungsmodus

Allgemeine Regeln zur Bearbeitung von Visualisierungen

Bearbeitungsmodus

Alle Bearbeitungen finden im Bearbeitungsmodus statt. Wählen Sie zum Aktivieren des Bearbeitungsmodus die folgenden Befehle aus den Menüs aus:
Ansicht > Bearbeitungsmodus

Auswahl

Die für die Bearbeitung verfügbaren Optionen hängen von der Auswahl ab. Je nach Auswahl werden andere Optionen der Symbolleiste und der Eigenschaftenpalette aktiviert. Auf die aktuelle Auswahl können nur die aktivierten Elemente angewendet werden. Wenn z. B. eine Achse ausgewählt ist, sind in der Eigenschaftenpalette die Registerkarten “Skala”, “Hauptteilstriche” und “Hilfsteilstriche” verfügbar.

Hier finden Sie einige Tipps zur Auswahl von Objekten in der Visualisierung:

- Klicken Sie auf ein Element, um es auszuwählen.
- Wählen Sie ein grafisches Element (wie Punkte in einem Streudiagramm oder Balken in einem Balkendiagramm) mit einem einzelnen Klick aus. Klicken Sie nach der anfänglichen Auswahl erneut, um die Auswahl auf Gruppen von Grafikelementen oder ein einziges Grafikelement einzuschränken.
- Drücken Sie auf ESC, um die gesamte Auswahl aufzuheben.

Paletten

Wenn ein Objekt in der Visualisierung ausgewählt ist, werden die verschiedenen Paletten passend zur Auswahl aktualisiert. Die Paletten enthalten Steuerungen für die Bearbeitung der Auswahl. Paletten können Symbolleisten oder ein Fenster mit mehreren Steuerungen und Registerkarten sein. Paletten können ausgeblendet sein, stellen Sie also sicher, dass die erforderliche Palette für die jeweilige Bearbeitung angezeigt ist. Prüfen Sie im Menü “Ansicht”, welche Paletten derzeit angezeigt werden.

Sie können die Paletten umpositionieren, indem Sie in den leeren Bereich in der Symbolleistenpalette oder an der linken Seite anderer Paletten klicken und ihn an eine andere Stelle ziehen. Die Stellen, an denen Sie die Palette andocken können, werden visuell gekennzeichnet. Für Paletten, die keine Systemleisten sind, können Sie auch auf das Schließfeld klicken, um die Palette auszublenden, und auf die Schaltfläche zum Loslösen, um die Palette in einem separaten Fenster zu zeigen. Klicken Sie auf die Schaltfläche “Hilfe”, um Hilfe zur jeweiligen Palette zu erhalten.

Automatische Einstellungen

Für einige Einstellungen ist die Option -automatisch- verfügbar. Diese gibt an, dass automatisch Werte angewendet werden. Welche automatischen Einstellungen benutzt werden, hängt von der jeweiligen Visualisierung und den Datenwerten ab. Um die automatische Einstellung zu überschreiben, können Sie einen Wert eingeben. Wenn Sie die automatische Einstellung

wiederherstellen möchten, löschen Sie den aktuellen Wert und drücken Sie die EINGABETASTE. Für die Einstellungen wird erneut -automatisch- angezeigt.

Entfernen/Ausblenden von Elementen

Sie können zahlreiche Elemente in der Visualisierung ein- oder ausblenden. Sie können beispielsweise die Legende oder die Achsenbeschriftung ausblenden. Um ein Element zu löschen, wählen Sie es aus und drücken Sie auf ENTF. Wenn das Element nicht löschbar ist, geschieht nichts. Wenn Sie ein Element versehentlich gelöscht haben, drücken Sie die Tasten STRG+Z, um das Löschen rückgängig zu machen.

State

Einige Symbolleisten zeigen den Status der aktuellen Auswahl an, andere nicht. Die Eigenschaftenpalette zeigt immer den Status an. Wenn eine Symbolleiste den Status *nicht* berücksichtigt, wird dies unter dem Punkt, der die Symbolleiste beschreibt, erwähnt.

Bearbeiten und Formatieren von Text

Sie können Texte an Ort und Stelle bearbeiten und die Formatierung eines gesamten Textblocks ändern. Beachten Sie, dass Sie keine Texte bearbeiten können, die direkt mit Datenwerten verknüpft sind. Sie können z. B. keine Teilstrichbeschriftungen bearbeiten, da der Inhalt der Beschriftung aus den zugrunde liegenden Daten abgeleitet wird. Sie können jedoch jeden Text in der Visualisierung formatieren.

So bearbeiten Sie Text im Diagramm

- Doppelklicken Sie auf den Textblock. Durch diese Aktion wird der gesamte Text ausgewählt. Alle Symbolleisten werden deaktiviert, da Sie während der Bearbeitung von Text keinen anderen Bereich der Visualisierung ändern können.
- Geben Sie den Text ein, der den vorhandenen ersetzen soll. Sie können auch erneut auf den Text klicken, damit ein Cursor angezeigt wird. Positionieren Sie den Cursor an der gewünschten Stelle und geben Sie weiteren Text ein.

So formatieren Sie Text

- Wählen Sie den Rahmen aus, der den Text enthält. Doppelklicken Sie nicht auf den Text.
- Formatieren Sie den Text mithilfe der Symbolleiste für Schriftarten. Wenn die Symbolleiste nicht aktiviert ist, stellen Sie sicher, dass der den Text enthaltende *Rahmen* ausgewählt ist. Wenn der Text als solcher ausgewählt ist, wird die Symbolleiste deaktiviert.

Abbildung 5-102
Symbolleiste "Schriftart"

Sie können die Schriftart ändern:

- Farbe
- Schriftartfamilie (z. B. Arial oder Verdana)
- Schriftgrad (die Einheit ist Punkt (pt), sofern Sie keine andere Einheit angeben, wie beispielsweise Pica (pc))
- Stärke
- Relative Ausrichtung zum Textrahmen

Die Formatierung wird auf den gesamten Text im Rahmen angewendet. Die Formatierung einzelner Buchstaben oder Wörter kann in einem Textblock nicht geändert werden.

Ändern von Farben, Mustern, Strichmustern und Transparenz

Viele verschiedene Elemente in einer Visualisierung verfügen über eine Füllung und einen Rahmen. Das naheliegendste Beispiel ist ein Balken eines Balkendiagramms. Die Farbe des Balkens ist die Füllfarbe. Außen herum kann sich ein durchgängiger schwarzer Rand befinden.

In der Visualisierung gibt es weniger deutliche Beispiele mit Füllfarben. Wenn die Füllfarbe transparent ist, erkennen Sie diese womöglich nicht. Nehmen Sie z. B. den Text einer Achsenbeschriftung. Dieser Text erscheint als frei schwebender Text. Tatsächlich steht er allerdings in einem Rahmen, der eine transparente Füllfarbe besitzt. Den Rahmen können Sie sehen, wenn Sie die Achsenbeschriftung auswählen.

Jeder Rahmen in einer Visualisierung kann über ein Füllmuster und eine Rahmenart verfügen, auch der Rahmen um die gesamte Visualisierung. Zudem ist mit jedem Füllmuster ein Grad der Lichtdurchlässigkeit/Transparenz verbunden, der sich anpassen lässt.

So ändern Sie Farben, Muster, Strichmuster und Transparenz:

▶ Wählen Sie das Element aus, das Sie formatieren möchten. Wählen Sie beispielsweise die Balken eines Balkendiagramms oder einen Rahmen, der Text enthält, aus. Wenn die Visualisierung durch eine kategoriale Variable oder ein Feld getrennt ist, können Sie auch die Gruppe auswählen, die einer individuellen Kategorie entspricht. So können Sie das der Gruppe zugewiesene standardmäßige Aussehen ändern. Sie können z. B. in einem Stapeldiagramm die Farbe einer der Stapelgruppen ändern.

▶ Die Füllfarbe, die Rahmenfarbe oder das Füllmuster ändern Sie über die Symbolleiste für Farben.

Abbildung 5-103
Symbolleiste "Farbe"

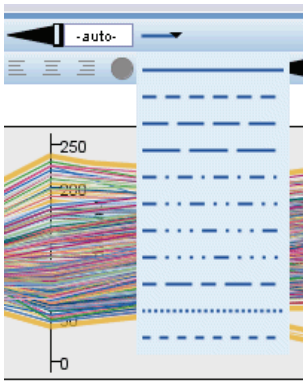
Hinweis: Diese Symbolleiste spiegelt den Status der aktuellen Auswahl nicht.

Sie können eine Farbe oder ein Füllmuster ändern, indem Sie auf die Schaltfläche klicken, um die angezeigte Option auszuwählen, oder auf den Dropdown-Pfeil, um eine andere Option zu wählen. Beachten Sie bei den Farben, dass eine davon weiß erscheint und durch eine rote diagonale Linie gekennzeichnet ist. Dies ist die Kennzeichnung für die transparente Farbe. Diese können Sie z. B. verwenden, um die Rahmen um Balken in einem Histogramm zu verbergen.

- Die erste Schaltfläche bestimmt die Füllfarbe.
- Die zweite Schaltfläche bestimmt die Rahmenfarbe.
- Die dritte Schaltfläche bestimmt das Füllmuster. Das Füllmuster verwendet die Rahmenfarbe. Daher ist das Füllmuster nur sichtbar, wenn eine sichtbare Rahmenfarbe ausgewählt ist.
- Die vierte Steuerung besteht aus einem Schieberegler und einem Textfeld und dient zur Steuerung der Lichtundurchlässigkeit der Füllfarbe und des Füllmusters. Ein geringerer Prozentsatz bedeutet weniger Undurchlässigkeit und mehr Transparenz. 100 % ist völlig undurchlässig (keine Transparenz).

▶ Die Strichelung eines Rahmens oder einer Linie ändern Sie mit der Symbolleiste für Linien.

Abbildung 5-104
Symbolleiste "Linie"



Hinweis: Diese Symbolleiste spiegelt den Status der aktuellen Auswahl nicht.

Sie können wie bei der anderen Symbolleiste auf die Schaltfläche klicken, um die angezeigte Option auszuwählen, oder Sie klicken auf den Pfeil der Dropdown-Liste, um eine andere Option auszuwählen.

Drehen und Verändern der Form und des Seitenverhältnisses von Punktelementen

Sie können Punktelemente drehen, ihnen eine andere vordefinierte Form zuweisen oder das Seitenverhältnis ändern (das Verhältnis von Breite zu Höhe).

So ändern Sie Punktelemente

▶ Wählen Sie die Punktelemente aus. Die Form und das Seitenverhältnis eines einzelnen Punktelements können nicht gedreht oder geändert werden.

▶ Ändern Sie die Punkte über die Symbolleiste für Symbole.

Abbildung 5-105
Symbolleiste "Symbol"

- Mit der ersten Schaltfläche können Sie die Form der Punkte ändern. Klicken Sie auf den Dropdown-Pfeil und wählen Sie eine vordefinierte Form aus.

- Mit der zweiten Schaltfläche können Sie die Punkte in eine bestimmte Position drehen. Klicken Sie auf den Pfeil nach unten und ziehen Sie dann die Nadel in die gewünschte Position.
- Mit der dritten Schaltfläche können Sie das Seitenverhältnis ändern. Klicken Sie auf den Dropdown-Pfeil, klicken Sie dann auf das Rechteck und ziehen Sie es. Die Form des Rechtecks stellt das Seitenverhältnis dar.

Die Größe grafischer Elemente ändern

Sie können die Größe der Grafikelemente in der Visualisierung ändern. Dazu gehören unter anderem Balken, Linien und Punkte. Wenn die Größe des grafischen Elements durch eine Variable oder ein Feld bestimmt wird, ist die festgelegte Größe die *minimale* Größe.

So ändern Sie die Größe grafischer Elemente

▶ Wählen Sie die grafischen Elemente aus, deren Größe Sie ändern möchten.

▶ Verwenden Sie den Schieberegler oder geben Sie für die in der Symbolleiste "Symbol" verfügbare Option eine bestimmte Größe ein. Sofern Sie keine andere Einheit angeben (unten finden Sie eine vollständige Liste mit Abkürzungen für Einheiten), handelt es sich um Pixel. Sie können auch einen Prozentwert angeben (wie 30 %). Dadurch verwendet das grafische Element den angegebenen Prozentsatz des verfügbaren Raums. Der verfügbare Platz hängt vom Typ des Grafikelements und der spezifischen Visualisierung ab.

Tabelle 5-3
Gültige Abkürzungen für Einheiten

Abkürzung	Einheit
cm	Zentimeter
in	Zoll
mm	Millimeter
pc	Pica
pt	Punkt
px	Pixel

Abbildung 5-106
Größenangabe auf der Symbolleiste "Symbol"

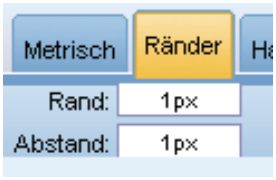
Festlegen von Rändern und Textabstand

Wenn um oder im Rahmen in der Visualisierung zu viel oder zu wenig Platz ist, können Sie seine Rand- und Abstandeinstellungen ändern. Beim **Rand** handelt es sich um den Abstand zwischen dem Rahmen und anderen darum befindlichen Elementen. Der **Textabstand** ist der Abstand zwischen dem Rand des Rahmens und dem *Inhalt* des Rahmens.

So legen Sie Ränder und Textabstand fest

- Wählen Sie den Rahmen aus, für den Sie Ränder und Textabstand festlegen möchten. Hierbei kann es sich um einen Textrahmen, einen Rahmen um eine Legende oder auch um einen Datenrahmen handeln, in dem die grafischen Elemente (wie Balken und Punkte) angezeigt werden.
- Verwenden Sie die Registerkarte “Ränder” der Eigenschaftenpalette, um die Einstellungen vorzunehmen. Bei allen Größenangaben handelt es sich um Pixel, sofern Sie keine andere Einheit angeben (z. B. “cm” oder “in”).

Abbildung 5-107
Registerkarte “Ränder”



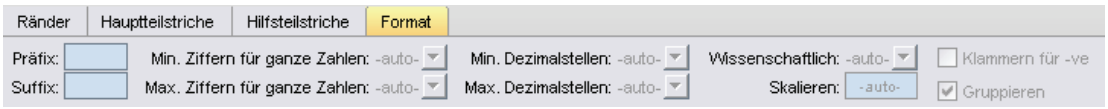
Formatieren von Zahlen

Sie können das Format für Zahlen in Teilstrichbeschriftungen auf einer fortlaufenden Achse sowie bei Datenwertelabels mit Zahlen angeben. Beispielsweise können Sie festlegen, dass an den Hauptteilstrichen angegebene Zahlen in Tausendern gezeigt werden.

So geben Sie Zahlenformate an:

- Wählen Sie die Teilstrichbeschriftungen der fortlaufenden Achse oder die Datenwertelabels aus, wenn sie Zahlen enthalten.
- Klicken Sie im Fenster “Eigenschaften” auf die Registerkarte Format.

Abbildung 5-108
Registerkarte “Format”



- Wählen Sie die gewünschten Zahlenformatoptionen aus.

Präfix. Ein Zeichen, das vor der Zahl angezeigt werden soll. Geben Sie beispielsweise ein Dollarzeichen ($) ein, wenn es sich bei den Zahlen um Gehälter in US-Dollar handelt.

Suffix. Ein Zeichen, das nach der Zahl angezeigt werden soll. Geben Sie beispielsweise ein Prozentzeichen (%) ein, wenn es sich bei den Zahlen um Prozentsätze handelt.

Min. Ganzzahlstellen. Mindestanzahl an Stellen, die im ganzzahligen Teil der Dezimaldarstellung angezeigt werden sollen. Wenn der tatsächliche Wert nicht über die Mindestanzahl an Stellen verfügt, wird der ganzzahlige Teil des Wertes mit Nullen aufgefüllt.

Max. Ganzzahlstellen. Höchstanzahl an Stellen, die im ganzzahligen Teil der Dezimaldarstellung angezeigt werden sollen. Wenn der tatsächliche Wert die Höchstanzahl an Stellen überschreitet, wird der ganzzahlige Teil des Wertes durch Sternchen ersetzt.

Min. Dezimalstellen. Mindestanzahl an Stellen, die im Dezimalteil der Dezimal- oder wissenschaftlichen Darstellung angezeigt werden sollen. Wenn der tatsächliche Wert nicht über die Mindestanzahl an Stellen verfügt, wird der Dezimalteil des Wertes mit Nullen aufgefüllt.

Max. Dezimalstellen. Höchstanzahl an Stellen, die im Dezimalteil der Dezimal- oder wissenschaftlichen Darstellung angezeigt werden sollen. Wenn der tatsächliche Wert die Höchstanzahl an Stellen überschreitet, wird der Dezimalwert auf die passende Anzahl an Stellen gerundet.

Wissenschaftlich. Ob Zahlen in wissenschaftlicher Notation angezeigt werden sollen. Die wissenschaftliche Notation ist für sehr große oder sehr kleine Zahlen sinnvoll. -auto- überlässt es der Anwendung zu entscheiden, ob wissenschaftliche Notation angemessen ist.

Skalierung. Ein Skalierungsfaktor, der eine Zahl ist, durch die der Originalwert dividiert wird. Verwenden Sie einen Skalierungsfaktor, wenn die Zahlen groß sind, Sie jedoch nicht möchten, dass die Teilstrichbeschriftung durch die Anzeige der großen Zahl übermäßig lang wird. Wenn Sie das Zahlenformat der Teilstrichbeschriftungen ändern, bearbeiten Sie auch den Achsentitel, um anzugeben, wie die Zahl zu interpretieren ist. Nehmen wir an, auf der Skalenachse werden Gehälter angezeigt und die Labels lauten 30.000, 50.000 und 70.000. Hier können Sie einen Skalierungsfaktor von 1.000 eingeben, um 30, 50 und 70 anzuzeigen. In diesem Fall wäre es sinnvoll, den Skalenachsentitel zu bearbeiten, um den Text in Tausend einzuschließen.

Klammern für -ve. Ob negative Werte eingeklammert werden sollen.

Gruppierung. Ob ein bestimmtes Zeichen zwischen Zifferngruppen angezeigt werden soll. Das aktuelle Gebietsschema Ihres Computers bestimmt, welches Zeichen zur Zifferngruppierung verwendet wird.

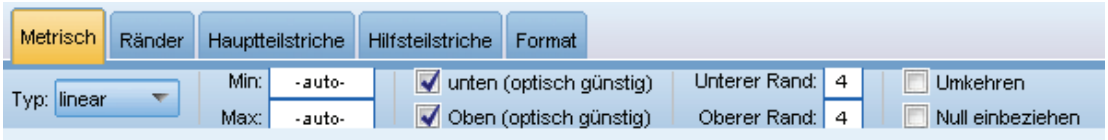
Änderung der Achsen- und Skaleneinstellungen.

Für die Änderung der Achsen und Skalen sind verschiedene Optionen verfügbar.

So ändern Sie die Achsen- und Skaleneinstellungen

- Wählen Sie einen beliebigen Teil der Achse aus (z. B. die Achsen- oder die Teilstrichbeschriftungen).
- Verwenden Sie die Registerkarten “Skala”, “Hauptteilstriche” und “Hilfsteilstriche” der Eigenschaftenpalette, um die Achsen- und Skaleneinstellungen zu ändern.

Abbildung 5-109
Eigenschaftenpalette



Registerkarte “Skala”

Anmerkung: Die Registerkarte “Skala” wird bei Diagrammen mit voraggregierten Daten (z. B. Histogrammen) nicht angezeigt.

Typ. Legt fest, ob die Skala linear oder transformiert ist. Skalentransformationen erleichtern das Verständnis der Daten und der für statistische Rückschlüsse notwendigen Annahmen. In Streudiagrammen haben Sie möglicherweise eine transformierte Skala verwendet, wenn die Beziehung zwischen den unabhängigen und den abhängigen Variablen oder Feldern nicht linear ist. Skalentransformationen können auch verwendet werden, um ein verzerrtes Histogramm symmetrischer zu machen und es einer Normalverteilung anzugleichen. Beachten Sie, dass Sie nur die Skala transformieren, auf der die Daten angezeigt werden. Die tatsächlichen Daten werden nicht transformiert.

- **Linear.** Legt eine lineare, nicht transformierte Skala fest.
- **Log.** Legt eine mit Logarithmus zur Basis 10 transformierte Skala fest. Um Nullwerte und negative Werte zu berücksichtigen, verwendet diese Transformation eine angepasste Version der Log-Funktion. Diese "safeLog"-Funktion wird als sign(x) * log(1 + abs(x)) definiert, weshalb safeLog(-99) Folgendem entspricht:

  sign(-99) * log(1 + abs(-99)) = -1 * log(1 + 99) = -1 * 2 = -2
- **Exponent.** Legt eine Skala mit Potenz-Transformation fest, die einen Exponenten von 0,5 verwendet. Um negative Werte zu berücksichtigen, verwendet diese Transformation eine angepasste Version der Potenzfunktion. Diese "safePower"-Funktion wird als sign(x) * pow(abs(x), 0.5) definiert, weshalb safePower(-100) Folgendem entspricht:

  sign(-100) * pow(abs(-100), 0.5) = -1* pow(100, 0.5) = -1 * 10 = -10

Min/Max/Optisch günstig unten/oben. Legt den Bereich der Skala fest. Wenn Sie Unten (optisch günstig) und Oben (optisch günstig) auswählen, kann die Anwendung auf der Grundlage der Daten eine geeignete Skala auswählen. Das Minimum und das Maximum ist "optisch günstig", da es sich in der Regel um ganze Werte größer oder kleiner als die maximalen oder minimalen Datenwerte handelt. Wenn die Daten z. B. von 4 bis 92 reichen, kann eine unten oder oben optisch günstige Skala anstatt des tatsächlichen Minimum- und Maximumwerts der Daten 0 und 100 sein. Achten Sie darauf, dass Sie den Bereich nicht zu klein festlegen, wodurch wichtige Elemente eventuell verborgen werden. Beachten Sie außerdem, dass Sie kein explizites Minimum oder Maximum festlegen können, wenn die Option Null einbeziehen ausgewählt ist.

Unterer Rand/Oberer Rand. Erstellt Ränder am unteren und/oder oberen Ende der Achse. Der Rand erscheint senkrecht zur ausgewählten Achse. Sofern Sie keine andere Einheit angeben (z. B. "cm" oder "in"), handelt es sich um Pixel. Wenn Sie als für die vertikale Achse unter Oberer Rand beispielsweise 5 festlegen, verläuft oben im Datenrahmen ein horizontaler Rand von 5 Pixel.

Umkehren. Legt fest, ob die Skala umgekehrt ist.

Null einbeziehen. Gibt an, dass die Skala 0 einschließen soll.Diese Option wird in der Regel für Balkendiagramme verwendet, um sicherzustellen, dass die Balken bei 0 beginnen anstatt bei einem Wert im Bereich der Höhe des kleinsten Balkens. Wenn diese Option ausgewählt ist, sind Min und Max deaktiviert, da für den Skalenbereich kein benutzerdefinierter Minimal- und Maximalwert festgelegt werden kann.

Registerkarte "Hauptteilstriche"/"Hilfsteilstriche"

Teilstriche oder **Achsenteilstriche** sind die Linien, die auf einer Achse erscheinen. Diese zeigen Werte bei bestimmten Intervallen oder Kategorien an. **Hauptteilstriche** sind Achsenteilstriche mit Beschriftungen. Diese sind außerdem länger als andere Teilstriche. **Hilfsteilstriche** sind

Achsenteilstriche, die zwischen den Hauptteilstrichen erscheinen. Einige Optionen sind für den Teilstrichtyp spezifisch. Die meisten Optionen sind aber sowohl für Hauptteilstriche als auch für Hilfsteilstriche verfügbar.

Teilstriche anzeigen. Gibt an, ob in einem Diagramm Haupt- oder Hilfsteilstriche angezeigt werden sollen.

Gitterlinien anzeigen. Legt fest, ob mit den Haupt- oder Hilfsteilstrichen Gitterlinien angezeigt werden. **Gitterlinien** sind Linien, die auf einem ganzen Diagramm parallel zu den Achsen verlaufen.

Position. Legt die Position der Teilstriche in Bezug auf die Achse fest.

Länge. Legt die Länge der Teilstriche fest. Sofern Sie keine andere Einheit angeben (z. B. “cm” oder “in”), handelt es sich um Pixel.

Basis. *Gilt nur für Hauptteilstriche.* Legt den Wert fest, bei dem der erste Hauptteilstrich erscheint.

Delta. *Gilt nur für Hauptteilstriche.* Legt die Differenz zwischen Hauptteilstrichen fest. Dadurch erscheinen Hauptteilstriche bei jedem *n*-ten Wert, wobei *n* der Delta-Wert ist.

Unterteilungen. *Gilt nur für Hilfsteilstriche.* Legt die Anzahl der Hilfsteilstriche zwischen Hauptteilstrichen fest. Die Anzahl der Hilfsteilstriche ist um eins kleiner als die Anzahl der Unterteilungen. Angenommen, bei 0 und 100 befinden sich Hauptteilstriche. Wenn Sie für die Anzahl der Hilfsteilstrichunterteilungen 2 eingeben, wird *ein* Hilfsteilstrich bei 50 erscheinen, der den Bereich von 0 bis 100 in *zwei* Bereiche unterteilt.

Bearbeiten von Kategorien

Die Kategorien auf einer Kategorienachse können auf verschiedene Weise bearbeitet werden:

- Ändern der Sortierreihenfolge für die Anzeige der Kategorien.
- Ausschließen bestimmter Kategorien.
- Sie können Kategorien hinzufügen, die nicht im Daten-Set angezeigt werden.
- Kleinere Kategorien lassen sich in einer Kategorie zusammenfassen oder kombinieren.

So ändern Sie die Sortierreihenfolge von Kategorien:

▶ Wählen Sie eine Kategorienachse aus. Die Kategorienpalette zeigt die Kategorien auf der Achse an.

Hinweis: Sollte die Palette nicht sichtbar sein, überprüfen Sie, ob Sie sie aktiviert haben. Wählen Sie in IBM® SPSS® ModelerKategorien aus dem Menü “Ansicht”.

▶ Wählen Sie in der Kategorienpalette eine Sortierreihenfolge in der Dropdown-Liste aus.

Benutzerdefiniert. Sortiert die Kategorien anhand der Reihenfolge, in der sie in der Palette angezeigt werden. Mit den Pfeilschaltflächen können Sie die Kategorien an die Spitze oder an das Ende der Liste bzw. innerhalb der Liste nach oben und unten verschieben.

Daten. Sortiert die Kategorien anhand der Reihenfolge, in der sie im Daten-Set vorkommen.

Name. Sortiert die Kategorien alphabetisch anhand der in der Palette angezeigten Namen. Hierbei kann es sich entweder um den Wert oder um die Beschriftung handeln, je nachdem ob Symbolleistenschaltfläche zur Anzeige von Werten und Beschriftungen ausgewählt wurde.

Wert. Sortierung der Kategorien nach dem zugrunde liegenden Datenwert unter Verwendung der Werte, die im Fenster eingeklammert angezeigt werden. Nur Datenquellen mit Metadaten (z. B. IBM® SPSS® Statistics Datendateien) unterstützen diese Option.

Statistik. Sortiert die Kategorien anhand der berechneten Statistik für die jeweilige Kategorie. Beispiele für Statistiken sind Anzahl, Prozentsatz und Mittelwert. Diese Option ist nur verfügbar, wenn im Diagramm eine Statistik verwendet wird.

So fügen Sie eine Kategorie hinzu:

Standardmäßig sind nur Kategorien verfügbar, die im Daten-Set angezeigt werden. Sie können der Visualisierung bei Bedarf eine Kategorie hinzufügen.

- Wählen Sie eine Kategorienachse aus. Die Kategorienpalette zeigt die Kategorien auf der Achse an.

  Hinweis: Sollte die Palette nicht sichtbar sein, überprüfen Sie, ob Sie sie aktiviert haben. Wählen Sie in SPSS ModelerKategorien aus dem Menü “Ansicht”.

- Klicken Sie im Fenster “Kategorien” auf die Schaltfläche “Kategorie hinzufügen”:

  Abbildung 5-110
  Schaltfläche “Kategorie hinzufügen”

- Geben Sie im Dialogfeld “Neue Kategorie hinzufügen” einen Namen für die Kategorie ein.
- Klicken Sie auf OK.

So können Sie bestimmte Kategorien ausschließen:

- Wählen Sie eine Kategorienachse aus. Die Kategorienpalette zeigt die Kategorien auf der Achse an.

  Hinweis: Sollte die Palette nicht sichtbar sein, überprüfen Sie, ob Sie sie aktiviert haben. Wählen Sie in SPSS ModelerKategorien aus dem Menü “Ansicht”.

- Wählen Sie in der Kategorienpalette einen Kategoriennamen in der Liste “Einschließen” aus und klicken Sie dann auf die Schaltfläche “X”. Um die Kategorie zurückzuverschieben, wählen Sie ihren Namen in der Liste der ausgeschlossenen Elemente aus und klicken Sie auf den Pfeil rechts neben der Liste.

So kombinieren Sie kleinere Kategorien bzw. fassen sie zusammen:

Sie können Kategorien zusammenfassen, die so klein sind, dass sie nicht gesondert angezeigt zu werden brauchen. Bei einem Kreisdiagramm mit vielen Kategorien könnten beispielsweise alle Kategorien mit einem Prozentsatz von weniger als 10 zusammengefasst werden. Das

Zusammenfassen ist nur für additive Statistiken verfügbar. Sie können beispielsweise keine Mittelwerte addieren, da Mittelwerte nicht additiv sind. Daher ist das Kombinieren bzw. Zusammenfassen von Kategorien mithilfe eines Mittelwerts nicht möglich.

- Wählen Sie eine Kategorienachse aus. Die Kategorienpalette zeigt die Kategorien auf der Achse an.

  Hinweis: Sollte die Palette nicht sichtbar sein, überprüfen Sie, ob Sie sie aktiviert haben. Wählen Sie in SPSS ModelerKategorien aus dem Menü "Ansicht".

- Wählen Sie in der Kategorienpalette die Option Reduzieren und geben Sie einen Prozentsatz an. Alle Kategorien, deren Gesamtprozentsatz unter dem angegebenen Wert liegt, werden zu einer Kategorie zusammengefasst. Der Prozentsatz beruht auf der im Diagramm gezeigten Statistik. Die Reduzierung steht nur für Anzahl- und Summierungsstatistiken zur Verfügung.

Ändern der Orientierung von Feldern

Wenn Sie in Ihrer Visualisierung Felder verwenden, können Sie deren Orientierung ändern.

So ändern Sie die Ausrichtung der Fenster

- Klicken Sie auf einen beliebigen Bereich der Visualisierung.
- Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte Felder.

Abbildung 5-111
Registerkarte "Fenster"

- Wählen Sie eine Option für Layout:

Tabelle. Ordnet Fenster wie eine Tabelle an, in der jeder Zeile oder Spalte ein einzelner Wert zugeordnet ist.

Transponiert. Ordnet Fenster wie eine Tabelle an und vertauscht außerdem die Daten der ursprünglichen Zeilen und Spalten. Diese Option entspricht nicht dem Transponieren des Diagramms. Beachten Sie, dass die *x*-Achse und die *y*-Achse bei der Auswahl dieser Option nicht verändert werden.

Liste. Ordnet Fenster wie eine Liste an, in der jede Zelle eine Kombination von Werten darstellt. Spalten und Zeilen sind keinen einzelnen Werten mehr zugewiesen. Mit dieser Option können die Fenster bei Bedarf umbrochen werden.

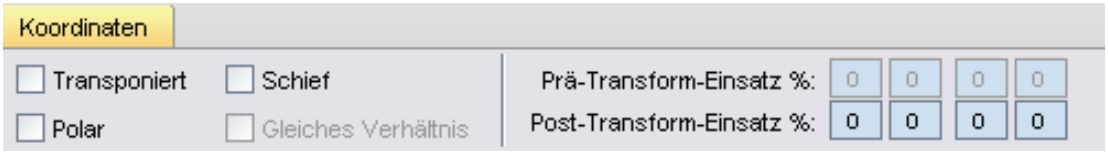
Transformieren des Koordinatensystems

Viele Visualisierungen werden in einem flachen, rechteckigen Koordinatensystem angezeigt. Sie können das Koordinatensystem wie erforderlich umformen. Sie können dem Koordinatensystem beispielsweise eine polare Transformation zuweisen, schräge Schatteneffekte hinzufügen und die Achsen transponieren. Sie können diese Transformationen auch rückgängig machen, wenn sie der aktuellen Visualisierung bereits zugewiesen sind. Beispiel: Ein Kreisdiagramm ist in einem Polarkoordinatensystem gezeichnet. Nach Wunsch können Sie die Polartransformation rückgängig machen und das Kreisdiagramm als einzelnes gestapeltes Balkendiagramm in einem rechteckigen Koordinatensystem anzeigen.

So transformieren Sie das Koordinatensystem:

- Wählen Sie das zu transformierende Koordinatensystem aus. Sie wählen das Koordinatensystem aus, indem Sie den Rahmen um das individuelle Diagramm auswählen.
- Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte Koordinaten.

Abbildung 5-112
Registerkarte "Koordinaten"



- Wählen Sie die Transformationen aus, die Sie dem Koordinatensystem zuweisen möchten. Sie können die Auswahl einer Transformation auch aufheben, um sie rückgängig zu machen.

Transponiert. Die Änderung der Ausrichtung der Achsen wird als **Transponieren** bezeichnet. Es gleicht dem Vertauschen der vertikalen und horizontalen Achsen in einer 2D-Visualisierung.

Polar. Eine Polartransformation zeichnet die Grafikelemente in einem bestimmten Winkel und Abstand vom Mittelpunkt des Diagramms. Ein Kreisdiagramm ist eine 1D-Visualisierung mit einer Polartransformation, die einzelne Balken in bestimmten Winkeln zeichnet. Ein Radardiagramm ist eine 2D-Visualisierung mit einer Polartransformation, die Grafikelemente in bestimmten Winkeln und Abständen von der Mitte des Diagramms zeichnet. Eine 3D-Visualisierung würde eine zusätzliche Tiefendimension umfassen.

Schräg. Eine schräge Transformation fügt den Grafikelementen einen 3D-Effekt hinzu. Diese Transformation verleiht den Grafikelementen mehr Tiefe, die aber rein dekorativen Zwecken dient. Sie wird durch keine bestimmten Datenwerte beeinflusst.

Gleiches Verhältnis. Wenn Sie dasselbe Verhältnis zuweisen, repräsentiert derselbe Abstand auf jeder Skale denselben Abstand in den Datenwerten. Beispielsweise repräsentieren 2 cm auf beiden Skalen einen Abstand von 1.000.

% Einsatz vor der Transformation. Wenn Achsen nach der Transformation abgeschnitten sind, sollten Sie dem Diagramm Einsätze hinzufügen, bevor Sie die Transformation zuweisen. Die Einsätze schrumpfen die Abmessungen um einen bestimmten Prozentsatz, bevor dem Koordinatensystem etwaige Transformationen zugewiesen werden. Sie können die unteren *x*-, oberen *x*-, unteren *y*- und oberen *y*-Abmessungen in dieser Reihenfolge steuern.

% Einsatz nach der Transformation. Wenn Sie das Seitenverhältnis eines Diagramms ändern möchten, können Sie dem Diagramm Einsätze hinzufügen, nachdem die Transformation zugewiesen wurde. Die Einsätze schrumpfen die Abmessungen um einen bestimmten Prozentsatz, nachdem dem Koordinatensystem etwaige Transformationen zugewiesen wurden. Diese Einsätze können auch zugewiesen werden, wenn dem Diagramm keine Transformation zugewiesen wird. Sie können die unteren *x*-, oberen *x*-, unteren *y*- und oberen *y*-Abmessungen in dieser Reihenfolge steuern.

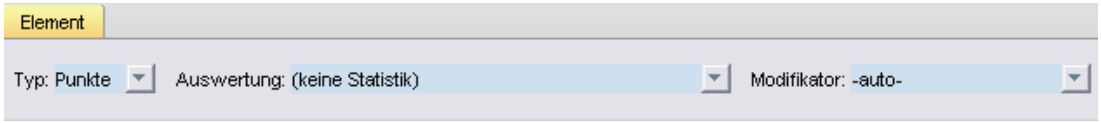
Ändern von Statistiken und Grafikelementen

Sie können ein in einen anderen Typ umwandeln, die zum Zeichnen des Grafikelements verwendete Statistik ändern oder den Kollisionsmodifikator angeben, der bestimmt, was geschieht, wenn sich Grafikelemente überlagern.

So konvertieren Sie ein Grafikelement:

- Wählen Sie das Grafikelement aus, das Sie konvertieren möchten.
- Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte Element.

Abbildung 5-113
Registerkarte "Element"



- Wählen Sie einen neuen Grafikelementtyp aus der Liste "Typ".

Grafikelementstypen	Beschreibung
Punkt	Eine Markierung, die den jeweiligen Datenpunkt identifiziert. Ein Punktelement wird in Streudiagrammen und anderen verwandten Visualisierungen verwendet.
Intervall	Eine rechteckige Form, die über einen bestimmten Datenwert gezogen wird und den Bereich zwischen einem Ursprung und einem anderen Datenwert ausfüllt. Ein Intervallelement wird in Balkendiagrammen und Histogrammen verwendet.
Linie	Eine Linie, die Datenwerte verbindet.
Pfad	Eine Linie, die Datenwerte in der Reihenfolge ihres Erscheinens im Daten-Set verbindet.
Bereich	Eine Linie, die Datenelemente mit dem ausgefüllten Bereich zwischen der Linie und einem Ursprung verbindet.
Polygon	Eine mehrseitige Form, die einen Datenbereich umschließt. Ein Polygonelement kann etwa in klassierten Streudiagrammen oder in einer Karte verwendet werden.
Schema	Ein Element, bestehend aus einer Box mit sogenannten Whiskers und Markierungen, die auf Ausreißer hinweisen. Eine Schema wird für Boxplots verwendet.

So ändern Sie die Statistik:

- ▶ Wählen Sie das Grafikelement aus, dessen Statistik Sie ändern möchten.
- ▶ Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte Element.
- ▶ Wählen Sie aus der Dropdown-Liste "Auswertung" eine neue Statistik. Beachten Sie, dass bei der Auswahl einer Statistik die Daten zusammengefasst werden. Wenn die Visualisierung stattdessen nicht zusammengefasste Daten anzeigen soll, wählen Sie (keine Statistik) aus der Liste "Auswertung".

Anhand eines stetigen Felds berechnete statistische Funktionen

- **Mittelwert.** Ein Lagemaß. Die Summe der Ränge, geteilt durch die Zahl der Fälle.
- **Median.** Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).
- **Modalwert.** Der am häufigsten auftretende Wert. Wenn mehrere Werte gleichermaßen die größte Häufigkeit aufweisen, ist jeder von ihnen ein Modalwert.
- **Minimum.** Der kleinste Wert einer numerischen Variablen.
- **Maximum.** Der größte Wert einer numerischen Variablen.
- **Bereich**. Differenz zwischen Mindest- und Höchstwert.
- **Mittelbereich**. Die Mitte des Bereichs, d. h. der Wert, dessen Differenz zum Minimum gleich seiner Differenz zum Maximum ist.
- **Summe.** Die Summe der Werte über alle Fälle mit nichtfehlenden Werten.
- **Kumulierte Summe**. Die kumulierte Summe der Werte. Jedes Grafikelement zeigt die Summe für eine Untergruppe und die Gesamtsumme aller vorherigen Gruppen an.
- **Prozentsumme**. Die Prozentzahl innerhalb jeder Untergruppe basierend auf einem summierten Feld im Vergleich zur Summe über alle Gruppen hinweg.
- **Kumulierte Prozentsumme**. Die kumulative Prozentzahl innerhalb jeder Untergruppe basierend auf einem summierten Feld im Vergleich zur Summe über alle Gruppen hinweg. Jedes Grafikelement zeigt die Prozentzahl für eine Untergruppe und die Gesamtprozentzahl aller vorherigen Gruppen an.
- **Variance.** Ein Maß der Streuung um den Mittelwert. Es ist gleich dem Quotienten aus der Summe der quadrierten Abweichung vom Mittelwert und der um 1 verringerten Fallanzahl. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.
- **Standardabweichung.** Ein Maß für die Streuung um den Mittelwert. In einer Normalverteilung liegen 68% der Fälle innerhalb von einer Standardabweichung des Mittelwerts und 95% der Fälle innerhalb von zwei Standardabweichungen. Wenn beispielsweise für das Alter der Mittelwert 45 und die Standardabweichung 10 beträgt, liegen bei einer Normalverteilung 95 % der Fälle im Bereich zwischen 25 und 65.
- **Standardfehler.** Ein Maß für die Abweichung des Werts einer Teststatistik zwischen Stichproben. Dies ist die Standardabweichung der Stichprobenverteilung einer Statistik. So ist z. B. der Standardfehler des Mittelwerts die Standardabweichung des Stichprobenmittelwerts.

- **Kurtosis.** Ein Maß dafür, wie sich die Beobachtungen um einen zentralen Punkt gruppieren. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremverten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.
- **Schiefe.** Ein Maß für die Asymmetrie einer Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefe-Wert, der mehr als doppelt so groß ist wie sein Standardfehler, für eine Abweichung von der Symmetrie spricht.

Bei folgenden Regionsstatistiken entsteht unter Umständen mehr als ein Grafikelement pro Untergruppe. Bei der Verwendung der Grafikelemente für Intervalle, Bereiche oder Kanten entsteht in einer Regionsstatistik mehr als ein Grafikelement, das den Bereich anzeigt. Alle anderen Grafikelemente erzeugen zwei getrennte Elemente, wobei das eine den Beginn und das andere das Ende des Bereichs anzeigt.

- **Region: Bereich**. Wertebereich zwischen Mindest- und Höchstwert.
- **Region: 95-%-Konfidenzintervall für den Mittelwert**. Ein Wertebereich mit einer 95%igen Wahrscheinlichkeit, die Grundgesamtheit zu enthalten.
- **Region: 95-%-Konfidenzintervall für einzelne Fälle**. Ein Wertebereich mit einer 95%igen Wahrscheinlichkeit, den vorhergesagten Wert angesichts des individuellen Falls zu enthalten.
- **Region: 1 Standardabweichung über/unter dem Mittelwert**. Ein Wertebereich zwischen 1 **Standardabweichung** über und unter dem **Mittelwert**.
- **Region: 1 Standardfehler über/unter dem Mittelwert**. Ein Wertebereich zwischen 1 **Standardfehler** über und unter dem **Mittelwert**.

Anzahlbasierte statistische Funktionen

- **Anzahl**. Die Anzahl der Zeilen/Fälle.
- **Kumulierte Anzahl**. Die kumulierte Anzahl der Zeilen/Fälle. Jedes Grafikelement zeigt die Anzahl für eine Untergruppe und die Gesamtanzahl aller vorherigen Gruppen an.
- **Häufigkeitsprozent**. Die Prozentzahl an Zeilen/Fällen in jeder Untergruppe im Vergleich zur Gesamtzahl an Zeilen/Fällen.
- **Kumulierte Häufigkeitsprozente**. Die kumulierte Prozentzahl an Zeilen/Fällen in jeder Untergruppe im Vergleich zur Gesamtzahl an Zeilen/Fällen. Jedes Grafikelement zeigt die Prozentzahl für eine Untergruppe und die Gesamtprozentzahl aller vorherigen Gruppen an.

So geben Sie den Kollisionsmodifikator an:

Der Kollisionsmodifikator bestimmt, was geschieht, wenn sich Grafikelemente überlagern.

- Wählen Sie das Grafikelement aus, für das Sie den Kollisionsmodifikator angeben möchten.
- Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte Element.
- Wählen Sie aus dem Dropdown-Listenfeld "Modifikator" einen Kollisionsmodifikator aus. -auto- überlässt es der Anwendung zu bestimmen, welcher Kollisionsmodifikator für den Grafikelementtyp und die Statistik geeignet ist.

  Überlagert. Zeichnet Grafikelemente mit demselben Wert übereinander.

  Stapeln. Stapelt Grafikelemente, die einander normalerweise überdecken würden, wenn sie dieselben Datumswerte besitzen.

  Ausweichen. Verschiebt Grafikelemente neben andere Grafikelemente, die am gleichen Wert erscheinen, anstatt sie übereinanderzudecken. Die Grafikelemente werden symmetrisch angeordnet. D. h. die Grafikelemente werden an entgegengesetzten Seiten einer zentralen Position verschoben. Ausweichen ist dem Clustering sehr ähnlich.

  Übereinander. Verschiebt Grafikelemente neben andere Grafikelemente, die am gleichen Wert erscheinen, anstatt sie übereinanderzudecken. Die Grafikelemente werden asymmetrisch angeordnet. D. h., die Grafikelemente werden schräg übereinander angeordnet, wobei das untere Grafikelement an einem bestimmten Wert auf der Skala positioniert ist.

  Streuen (normal). Positioniert Grafikelemente am selben Datenwert anhand einer Normalverteilung zufällig um.

  Streuen (uniform). Positioniert Grafikelemente am selben Datenwert anhand einer Uniformverteilung zufällig um.

Ändern der Position der Legende

Wenn ein Diagramm eine Legende umfasst, wird diese in der Regel rechts neben dem Diagramm angezeigt. Diese Position können Sie verändern.

So ändern Sie die Position der Legende

- Wählen Sie die Legende aus.
- Klicken Sie im Fenster "Eigenschaften" auf die Registerkarte Legende.

Abbildung 5-114
Registerkarte "Legende"

- Wählen Sie eine Position.

Kopieren von Visualisierungen und Visualisierungsdaten

Die Palette "Allgemein" enthält Schaltflächen zum Kopieren der Visualisierung und ihrer Daten.

Abbildung 5-116
Schaltfläche "Visualisierung kopieren"

Kopieren der Visualisierung. Diese Aktion kopiert die Visualisierung als Bild in die Zwischenablage. Mehrere Bildformate stehen zur Verfügung. Wenn Sie das Bild in eine andere Anwendung einfügen, können Sie eine "Einfügen-Spezial"-Option wählen, um eines der verfügbaren Bildformate zum Einfügen festzulegen.

Abbildung 5-117
Schaltfläche "Visualisierungsdaten kopieren"

Kopieren der Visualisierungsdaten. Diese Aktion kopiert die zugrundeliegenden Daten, anhand denen die Visualisierung gezeichnet wird. Die Daten werden als Standardtext oder HTML-Text in die Zwischenablage kopiert. Wenn Sie die Daten in eine andere Anwendung einfügen, können Sie eine "Einfügen-Spezial"-Option wählen, um eines dieser Formate zum Einfügen festzulegen.

Direktzugriffstasten

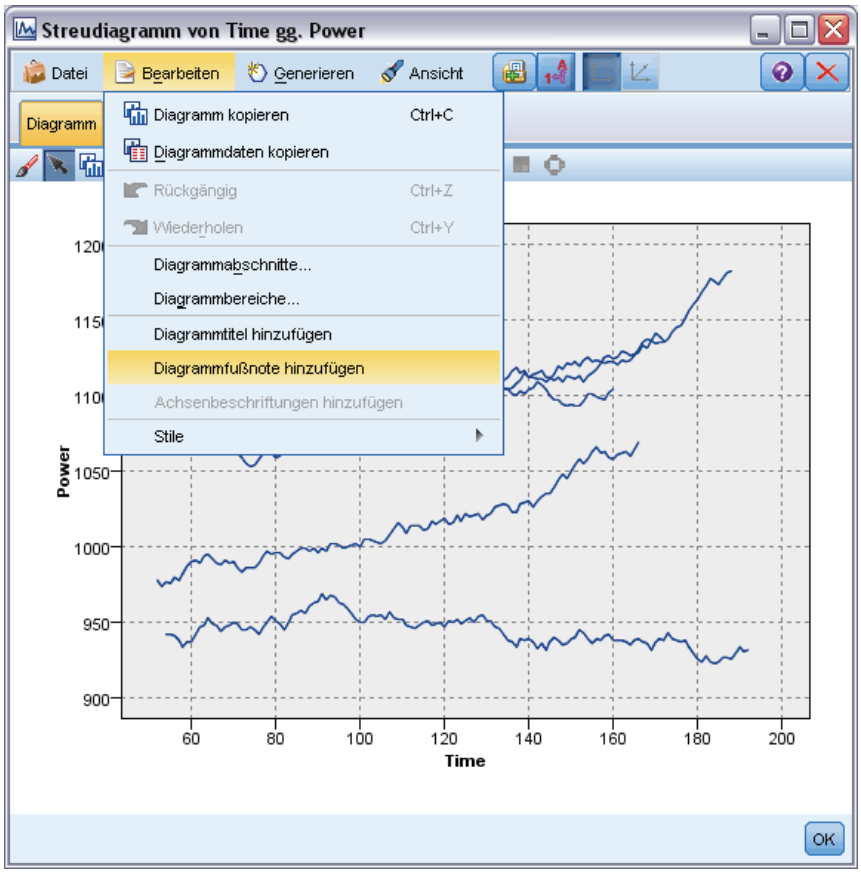
Tabelle 5-4
Direktzugriffstasten

Tastenkürzel	Function
Strg+Leerzeichen	Umschalten zwischen Sondierungs- und Bearbeitungsmodus
Delete	Löschen eines Visualisierungselements
Strg+Z	Rückgängig
Strg+Y	Wiederholen
F2	Gliederung für ausgewählte Elemente im Diagramm anzeigen

Hinzufügen von Titeln und Fußnoten

Bei allen Diagrammtypen können Sie einen eindeutigen Titel, eine Fußnote oder Achsenbeschriftungen hinzufügen, um deutlicher zu machen, was im Diagramm angezeigt wird.

Abbildung 5-118
Hinzufügen einer Diagrammfußnote



Hinzufügen von Titeln zu Diagrammen

- Wählen Sie in den Menüs die Optionsfolge Bearbeiten > Diagrammtitel hinzufügen aus. Ein Textfeld, das <TITLE> enthält, wird oberhalb des Diagramms angezeigt.
- Vergewissern Sie sich, dass Sie sich im Bearbeitungsmodus befinden. Wählen Sie in den Menüs die Optionsfolge Ansicht > Bearbeitungsmodus.
- Doppelklicken Sie auf den Text <TITLE>.
- Geben Sie den gewünschten Titel ein und drücken Sie die Eingabetaste.

Hinzufügen von Fußnoten zu Diagrammen

- Wählen Sie in den Menüs die Optionsfolge Bearbeiten > Diagrammfußnote hinzufügen aus. Ein Textfeld, das <FOOTNOTE> enthält, wird unterhalb des Diagramms angezeigt.
- Vergewissern Sie sich, dass Sie sich im Bearbeitungsmodus befinden. Wählen Sie in den Menüs die Optionsfolge Ansicht > Bearbeitungsmodus.
- Doppelklicken Sie auf den Text <FOOTNOTE>.

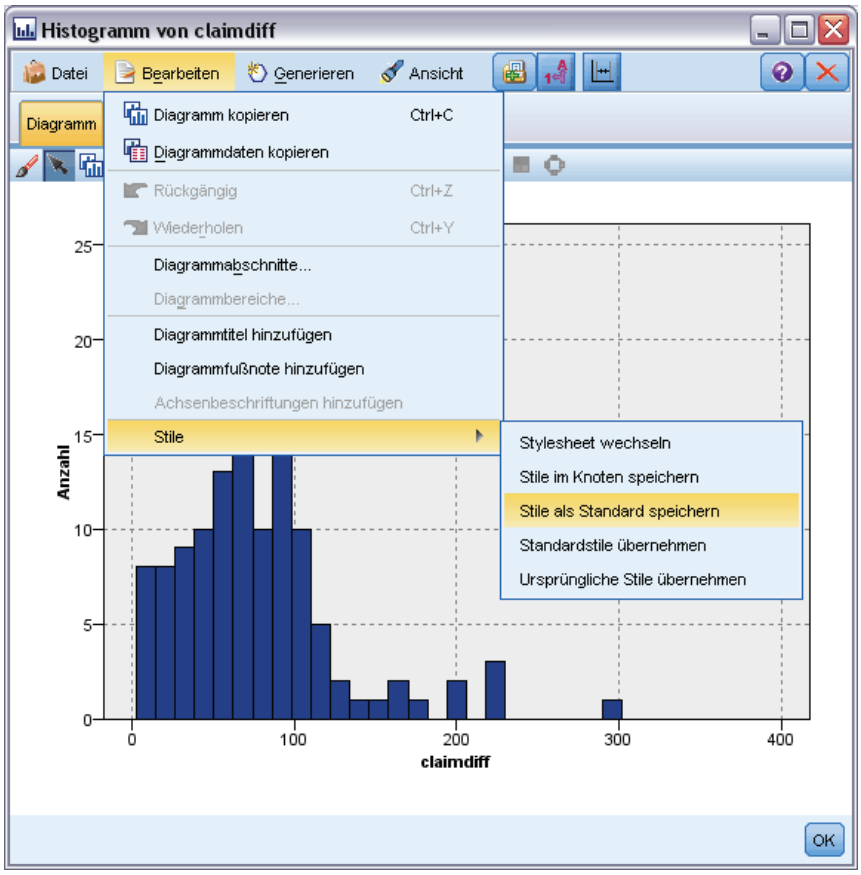
▶ Geben Sie den gewünschten Titel ein und drücken Sie die Eingabetaste.

Verwenden von Diagramm-Stylesheets

Die Grundlegenden Informationen zur Anzeige von Diagrammen, wie Farben, Schriftarten, Symbole und Linienstärke werden über ein Stylesheet festgelegt. Im Lieferumfang von IBM® SPSS® Modeler ist ein Standard-Stylesheet enthalten; Sie können jedoch, falls erforderlich, Änderungen daran vornehmen. Beispielsweise gilt möglicherweise in Ihrem Unternehmen ein Farbschema für Präsentationen, das Sie auch in Ihren Diagrammen verwenden möchten. Für weitere Informationen siehe Thema Bearbeiten von Visualisierungen auf S. 372.

In den Diagrammknoten können Sie mithilfe des Bearbeitungsmodus Stiländerungen am Erscheinungsbild eines Diagramms vornehmen. Anschließend können Sie im Menü Bearbeiten > Stile die Änderungen als Stylesheet speichern, das auf alle Diagramme angewendet wird, die Sie danach aus dem Diagrammknoten generieren, oder als neues Standard-Stylesheet, das für alle Diagramme gilt, die Sie mit SPSS Modeler erstellen.

Abbildung 5-119
Auswählen von Diagrammstilen



Im Menü “Bearbeiten” stehen über die Option Stile fünf Stylesheet-Optionen zur Verfügung:

- **Stylesheet wechseln.** Damit wird eine Liste von verschiedenen gespeicherten Stylesheets angezeigt, die wählen können, um das Erscheinungsbild Ihrer Diagramme zu ändern. Für weitere Informationen siehe Thema Stylesheets anwenden auf S. 393.
- **Stile im Knoten speichern.** Dadurch werden Änderungen an den Stilen des ausgewählten Diagramms gespeichert, um sie auf alle zukünftigen Diagramme anzuwenden, die über denselben Diagrammknoten im aktuellen Stream erstellt werden.
- **Stile als Standard speichern.** Dadurch werden Änderungen an den Stilen des ausgewählten Diagramms gespeichert, um sie auf alle zukünftigen Diagramme anzuwenden, die über beliebigen Diagrammknoten in einem beliebigen Stream erstellt werden. Nach Auswahl dieser Option können Sie mit Standardstile übernehmen den Stil auch auf alle bestehenden Diagramme anwenden.
- **Standardstile übernehmen.** Ändert die Stile des ausgewählten Diagramms auf die derzeit gespeicherten Standardstile.
- **Ursprüngliche Stile übernehmen.** Ändert die Stile eines Diagramms zurück auf die ursprünglichen, mitgelieferten Standardstile.

Stylesheets anwenden

Sie können ein Visualisierungs-Stylesheet anwenden, das stilistische Eigenschaften der Visualisierung festlegt. Zum Beispiel kann ein Stylesheet unter anderem Schriftarten, Strichmuster und Farben definieren. Bis zu einem gewissen Grad bieten Stylesheets einen Direktzugriff zu Änderungen, die Sie sonst manuell vornehmen müssten. Beachten Sie jedoch, dass sich mit einem Stylesheet nur Änderungen am *Stil* vornehmen lassen. Andere Änderungen wie etwa die Position der Legende oder der Skalenbereich werden nicht im Stylesheet gespeichert.

So wenden Sie ein Stylesheet an

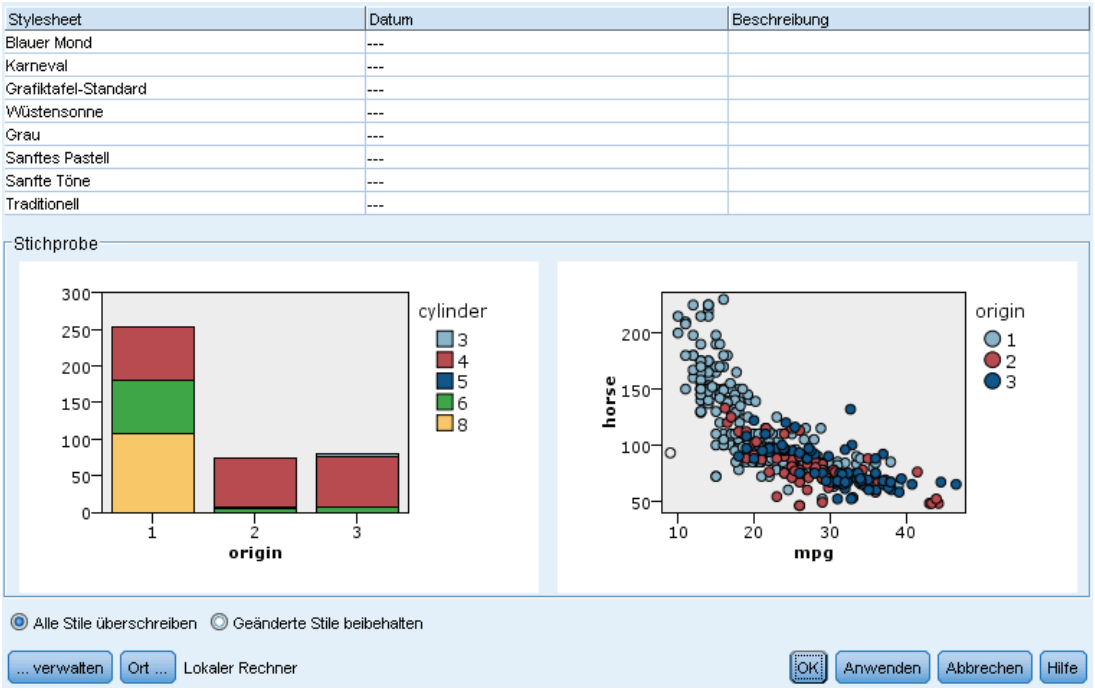
▶ Wählen Sie die folgenden Befehle aus den Menüs aus:
Edit (Bearbeiten) > Stile > Stylesheet wechseln

▶ Verwenden Sie das Dialogfeld “Stylesheet wechseln”, um ein Stylesheet auszuwählen.

▶ Klicken Sie auf Anwenden, um das Stylesheet auf die Visualisierung anzuwenden, ohne das Dialogdfeld zu schließen. Klicken Sie auf OK, um das Stylesheet anzuwenden und das Dialogfeld zu schließen.

Dialogfeld "Stylesheet wechseln/auswählen"

Abbildung 5-120
Dialogfeld "Stylesheet wechseln"



Die Tabelle im oberen Bereich des Dialogfelds enthält alle derzeit verfügbaren Visualisierungs-Stylesheets. Einige Stylesheets sind vorinstalliert, während andere unter Umständen erst in IBM® SPSS® Visualization Designer (ein separates Produkt) erstellt werden müssen.

Im unteren Bereich des Dialogfelds befinden sich Beispiel-Visualisierungen mit Beispieldaten. Wählen Sie ein Stylesheet aus, um dessen Stile auf die Beispiel-Visualisierungen anzuwenden. Mit Hilfe dieser Beispiele können Sie sehen, wie sich das Stylesheet auf Ihre Visualisierung auswirkt.

Das Dialogfeld bietet außerdem die folgenden Optionen:

Vorhandene Stile. Standardmäßig kann ein Stylesheet alle Stile der Visualisierung überschreiben. Sie können dieses Verhalten ändern.

- **Alle Stile überschreiben**. Bei der Anwendung des Stylesheets werden alle Stile der Visualisierung überschrieben, einschließlich jener Stile, die während des aktuellen Änderungsvorgangs in der Visualisierung bearbeitet wurden.
- **Geänderte Stile beibehalten**. Bei der Anwendung des Stylesheets werden jene Stile überschrieben, die *nicht* während des aktuellen Änderungsvorgangs in der Visualisierung bearbeitet wurden. Die Stile, die während des aktuellen Änderungsvorgangs bearbeitet wurden, werden beibehalten.

Verwalten. Verwalten von Visualisierungsvorlagen, Stylesheets und Karten auf dem Computer. Sie können Visualisierungsvorlagen , Stylesheets und Karten auf Ihrem lokalen Rechner importieren, exportieren, umbenennen und löschen. Für weitere Informationen siehe Thema Verwalten von Vorlagen, Stylesheets und Kartendateien auf S. 292.

Speicherort. Ändern des Speicherorts von Visualisierungsvorlagen, Stylesheets und Karten. Der aktuelle Speicherort wird rechts neben der Schaltfläche angezeigt. Für weitere Informationen siehe Thema Einstellen des Speicherorts für Vorlagen, Stylesheets und Karten auf S. 290.

Drucken, Speichern, Kopieren und Exportieren von Diagrammen

Jedes Diagramm weist eine Reihe von Optionen auf, mit denen Sie das Diagramm speichern oder drucken oder in ein anderes Format exportieren können. Die meisten dieser Optionen stehen über das Menü “Datei” zur Verfügung. Außerdem können Sie im Menü “Bearbeiten” auswählen, dass das Diagramm oder die enthaltenen Daten zur Verwendung in einer anderen Anwendung kopiert werden soll.

Abbildung 5-121
Menü “Datei” und Symbolleiste für Diagrammfenster

Drucke

▶ Zum Drucken des Diagramms können Sie das Menüelement bzw. die Schaltfläche Drucken verwenden. Vor dem Drucken können Sie mithilfe von Seite einrichten und Druckvorschau die Druckoptionen festlegen und eine Vorschau der Ausgabe anzeigen.

Speichern von Diagrammen

▶ Um ein Diagramm in einer IBM® SPSS® Modeler-Ausgabedatei (**.cou*) zu speichern, müssen Sie in den Menüs die Optionsfolge Datei > Speichern bzw. Datei > Speichern unter auswählen.

oder

Um das Diagramm im Repository zu speichern, wählen Sie in den Menüs die Optionsfolge Datei > Ausgabe speichern.

Kopieren von Diagrammen

▶ Um das Diagramm zur Verwendung in einer anderen Anwendung, beispielsweise in MS Word oder MS PowerPoint, zu kopieren, wählen Sie in den Menüs die Optionsfolge Bearbeiten > Diagramm kopieren.

Kopieren von Daten

▶ Um die Daten zur Verwendung in einer anderen Anwendung, beispielsweise in MS Excel oder MS Word, zu kopieren, wählen Sie in den Menüs die Optionsfolge Bearbeiten > Daten kopieren. Standardmäßig werden die Daten als HTML formatiert. Verwenden Sie in der anderen Anwendung die Option Inhalte einfügen, damit beim Einfügen andere Formatierungsoptionen angezeigt werden.

Exportieren von Diagrammen

Mit der Option Diagramm exportieren können Sie das Diagramm in einem der folgenden Formate exportieren: Bitmap (*.bmp*), JPEG (*.jpg*), PNG (*.png*), HTML (*.html*) oder ViZml-Dokument (*.xml*) zur Verwendung in anderen IBM® SPSS® Statistics-Anwendungen.

▶ Wählen Sie zum Exportieren von Diagrammen in den Menüs die Optionsfolge Datei > Diagramm exportieren aus und wählen Sie dann das Format aus.

Exportieren von Tabellen

Mit der Option Tabelle exportieren können Sie die Tabelle in einem der folgenden Formate exportieren: tabulatorgetrennt (*.tab*), kommagetrennt (*.csv*) oder HTML (*.html*)

▶ Wählen Sie zum Exportieren von Tabellen in den Menüs die Optionsfolge Datei > Tabelle exportieren aus und wählen Sie dann das Format aus.

Kapitel 6

Ausgabeknoten

Überblick über Ausgabeknoten

Mit Ausgabeknoten erhalten Sie Informationen zu Ihren Daten und Modellen. Sie bieten außerdem einen Mechanismus zum Exportieren von Daten in verschiedenen Formaten, sodass Sie diese Daten auch mit anderen Software-Tools nutzen können.

Folgende Ausgabeknoten stehen zur Verfügung:

Der Tabellenknoten zeigt die Daten in Tabellenform an, die auch in eine Datei geschrieben werden kann. Diese Vorgehensweise empfiehlt sich immer dann, wenn die Datenwerte überprüft oder in leicht lesbarer Form exportiert werden sollen. Für weitere Informationen siehe Thema Tabellenknoten auf S. 405.

Der Matrixknoten erstellt eine Tabelle, die die Beziehungen zwischen den Feldern aufzeigt. Dieser Knoten dient am häufigsten zur Darstellung der Beziehung zwischen zwei symbolischen Feldern, kann jedoch auch zum Aufzeigen der Beziehungen zwischen Flag-Feldern oder numerischen Feldern herangezogen werden. Für weitere Informationen siehe Thema Matrixknoten auf S. 410.

Der Analyseknoten evaluiert die Fähigkeit von Vorhersagemodellen, genaue Vorhersagen zu generieren. Mit Analyseknoten werden verschiedene Vergleiche zwischen den vorhergesagten Werten und den tatsächlichen Werten für ein oder mehrere Modell-Nuggets angestellt. Sie können außerdem Vorhersagemodelle miteinander vergleichen. Für weitere Informationen siehe Thema Analyseknoten auf S. 415.

Der Data Audit-Knoten bietet einen umfassenden ersten Einblick in die Daten mit statistischen Funktionen, Histogrammen und der Verteilung für die einzelnen Felder sowie Informationen zu Ausreißern, fehlenden Werten und Extremwerten. Die Ergebnisse werden in einer übersichtlichen Matrix dargestellt, die sortiert werden kann und als Grundlage für die Erzeugung normal großer Diagramme und Datenvorbereitungsknoten dient. Für weitere Informationen siehe Thema Data Audit-Knoten auf S. 420.

Mit dem Transformationsknoten können Sie die Ergebnisse von Transformationen auswählen und in einer Vorschau anzeigen, bevor Sie sie auf ausgewählte Felder anwenden. Für weitere Informationen siehe Thema Transformationsknoten auf S. 436.

Der Statistikknoten liefert grundlegende Übersichtsdaten zu numerischen Feldern. Er berechnet Übersichtsstatistiken für einzelne Felder und für die Korrelationen zwischen den Feldern. Für weitere Informationen siehe Thema Statistikknoten auf S. 442.

Der Mittelwertknoten vergleicht die Mittelwerte zwischen unabhängigen Gruppen oder zwischen Paaren von in Bezug stehenden Feldern, um zu testen, ob ein signifikanter Unterschied vorliegt. So können Sie beispielsweise die Einnahmen vor und nach der Durchführung einer Werbeaktion vergleichen oder die Einnahmen, die von Kunden stammen, die keine Werbezettel erhielten, mit den Einnahmen von Kunden vergleichen, die von der Werbeaktion erreicht wurden. Für weitere Informationen siehe Thema Mittelwertknoten auf S. 447.

Der Berichtknoten erstellt formatierte Berichte, die sowohl festen Text als auch Daten und andere aus den Daten abgeleitete Ausdrücke enthalten. Das Format des Berichts wird mithilfe von Textvorlagen festgelegt, mit denen der feste Text und die Datenausgabekonstruktionen definiert werden. Sie können eine benutzerdefinierte Textformatierung angeben; hierzu stehen HTML-Tags in der Vorlage sowie Optionen auf der Registerkarte "Ausgabe" zur Verfügung. Sie können Datenwerte und andere bedingte Ausgaben mithilfe von CLEM-Ausdrücken in der Vorlage aufnehmen. Für weitere Informationen siehe Thema Berichtknoten auf S. 453.

Mit dem Globalwerteknoten werden die Daten gescannt und Übersichtswerte berechnet, die in CLEM-Ausdrücken herangezogen werden können. Mit diesem Knoten können Sie beispielsweise die Statistiken für das Feld *Alter* berechnen und dann den Gesamtmittelwert für *Alter* in CLEM-Ausdrücken verwenden. Fügen Sie hierzu die Funktion @GLOBAL_MEAN(alter) ein. Für weitere Informationen siehe Thema Globalwerteknoten auf S. 456.

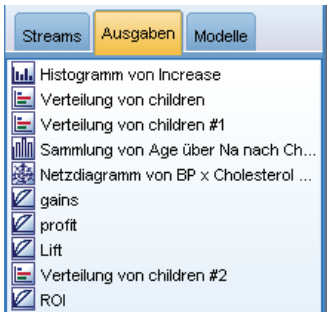
Verwalten der Ausgabe

Der Ausgabe-Manager zeigt die Diagramme, Grafiken und Tabellen an, die während einer IBM® SPSS® Modeler-Sitzung erstellt wurden. Sie können eine Ausgabe jederzeit erneut öffnen, indem Sie im Manager darauf doppelklicken. Der entsprechende Stream bzw. Knoten muss nicht erneut ausgeführt werden.

So zeigen Sie den Ausgabe-Manager an:

▶ Öffnen Sie das Menü "Ansicht" und wählen Sie Manager. Klicken Sie auf die Registerkarte Ausgaben.

Abbildung 6-1
Ausgabe-Manager



Im Ausgabe-Manager haben Sie folgende Möglichkeiten:

- Vorhandene Ausgabeobjekte anzeigen, beispielsweise Histogramme, Evaluationsdiagramme oder Tabellen.

- Ausgabeobjekte umbenennen.
- Ausgabeobjekte auf Datenträger oder im IBM® SPSS® Collaboration and Deployment Services Repository (sofern verfügbar) speichern.
- Ausgabedateien zum aktuellen Projekt hinzufügen.
- Ungespeicherte Ausgabeobjekte aus der aktuellen Sitzung löschen.
- Gespeicherte Ausgabeobjekte öffnen oder aus dem IBM SPSS Collaboration and Deployment Services Repository (sofern verfügbar) abrufen.

Um auf diese Optionen zuzugreifen, klicken Sie mit der rechten Maustaste auf eine beliebige Stelle auf der Registerkarte “Ausgaben”.

Anzeigen der Ausgabe

Die Bildschirmausgabe wird in einem Ausgabe-Browser-Fenster angezeigt. Das Ausgabe-Browser-Fenster weist einen eigenen Satz an Menüs auf, mit denen Sie die Ausgabe drucken bzw. speichern bzw. in ein anderes Format exportieren können. Beachten Sie, dass die spezifischen Optionen je nach Ausgabetyp unterschiedlich sein können.

Drucken, Speichern und Exportieren von Daten. Folgende weitere Informationen sind verfügbar:

- Zum Drucken der Ausgabe können Sie die Menüoption bzw. Schaltfläche Drucken verwenden. Vor dem Drucken können Sie mithilfe von Seite einrichten und Druckvorschau die Druckoptionen festlegen und eine Vorschau der Ausgabe anzeigen.
- Zum Speichern der Ausgabe in einer IBM® SPSS® Modeler-Ausgabedatei (*.cou*) wählen Sie im Menü “Datei” die Option Speichern bzw. Speichern unter.
- Um die Ausgabe in einem anderen Format zu speichern, beispielsweise als Text oder HTML, wählen Sie im Menü “Datei” die Option Exportieren. Für weitere Informationen siehe Thema Exportieren von Ausgaben auf S. 403.
- Um die Ausgabe in einem gemeinsam benutzten Repository zu speichern, damit andere Benutzer sie mit dem IBM® SPSS® Collaboration and Deployment Services Deployment Portal anzeigen können, wählen Sie Im Web veröffentlichen aus dem Menü “Datei”. Beachten Sie, dass für IBM® SPSS® Collaboration and Deployment Services eine gesonderte Lizenz erforderlich ist.

Auswählen von Zellen und Spalten. Das Menü “Bearbeiten” enthält verschiedene Optionen, mit denen Zellen und Spalten, ausgewählt oder kopiert werden können bzw. ihre Auswahl aufgehoben werden, je nachdem was für den aktuellen Ausgabetyp geeignet ist. Für weitere Informationen siehe Thema Auswählen von Zellen und Spalten auf S. 404.

Erzeugen neuer Knoten. Im Menü “Generieren” können Sie neue Knoten auf der Grundlage des Inhalts des Ausgabe-Browsers erzeugen. Die Optionen variieren je nach Ausgabetyp und aktuell in der Ausgabe ausgewählten Elementen. Einzelheiten zu den Knotengenerierungsoptionen für einen bestimmten Ausgabetyp finden Sie in der Dokumentation für die betreffende Ausgabe.

Im Web veröffentlichen

Mit der Funktion "Im Web veröffentlichen" können Sie bestimmte Typen von Stream-Ausgaben in einem zentralen, gemeinsam benutzten IBM® SPSS® Collaboration and Deployment Services Repository veröffentlichen, das die Grundlage von IBM® SPSS® Collaboration and Deployment Services bildet. Wenn Sie diese Option verwenden, können andere Benutzer diese Ausgabe über einen Internet-Zugang und ein IBM SPSS Collaboration and Deployment Services-Konto ansehen; sie müssen nicht über eine Installation von IBM® SPSS® Modeler verfügen.

Hinweis: Für den Zugriff auf ein IBM SPSS Collaboration and Deployment Services-Repository ist eine separate Lizenz erforderlich. Weitere Informationen finden Sie im Dokument *http://www.ibm.com/software/analytics/spss/products/deployment/cds/*

Die folgende Tabelle listet die SPSS Modeler-Knoten auf, die die Funktion "Im Web veröffentlichen" unterstützen. Ausgabe von diesen Knoten wird im IBM SPSS Collaboration and Deployment Services Repository in Ausgabeobjekt-Format (*.cou*) gespeichert und kann direkt in der IBM® SPSS® Collaboration and Deployment Services Deployment Portal angezeigt werden.

Andere Ausgabetypen können nur angezeigt werden, wenn die betreffende Anwendung (z. B. SPSS Modeler für Stream-Objekte) auf dem Rechner des Benutzers installiert ist.

Tabelle 6-1
Knoten, die "Im Web veröffentlichen" unterstützen

Knotentyp	**Knoten**
Grafiken	all
Ausgabe	Tabelle
	Matrix
	Data Audit
	Transformieren
	Mittelwerte
	Analyse
	Statistics
	Bericht (HTML)
IBM® SPSS® Statistics	Statistikausgabe

Veröffentlichen von Ausgabeobjekten im Web

So veröffentlichen Sie Ausgabeobjekte im Web:

- Führen Sie in einem IBM® SPSS® Modeler-Stream einen der in der Tabelle aufgeführten Knoten aus. Damit wird ein Ausgabeobjekt (z. B. eine Tabelle, eine Matrix oder ein Berichtobjekt) in einem neuen Fenster erstellt.

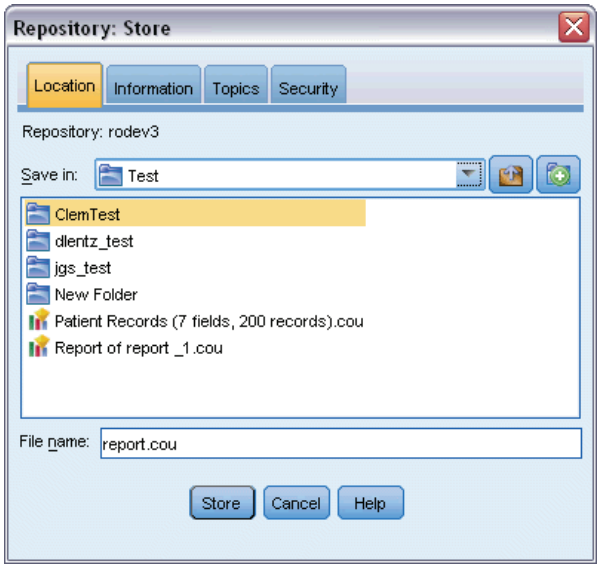
▶ Treffen Sie im Ausgabeobjektfenster folgende Auswahl:
Datei > Im Web veröffentlichen

Anmerkung: Sollen einfache HTML-Dateien zur Verwendung mit einem Standard-Webbrowser exportiert werden, wählen Sie Exportieren aus dem Menü “Datei” und dann HTML.

▶ Bauen Sie eine Verbindung zum IBM® SPSS® Collaboration and Deployment Services Repository auf.

Wenn die Verbindung erfolgreich aufgebaut wurde, wird das Dialogfeld “Repository: Speichern” angezeigt, in dem Sie zwischen verschiedenen Speicheroptionen wählen können.

Abbildung 6-2
Dialogfeld “Repository: speichern”



▶ Wenn Sie die gewünschten Speicheroptionen ausgewählten haben, klicken Sie auf Speichern.

Anzeigen von veröffentlichten Ausgabedaten im Web

Für die Verwendung dieser Funktion muss ein IBM SPSS Collaboration and Deployment Services-Konto eingerichtet sein. Wenn die entsprechende Anwendung für den anzuzeigenden Objekttyp installiert ist (z. B. IBM® SPSS® Modeler oder IBM® SPSS® Statistics), wird die Ausgabe in der Anwendung, nicht im Browser angezeigt.

*Anmerkung:*Für den Zugriff auf IBM® SPSS® Collaboration and Deployment Services ist eine separate Lizenz erforderlich. Weitere Informationen finden Sie unter *http://www.ibm.com/software/analytics/spss/products/deployment/cds/*.

So zeigen Sie veröffentlicht Ausgabeobjekte im Web an:

- Geben Sie in Ihrem Browser die Adresse *http://<repos_host>:<repos_port>/peb* ein.

  Dabei bezeichnen *repos_host* und *repos_port* den Hostnamen bzw. die Portnummer für den IBM SPSS Collaboration and Deployment Services-Host.

- Geben Sie die Anmeldedaten für Ihr IBM SPSS Collaboration and Deployment Services-Konto ein.

- Klicken Sie auf Inhalt Repository.

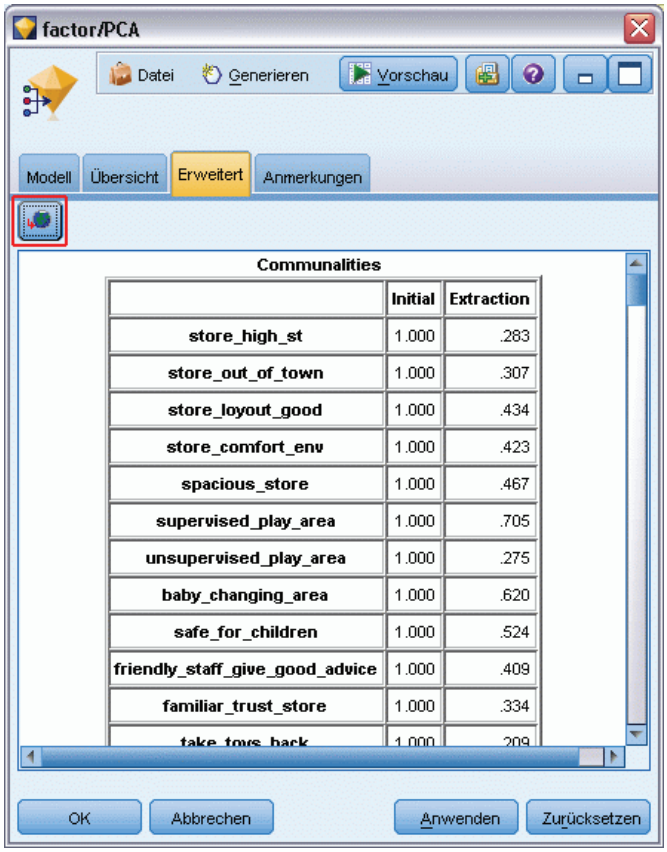
- Navigieren Sie zu dem Objekt, das Sie anzeigen möchten, oder suchen Sie nach dem Objekt.

- Klicken Sie auf den Objektnamen. Für einige Objekttypen wie z. B. Diagramme tritt eventuell eine Verzögerung ein, während das Objekt im Browser gerendert wird.

Anzeigen der Ausgabe in einem HTML-Browser

Auf der Registerkarte “Erweitert” der Modell-Nuggets “Lineare Regression”, “Logistische Regression” und “Faktor” können Sie die angezeigten Informationen in einem separaten Browser öffnen, beispielsweise im Internet Explorer. Die Informationen werden als HTML ausgegeben, was es ermöglicht, sie zu speichern und an anderer Stelle wiederzuverwenden, beispielsweise in einem unternehmenseigenen Intranet oder auf einer Internet-Site.

Abbildung 6-3
Schaltfläche "Starten" in der Registerkarte "Erweitert" des Modell-Nuggets



	Initial	Extraction
store_high_st	1.000	.283
store_out_of_town	1.000	.307
store_layout_good	1.000	.434
store_comfort_env	1.000	.423
spacious_store	1.000	.467
supervised_play_area	1.000	.705
unsupervised_play_area	1.000	.275
baby_changing_area	1.000	.620
safe_for_children	1.000	.524
friendly_staff_give_good_advice	1.000	.409
familiar_trust_store	1.000	.334
take_toys_back	1.000	.209

Um die Informationen in einem Browser anzuzeigen, klicken Sie auf die Startschaltfläche unterhalb des Modellsymbols links oben auf der Registerkarte "Erweitert" des Modell-Nuggets.

Exportieren von Ausgaben

Im Ausgabe-Browser-Fenster können Sie auswählen, dass die Ausgabe in einem anderen Format, beispielsweise Text oder HTML, exportiert werden soll. Die Exportformate variieren je nach Ausgabetyp, im Allgemeinen sind ähneln sie jedoch den Optionen für den Dateityp, die verfügbar sind, wenn Sie in dem zur Generierung der Ausgabe verwendeten Knoten die Option In Datei speichern auswählen.

So exportieren Sie Ausgaben:

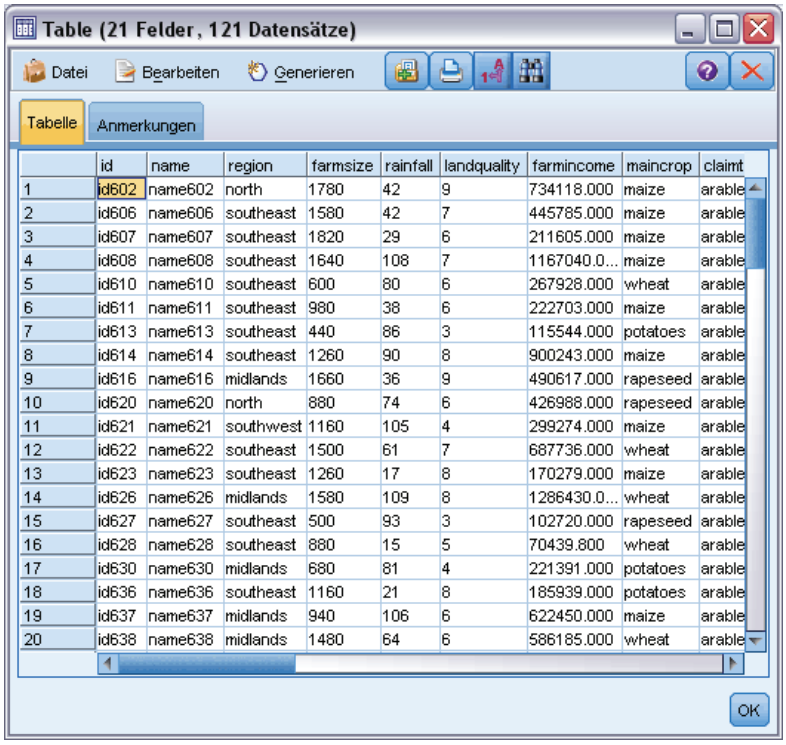
▶ Öffnen Sie im Ausgabe-Browser das Menü "Datei" und wählen Sie die Option Exportieren. Wählen Sie anschließend den zu erstellenden Dateityp:

- **Tabulator-getrennt (*.tab).** Mit dieser Option erzeugen Sie eine formatierte Textdatei mit den Datenwerten. Dieser Stil eignet sich häufig für das Erzeugen einer Textdarstellung der Daten, die dann in andere Anwendungen importiert werden kann. Diese Option ist für Tabellen-, Matrix- und Mittelwertknoten verfügbar.

- **Komma getrennt (*.dat).** Mit dieser Option erzeugen Sie eine Textdatei mit den Datenwerten; diese Werte sind durch Komma voneinander getrennt. Dieser Stil eignet sich häufig für die rasche Erzeugung einer Datendatei, die in Tabellenkalkulationen oder andere Anwendungen zur Datenanalyse importiert werden kann. Diese Option ist für Tabellen-, Matrix- und Mittelwertknoten verfügbar.
- **Transponiert Tabulator getrennt (*.tab).** Diese Option ist mit der Option "Tabulator getrennt" identisch, die Daten werden jedoch transponiert, sodass die Zeilen Felder und die Spalten Datensätze darstellen.
- **Transponiert Komma getrennt (*.tab)** Diese Option ist mit der Option "Komma getrennt" identisch, die Daten werden jedoch transponiert, sodass die Zeilen Felder und die Spalten Datensätze darstellen.
- **HTML (*.html).** Mit dieser Option wird die Ausgabe im HTML-Format in eine oder mehrere Dateien geschrieben.

Auswählen von Zellen und Spalten

Abbildung 6-4
Fenster des Tabellen-Browsers



	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimt
1	id602	name602	north	1780	42	9	734118.000	maize	arable
2	id606	name606	southeast	1580	42	7	445785.000	maize	arable
3	id607	name607	southeast	1820	29	6	211605.000	maize	arable
4	id608	name608	southeast	1640	108	7	1167040.0...	maize	arable
5	id610	name610	southeast	600	80	6	267928.000	wheat	arable
6	id611	name611	southeast	980	38	6	222703.000	maize	arable
7	id613	name613	southeast	440	86	3	115544.000	potatoes	arable
8	id614	name614	southeast	1260	90	8	900243.000	maize	arable
9	id616	name616	midlands	1660	36	9	490617.000	rapeseed	arable
10	id620	name620	north	880	74	6	426988.000	rapeseed	arable
11	id621	name621	southwest	1160	105	4	299274.000	maize	arable
12	id622	name622	southeast	1500	61	7	687736.000	wheat	arable
13	id623	name623	southeast	1260	17	8	170279.000	maize	arable
14	id626	name626	midlands	1580	109	8	1286430.0...	wheat	arable
15	id627	name627	southeast	500	93	3	102720.000	rapeseed	arable
16	id628	name628	southeast	880	15	5	70439.800	wheat	arable
17	id630	name630	midlands	680	81	4	221391.000	potatoes	arable
18	id636	name636	southeast	1160	21	8	185939.000	potatoes	arable
19	id637	name637	midlands	940	106	6	622450.000	maize	arable
20	id638	name638	midlands	1480	64	6	586185.000	wheat	arable

Eine Reihe von Knoten, darunter der Tabellenknoten, der Matrixknoten und der Mittelwertknoten generieren eine Tabellenausgabe. Diese Ausgabetabellen können auf ähnliche Weise angezeigt und bearbeitet werden. Zu den Bearbeitungsmöglichkeiten gehören die Aufnahme ausgewählter Zellen, das Kopieren der gesamten Tabelle oder von Teilen davon in die Zwischenablage, das Erstellen neuer Knoten auf der Grundlage der aktuellen Auswahl sowie das Speichern und Drucken der Tabelle.

Auswählen von Zellen. Um eine Zelle auszuwählen, klicken Sie darauf. Soll ein rechteckiger Zellbereich ausgewählt werden, klicken Sie auf eine Ecke des gewünschten Bereichs. Halten Sie die Maustaste gedrückt, ziehen Sie die Maus auf die diagonal gegenüberliegende Ecke des Bereichs und lösen Sie die Maustaste wieder. Um eine ganze Spalte auszuwählen, klicken Sie auf die Spaltenüberschrift. Mit Umschalt-Klicken bzw. Strg-Klicken auf Spaltenüberschriften können Sie mehrere Spalten gleichzeitig auswählen.

Sobald Sie eine neue Auswahl treffen, wird die bisherige Auswahl wieder aufgehoben. Wenn Sie die Strg-Taste beim Auswählen gedrückt halten, wird jedoch nicht die bisherige Auswahl aufgehoben, sondern die neue Auswahl wird zur vorhandenen Auswahl hinzugefügt. Auf diese Weise können Sie mehrere, nicht zusammenhängende Bereiche der Tabelle auswählen. Das Menü "Bearbeiten" enthält außerdem die Optionen Alles auswählen und Auswahl aufheben.

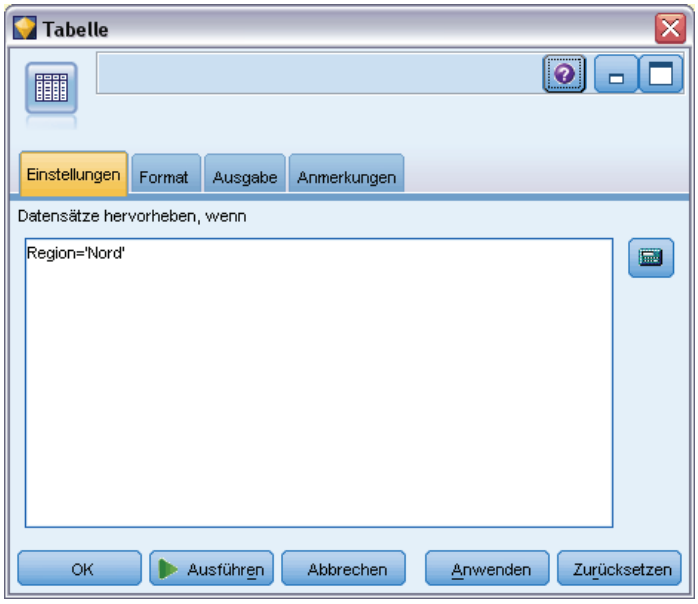
Ordnen von Spalten Mit den Ausgabe-Browsern des Tabellenknotens und des Mittelwertknotens können Sie Spalten in der Tabelle verschieben. Klicken Sie hierzu auf eine Spaltenüberschrift und ziehen Sie sie an die gewünschte Position. Es ist nicht möglich, mehrere Spalten gleichzeitig zu verschieben.

Tabellenknoten

Der Tabellenknoten erstellt eine Tabelle mit den in Ihren Daten enthaltenen Werten. Da alle Felder und alle Werte im Stream enthalten sind, können hiermit Datenwerte schnell und einfach überprüft oder in leicht lesbarer Form exportiert werden. Wahlweise können Sie auch Datensätze hervorheben, die eine bestimmte Bedingung erfüllen.

Registerkarte "Einstellungen" beim Tabellenknoten

Abbildung 6-5
Tabellenknoten: Registerkarte "Einstellungen"



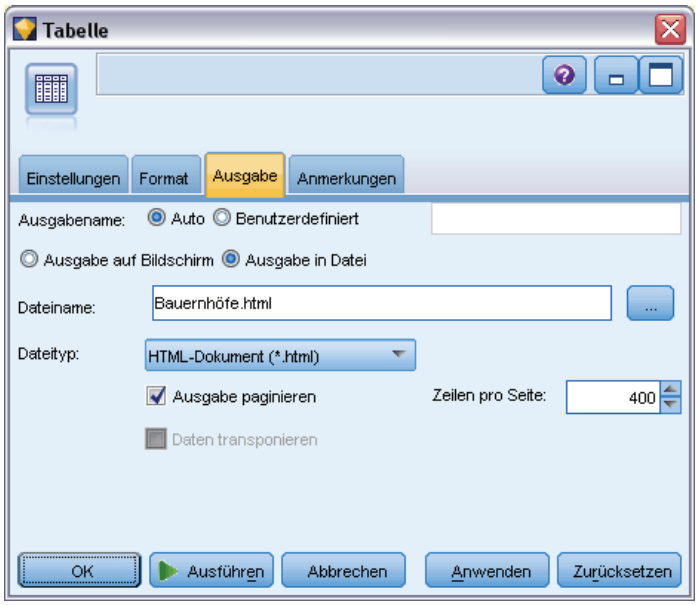
Datensätze hervorheben, wenn. Sie können Datensätze in der Tabelle hervorheben, indem Sie einen CLEM-Ausdruck eingeben, der für die betreffenden Datensätze wahr ist. Diese Option ist nur dann aktiviert, wenn Sie die Option Ausgabe auf Bildschirm aktiviert haben.

Registerkarte "Format" beim Tabellenknoten

Die Registerkarte "Format" enthält Optionen, mit denen Sie die Formatierung für die einzelnen Felder festlegen. Diese Registerkarte wird auch beim Typknoten verwendet. Für weitere Informationen siehe Thema Feldformat – Registerkarte "Einstellungen" auf S. 153.

Registerkarte "Ausgabe" beim Ausgabeknoten

Abbildung 6-6
Registerkarte "Ausgabe" beim Ausgabeknoten



Bei Knoten, die Ausgaben in Tabellenform generieren, können Sie mithilfe der Registerkarte "Ausgabe" Format und Standort der Ergebnisse angeben.

Ausgabename. Bestimmt den Namen der Ausgabe, die beim Ausführen des Knotens erstellt wird. Mit Auto wird ein Name auf der Grundlage des Knotens bestimmt, mit dem die Ausgabe erzeugt wird. Optional können Sie auch Angepasst auswählen und einen anderen Namen angeben.

Ausgabe auf Bildschirm (Standardeinstellung)**.** Erstellt ein Ausgabeobjekt für die Online-Anzeige. Das Ausgabeobjekt wird auf der Registerkarte "Ausgaben" im Manager-Fenster dargestellt, wenn der Ausgabeknoten ausgeführt wird.

Ausgabe in Datei. Speichert die Ausgabe in einer Datei, wenn der Knoten ausgeführt wird. Wenn Sie diese Option wählen, geben Sie einen Dateinamen an (oder wechseln Sie zu einem Verzeichnis und geben Sie einen Dateinamen mithilfe der Feldauswahl-Schaltfläche an) und wählen Sie

einen Dateityp aus. Beachten Sie, dass einige Dateitypen möglicherweise nicht für bestimmte Ausgabetypen verfügbar sind.

Daten werden im Standardkodierungsformat des Systems ausgegeben, das in der Windows-Systemsteuerung bzw. bei Ausführung im verteilten Modus auf dem Server-Computer angegeben wird.

- **Tabulator getrennte Daten (*.tab).** Mit dieser Option erzeugen Sie eine formatierte Textdatei mit den Datenwerten. Dieser Stil eignet sich häufig für das Erzeugen einer Textdarstellung der Daten, die dann in andere Anwendungen importiert werden kann. Diese Option ist für Tabellen-, Matrix- und Mittelwertknoten verfügbar.
- **Daten (kommagetrennt) (.dat).** Mit dieser Option erzeugen Sie eine Textdatei mit den Datenwerten; diese Werte sind durch Komma voneinander getrennt. Dieser Stil eignet sich häufig für die rasche Erzeugung einer Datendatei, die in Tabellenkalkulationen oder andere Anwendungen zur Datenanalyse importiert werden kann. Diese Option ist für Tabellen-, Matrix- und Mittelwertknoten verfügbar.
- **HTML (*.html).** Mit dieser Option wird die Ausgabe im HTML-Format in eine oder mehrere Dateien geschrieben. Bei Tabellenausgaben (aus dem Tabellen-, Matrix- oder Mittelwertknoten) enthält eine Reihe von HTML-Dateien einen Inhaltsbereich, in dem die Feldnamen aufgeführt werden; die Daten befinden sich in einer HTML-Tabelle. Die Tabelle wird ggf. auf mehrere HTML-Dateien aufgeteilt, wenn die Anzahl der Zeilen in der Tabelle die Angaben unter Zeilen pro Seite überschreitet. In diesem Fall enthält der Inhaltsbereich Links zu allen Tabellenseiten und dient als Mittel zur Navigation in der Tabelle. Bei einer nicht tabellenförmigen Ausgabe wird eine einzige HTML-Datei mit den Ergebnissen des Knotens erzeugt.

  Hinweis: Falls die HTML-Ausgabe nur die Formatierung für die erste Seite enthält, wählen Sie die Option Ausgabe paginieren und passen Sie die Angaben unter Zeilen pro Seite an, sodass die gesamte Ausgabe auf einer einzigen Seite erfolgt. Falls die Ausgabevorlage für die Knoten (z. B. für den Berichtknoten) benutzerdefinierte HTML-Tags enthält, können Sie alternativ den Formattyp Benutzerdefiniert auswählen.
- **Textdatei (*.txt).** Mit dieser Option erzeugen Sie eine Textdatei mit der Ausgabe. Dieser Stil eignet sich häufig zum Erzeugen einer Ausgabe, die dann in andere Anwendungen importiert werden kann, z. B. in eine Textverarbeitung oder in Präsentations-Software. Diese Option ist für einige Knoten nicht verfügbar.
- **Ausgabeobjekt (*.cou).** Die in diesem Format gespeicherten Ausgabeobjekte können in IBM® SPSS® Modeler geöffnet und angezeigt, zu Projekten hinzugefügt sowie mit demIBM® SPSS® Collaboration and Deployment Services Repository veröffentlicht und verfolgt werden.

Ausgabeansicht. Für den Mittelwertknoten können Sie angeben, ob standardmäßig eine einfache oder eine erweiterte Ausgabe angezeigt werden soll. Beachten Sie, dass Sie auch zwischen diesen beiden Ansichten umschalten können, während Sie die generierte Ausgabe durchsuchen. Für weitere Informationen siehe Thema Mittelwertknoten – Ausgabe-Browser auf S. 450.

Format. Beim Berichtknoten können Sie auswählen, ob die Ausgabe automatisch formatiert oder mit HTML aus der Vorlage formatiert werden soll. Mit der Option Angepasst ermöglichen Sie die HTML-Formatierung in der Vorlage.

Titel. Beim Berichtknoten können Sie optional einen Titeltext angeben, der oben in der Berichtausgabe eingefügt werden soll.

Eingefügten Text hervorheben. Beim Berichtknoten lassen Sie mit dieser Option den Text hervorheben, der durch CLEM-Ausdrücke in der Berichtvorlage erzeugt wurde. Für weitere Informationen siehe Thema Registerkarte "Vorlage" beim Berichtknoten auf S. 454. Diese Option wird nicht empfohlen, wenn Sie die Formatierung Angepasst verwenden.

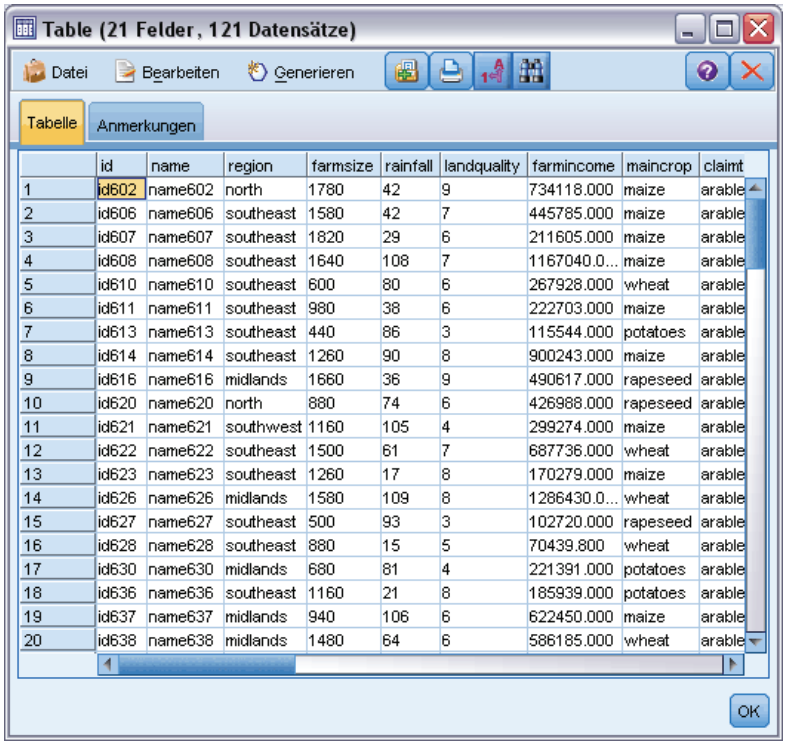
Zeilen pro Seite. Geben Sie beim Berichtknoten die Anzahl der Zeilen an, die bei der Formatierung des Ausgabeberichts mit der Option Auto auf jeder Seite untergebracht werden sollen.

Daten transponieren. Mit dieser Option werden die Daten vor dem Export transponiert, sodass die Zeilen Felder und die Spalten Datensätze darstellen.

Hinweis: Bei umfangreichen Tabellen sind die obigen Optionen eher unrationell; dies gilt insbesondere dann, wenn Sie mit einem Remote-Server arbeiten. In solchen Fällen liefert ein Dateiausgabeknoten deutlich bessere Leistungen. Für weitere Informationen siehe Thema Textdatei-Exportknoten in Kapitel 7 auf S. 482.

Tabellen-Browser

Abbildung 6-7
Fenster des Tabellen-Browsers



	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimt
1	id602	name602	north	1780	42	9	734118.000	maize	arable
2	id606	name606	southeast	1580	42	7	445785.000	maize	arable
3	id607	name607	southeast	1820	29	6	211605.000	maize	arable
4	id608	name608	southeast	1640	108	7	1167040.0...	maize	arable
5	id610	name610	southeast	600	80	6	267928.000	wheat	arable
6	id611	name611	southeast	980	38	6	222703.000	maize	arable
7	id613	name613	southeast	440	86	3	115544.000	potatoes	arable
8	id614	name614	southeast	1260	90	8	900243.000	maize	arable
9	id616	name616	midlands	1660	36	9	490617.000	rapeseed	arable
10	id620	name620	north	880	74	6	426988.000	rapeseed	arable
11	id621	name621	southwest	1160	105	4	299274.000	maize	arable
12	id622	name622	southeast	1500	61	7	687736.000	wheat	arable
13	id623	name623	southeast	1260	17	8	170279.000	maize	arable
14	id626	name626	midlands	1580	109	8	1286430.0...	wheat	arable
15	id627	name627	southeast	500	93	3	102720.000	rapeseed	arable
16	id628	name628	southeast	880	15	5	70439.800	wheat	arable
17	id630	name630	midlands	680	81	4	221391.000	potatoes	arable
18	id636	name636	southeast	1160	21	8	185939.000	potatoes	arable
19	id637	name637	midlands	940	106	6	622450.000	maize	arable
20	id638	name638	midlands	1480	64	6	586185.000	wheat	arable

Der Tabellen-Browser zeigt Daten in Tabellenform an und ermöglicht Ihnen die Durchführung von Standardoperationen: beispielsweise Auswahl und das Kopieren von Zellen, Neuordnen von Spalten und Speichern und Drucken der Tabelle. Für weitere Informationen siehe

Thema Auswählen von Zellen und Spalten auf S. 404.Dabei handelt es sich um die gleichen Operationen, die Sie bei der Vorschau der Daten in einem Knoten ausführen können.

Exportieren von Tabellendaten. Sie können Daten aus dem Tabellen-Browser über folgende Optionsfolge exportieren:

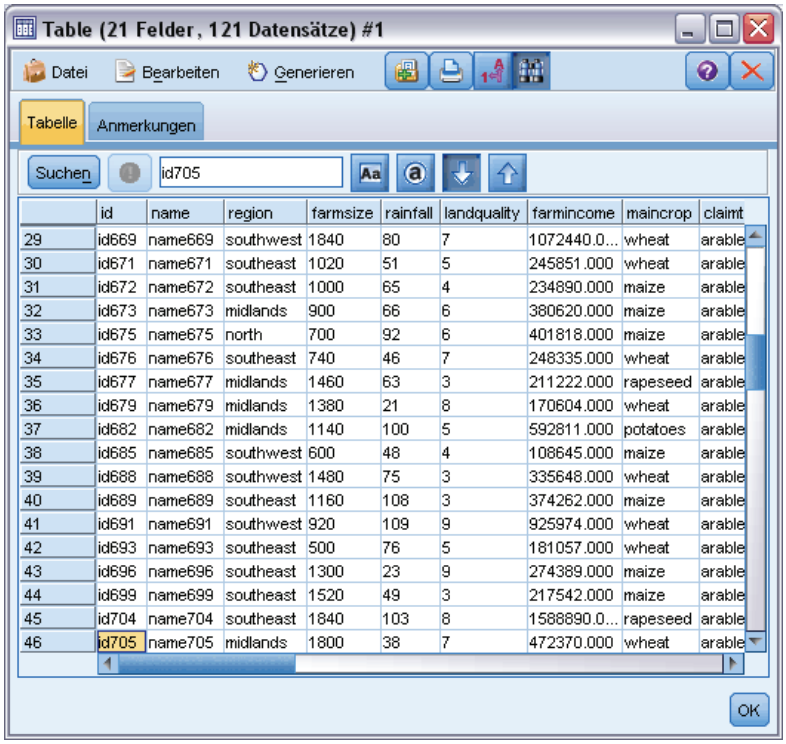
Datei > Exportieren

Für weitere Informationen siehe Thema Exportieren von Ausgaben auf S. 403.

Daten werden im Standardkodierungsformat des Systems exportiert, das in der Windows-Systemsteuerung bzw. bei Ausführung im verteilten Modus auf dem Server-Computer angegeben wird.

Durchsuchen der Tabelle. Mit der Schaltfläche “Suche” (das Fernglassymbol) in der Hauptsymbolleiste aktivieren Sie die Such-Symbolleiste, um so bestimmte Werte in der Tabelle zu suchen. Sie können vorwärts oder rückwärts in der Tabelle suchen, die Suche unter Beachtung der Groß-/Kleinschreibung starten (Schaltfläche Aa) sowie einen laufenden Suchvorgang mit der Schaltfläche zum Unterbrechen der Suche anhalten.

Abbildung 6-8
Tabelle mit aktivierten Steuerelementen für Suchvorgänge



	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimt
29	id669	name669	southwest	1840	80	7	1072440.0...	wheat	arable
30	id671	name671	southeast	1020	51	5	245851.000	wheat	arable
31	id672	name672	southeast	1000	65	4	234890.000	maize	arable
32	id673	name673	midlands	900	66	6	380620.000	maize	arable
33	id675	name675	north	700	92	6	401818.000	maize	arable
34	id676	name676	southeast	740	46	7	248335.000	wheat	arable
35	id677	name677	midlands	1460	63	3	211222.000	rapeseed	arable
36	id679	name679	midlands	1380	21	8	170604.000	wheat	arable
37	id682	name682	midlands	1140	100	5	592811.000	potatoes	arable
38	id685	name685	southwest	600	48	4	108645.000	maize	arable
39	id688	name688	southwest	1480	75	3	335648.000	wheat	arable
40	id689	name689	southeast	1160	108	3	374262.000	maize	arable
41	id691	name691	southwest	920	109	9	925974.000	wheat	arable
42	id693	name693	southeast	500	76	5	181057.000	wheat	arable
43	id696	name696	southeast	1300	23	9	274389.000	maize	arable
44	id699	name699	southeast	1520	49	3	217542.000	maize	arable
45	id704	name704	southeast	1840	103	8	1588890.0...	rapeseed	arable
46	id705	name705	midlands	1800	38	7	472370.000	wheat	arable

Erzeugen neuer Knoten. Das Menü “Generieren” enthält Funktionen zum Erzeugen von Knoten.

- **Auswahlknoten (“Datensätze”).** Erzeugt einen Auswahlknoten, mit dem die Datensätze ausgewählt werden, für die eine beliebige Zelle in der Tabelle markiert ist.
- **Auswahlknoten (“UND”).** Erzeugt einen Auswahlknoten, mit dem die Datensätze ausgewählt werden, die *alle* in der Tabelle markierten Werte enthält.

- **Auswahlknoten ("ODER").** Erzeugt einen Auswahlknoten, mit dem die Datensätze ausgewählt werden, die *einen* der in der Tabelle markierten Werte enthält.
- **Ableitungsknoten ("Datensätze").** Erzeugt einen Ableitungsknoten, mit dem ein neues Flag-Feld erstellt wird. Das Flag-Feld besitzt den Wert *T* für Datensätze, für die eine beliebige Zelle in der Tabelle ausgewählt ist, bzw. den Wert *F* für die verbleibenden Datensätze.
- **Ableitungsknoten ("UND").** Erzeugt einen Ableitungsknoten, mit dem ein neues Flag-Feld erstellt wird. Das Flag-Feld besitzt den Wert *T* für Datensätze, die *alle* in der Tabelle ausgewählten Werte enthalten, bzw. *F* für die verbleibenden Datensätze.
- **Ableitungsknoten ("ODER").** Erzeugt einen Ableitungsknoten, mit dem ein neues Flag-Feld erstellt wird. Das Flag-Feld besitzt den Wert *T* für Datensätze, die *einen* in der Tabelle ausgewählten Wert enthalten, bzw. *F* für die verbleibenden Datensätze.

Matrixknoten

Mit dem Matrixknoten können Sie eine Tabelle erstellen, die die Beziehungen zwischen den Feldern aufzeigt. Dieser Knoten dient am häufigsten zum Darstellen der Beziehung zwischen zwei kategorialen Feldern (Flag, nominal oder ordinal), kann jedoch auch zum Aufzeigen der Beziehungen zwischen stetigen Feldern (numerischer Bereich) herangezogen werden.

Registerkarte "Einstellungen" beim Matrixknoten

Auf der Registerkarte "Einstellungen" legen Sie Optionen für die Struktur der Matrix fest.

Abbildung 6-9
Matrixknoten: Registerkarte "Einstellungen"

Felder. Wählen Sie einen Feldauswahltyp aus den folgenden Optionen aus:

- **Ausgewählt.** Bei dieser Option können Sie je ein kategoriales Feld für die Zeilen und Spalten in der Matrix auswählen. Die Zeilen und Spalten der Matrix werden durch die Liste der Werte für das ausgewählte kategoriale Feld bestimmt. Die Zellen der Matrix enthalten die unten ausgewählten Übersichtsstatistiken.
- **Alle Flags (wahre Werte).** Bei dieser Option wird eine Matrix mit je einer Zeile und einer Spalte für jedes Flag-Feld in den Daten angefordert. Die Zellen der Matrix enthalten die Anzahl der doppeltpositiven Werte für jede Flag-Kombination. Beispiel: Bei einer Zeile für *Brot gekauft* und einer Spalte für *Käse gekauft* enthält die Zelle am Schnittpunkt dieser Zeile und Spalte die Anzahl der Datensätze, bei denen sowohl *Brot gekauft* als auch *Käse gekauft* wahr sind.
- **Alle numerischen Werte.** Bei dieser Option wird eine Matrix mit je einer Zeile und einer Spalte für jedes numerische Feld angefordert. Die Zellen der Matrix stellen die Summe der Kreuzprodukte für das entsprechende Feldpaar dar. Für jede Zelle in der Matrix werden also die Werte aus dem Zeilenfeld und dem Spaltenfeld für jeden Datensatz multipliziert und dann über die Datensätze hinweg summiert.

Fehlende Werte einschließen. Schließt benutzerdefiniert fehlende Werte (leer) und systemdefiniert fehlende Werte ($null$) in die Zeilen- und Spaltenausgabe ein. Wenn beispielsweise der Wert *N/A* für das ausgewählte Spaltenfeld als benutzerdefiniert fehlend definiert wurde, wird eine gesonderte Spalte mit der Beschriftung *N/A* wie jede andere Kategorie in die Tabelle aufgenommen (vorausgesetzt, dieser Wert kommt tatsächlich in den Daten vor). Wenn die Auswahl dieser Option aufgehoben wird, wird die Spalte *k. A* ausgeschlossen, egal wie oft sie vorkommt.

Anmerkung: Die Option zur Aufnahme fehlender Werte gilt nur, wenn die ausgewählten Felder als Kreuztabelle vorliegen. Leere Werte werden $null$ zugeordnet und aus der Aggregation für das Funktionsfeld ausgeschlossen, wenn der Modus Ausgewählt ist und der Inhalt auf Funktion gesetzt ist. Der Ausschluss für alle numerischen Felder erfolgt, wenn der Modus auf Alle numerischen Werte gesetzt ist.

Zelleninhalte. Wenn Sie oben die Option Ausgewählte Felder aktiviert haben, können Sie die Statistik angeben, die in den Zellen der Matrix verwendet werden soll. Wählen Sie eine anzahlbasierte Statistik aus oder wählen Sie ein Überlagerungsfeld aus, mit dem die Werte aus einem numerischen Feld auf der Grundlage der Werte der Zeilen- und Spaltenfelder zusammengefasst werden.

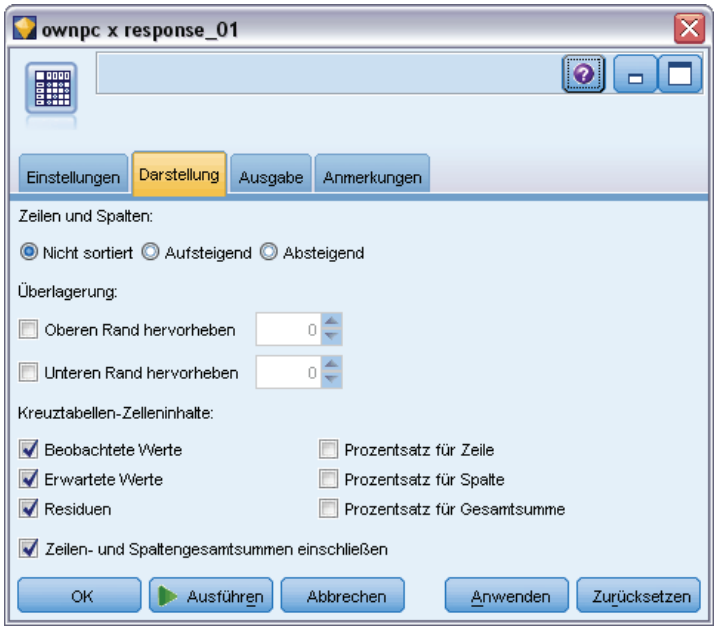
- **Kreuztabellen.** Die Zellenwerte bestehen aus der Anzahl und/oder dem Prozentsatz der Datensätze, die die entsprechende Wertekombination aufweisen. Mit den Optionen auf der Registerkarte “Darstellung” können Sie die gewünschten Kreuztabellenübersichten auswählen. Der globale Chi-Quadrat-Wert wird ebenso zusammen mit der Signifikanz angezeigt. Für weitere Informationen siehe Thema Matrixknoten – Ausgabe-Browser auf S. 413.
- **Funktion.** Wenn Sie eine Übersichtsfunktion auswählen, bilden die Zellenwerte eine Funktion der Werte aus dem ausgewählten Überlagerungsfeld für Fälle, die die entsprechenden Zeilen- und Spaltenwerte besitzen. Beispiel: Sie verwenden das Zeilenfeld *Region*, das Spaltenfeld *Produkt* und das Überlagerungsfeld *Einkommen*. In diesem Fall enthält die Zelle in der Zeile *Nordosten* und der Spalte *Dings* die Summe (bzw. den Durchschnitt, den Mindestwert oder den Höchstwert) für den Umsatz aus Geräten, die in der Region Nordost verkauft wurden. Die Standardeinstellung für die Übersichtsfunktion lautet Mittelwert. Sie können eine andere

Funktion auswählen, mit der das Funktionsfeld zusammengefasst werden soll. Die folgenden Optionen stehen zur Auswahl: Mittelwert, Summe, Std.Abw. (Standardabweichung), Max (Maximum) sowie Min (Minimum).

Registerkarte "Darstellung" beim Matrixknoten

Auf der Registerkarte "Darstellung" steuern Sie die Optionen zum Sortieren und Hervorheben für die Matrix, außerdem die Statistiken, die für Kreuztabellenmatrizen angezeigt werden.

Abbildung 6-10
Matrixknoten: Darstellung, Registerkarte



Zeilen und Spalten. Steuert die Sortierung der Zeilen- und Spaltenüberschriften in der Matrix. Die Standardeinstellung lautet Nicht sortiert. Mit der Option Aufsteigend oder Absteigend lassen Sie die Zeilen- und Spaltenüberschrift in die angegebene Richtung sortieren.

Überlagerung. Ermöglicht das Hervorheben von Extremwerten in der Matrix. Die Werte werden auf der Grundlage der Zellenanzahl (bei Kreuztabellenmatrizen) oder der berechneten Werte (bei Funktionsmatrizen) hervorgehoben.

- **Oberen Rand hervorheben.** Sie können die höchsten Werte in der Matrix hervorheben lassen (in Rot). Geben Sie die Anzahl der hervorzuhebenden Werte an.
- **Unteren Rand hervorheben.** Sie können auch die niedrigsten Werte in der Matrix hervorheben lassen (in Grün). Geben Sie die Anzahl der hervorzuhebenden Werte an.

Hinweis: Bei den beiden Hervorhebungsoptionen können Bindungen dazu führen, dass mehr Werte als angefordert hervorgehoben werden. Wenn Sie beispielsweise eine Matrix mit sechs Nullwerten in den Zellen verwenden und die Option Unteren Rand hervorheben mit dem Wert 5 auswählen, werden alle sechs Nullwerte hervorgehoben.

Kreuztabellen-Zelleninhalte. Bei Kreuztabellen können Sie die Übersichtsstatistiken in der Matrix für Kreuztabellenmatrizen angeben. Diese Optionen sind nicht verfügbar, wenn die Option Alle numerischen Werte oder Funktion auf der Registerkarte "Einstellungen" ausgewählt wurde.

- **Häufigkeiten.** Die Zellen umfassen die Anzahl der Datensätze mit dem Zeilenwert, die auch den zugehörigen Spaltenwert aufweisen. Dies gilt nur für den standardmäßigen Zelleninhalt.
- **Erwartete Werte.** Der erwartete Wert für die Anzahl der Datensätze in der Zelle, unter der Annahme, dass keine Beziehung zwischen Zeilen und Spalten besteht. Die erwarteten Werte beruhen auf der folgenden Formel:

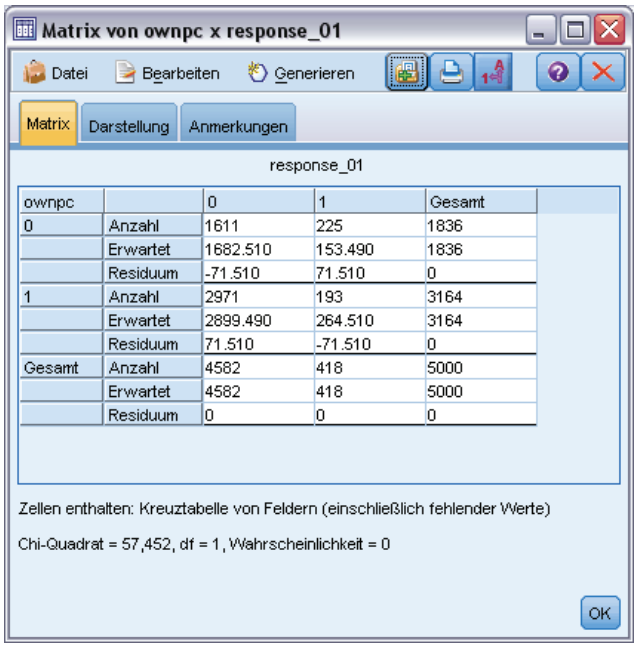
  p(*Zeilenwert*) * p(*Spaltenwert*) * *Gesamtanzahl der Datensätze*

- **Residuen.** Die Differenz zwischen beobachteten und erwarteten Werten.
- **Prozentsatz für Zeile.** Der Prozentsatz aller Datensätze mit dem Zeilenwert, die auch den zugehörigen Spaltenwert aufweisen. Die Prozentsätze für die Zeilen ergeben insgesamt den Wert 100.
- **Prozentsatz für Spalte.** Der Prozentsatz aller Datensätze mit dem Spaltenwert, die auch den zugehörigen Zeilenwert aufweisen. Die Prozentsätze für die Spalten ergeben insgesamt den Wert 100.
- **Prozentsatz für Gesamtsumme.** Der Prozentsatz aller Datensätze, die die angegebene Kombination aus Spaltenwert und Zeilenwert aufweisen. Die Prozentsätze in der gesamten Matrix ergeben insgesamt den Wert 100.
- **Zeilen- und Spaltengesamtsummen einschließen.** Fügt eine Reihe und eine Spalte zur Matrix hinzu, in denen die Gesamtsummen der Zeilen und Spalten eingetragen werden.
- **Einstellungen anwenden.** (nur Ausgabe-Browser) Ermöglicht die Vornahme von Änderungen am Erscheinungsbild der Ausgabe des Matrixknotens, ohne dass der Ausgabe-Browser geschlossen und erneut geöffnet werden muss. Nehmen Sie die Änderungen auf dieser Registerkarte des Ausgabe-Browsers vor, klicken Sie auf diese Schaltfläche und wählen Sie dann die Registerkarte "Matrix", um die Auswirkungen der Änderungen anzuzeigen.

Matrixknoten – Ausgabe-Browser

Im Matrix-Browser werden Kreuztabellendaten angezeigt. Hier können Sie verschiedene Vorgänge für die Matrix ausführen, z. B. Zellen auswählen, Matrix ganz oder teilweise in die Zwischenablage kopieren, neue Knoten auf der Grundlage der Auswahl in der Matrix erzeugen sowie die Matrix speichern und drucken. Mit dem Matrix-Browser können Sie außerdem die Ausgabe bestimmter Modelle anzeigen lassen, z. B. Naive Bayes-Modelle aus Oracle.

Abbildung 6-11
Matrix-Browser



ownpc		0	1	Gesamt
0	Anzahl	1611	225	1836
	Erwartet	1682.510	153.490	1836
	Residuum	-71.510	71.510	0
1	Anzahl	2971	193	3164
	Erwartet	2899.490	264.510	3164
	Residuum	71.510	-71.510	0
Gesamt	Anzahl	4582	418	5000
	Erwartet	4582	418	5000
	Residuum	0	0	0

Die Menüs “Datei” und “Bearbeiten” bieten die üblichen Optionen zum Drucken, Speichern und Exportieren von Ausgaben sowie zum Auswählen und Kopieren von Daten. Für weitere Informationen siehe Thema Anzeigen der Ausgabe auf S. 399.

Chi-Quadrat. Für eine Kreuztabelle zweier kategorialer Felder wird außerdem das globale Pearson’sche Chi-Quadrat unterhalb der Tabelle angezeigt. Dieser Test gibt die Wahrscheinlichkeit an, dass die beiden Felder unabhängig sind. Die Grundlage hierfür ist die Differenz zwischen der beobachteten Anzahl und den Anzahlwerten, die zu erwarten sind, wenn keine Beziehung vorhanden ist. Beispiel: Wenn es keinen Zusammenhang zwischen Kundenzufriedenheit und Geschäftsstandort gibt, sind ähnliche Zufriedenheitsquoten in allen Geschäften zu erwarten. Wenn jedoch die Kunden in bestimmten Geschäften durchgängig höhere Zufriedenheitsquoten aufweisen als andere, ist zu vermuten, dass es sich nicht um einen Zufall handelt. Je größer die Differenz, desto kleiner ist die Wahrscheinlichkeit, dass es sich lediglich um die Folge eines Fehlers bei der Zufallsstichprobennahme handelt.

- Der Chi-Quadrat-Test gibt die Wahrscheinlichkeit an, dass die beiden Felder unabhängig sind. In diesem Fall sind die Differenzen zwischen beobachteten und erwarteten Häufigkeiten ausschließlich auf den Zufall zurückzuführen. Wenn diese Wahrscheinlichkeit sehr gering ist – üblicherweise unter 5 % – wird die Beziehung zwischen den beiden Feldern als signifikant bezeichnet.
- Wenn es nur eine einzige Spalte oder Zeile gibt (einfacher Chi-Quadrat-Test) ist der Wert für die Freiheitsgrade die Anzahl der Zellen minus 1. Bei einem zweifachen Chi-Quadrat ist der Wert für die Freiheitsgrade gleich der Anzahl der Zeilen minus 1 mal der Anzahl der Spalten minus 1.

- Lassen Sie bei der Interpretation der Chi-Quadrat-Statistik Vorsicht walten, wenn eine der erwarteten Zellenhäufigkeiten unter 5 liegt.
- Der Chi-Quadrat-Test ist nur für eine Kreuztabelle aus zwei Feldern verfügbar. (Wenn Alle Flags oder Alle numerischen Werte auf der Registerkarte "Einstellungen" ausgewählt wurde, wird dieser Test nicht angezeigt.)

Menü "Generieren". Das Menü "Generieren" enthält Funktionen zum Erzeugen von Knoten. Diese Funktionen sind nur bei Kreuztabellenmatrizen verfügbar; in der Matrix muss dabei mindestens eine Zelle ausgewählt sein.

- **Auswahlknoten.** Erzeugt einen Auswahlknoten, mit dem die Datensätze ausgewählt werden, die mit einer beliebigen ausgewählten Zelle in der Matrix übereinstimmen.
- **Ableitungsknoten (Flag).** Erzeugt einen Ableitungsknoten, mit dem ein neues Flag-Feld erstellt wird. Das Flag-Feld besitzt den Wert *T* für Datensätze, die mit einer beliebigen Zelle in der Matrix übereinstimmen, bzw. den Wert *F* für die verbleibenden Datensätze.
- **Ableitungsknoten (Set).** Erzeugt einen Ableitungsknoten, mit dem ein neues nominales Feld erstellt wird. Das nominale Feld enthält je eine Kategorie für jedes zusammenhängende Set ausgewählter Zellen in der Matrix.

Analyseknoten

Mit dem Analyseknoten können Sie die Fähigkeit eines Modells zur Erzeugung genauer Vorhersagen evaluieren. Mit Analyseknoten werden verschiedene Vergleiche zwischen den vorhergesagten Werten und den tatsächlichen Werten (Ihr Zielfeld) für ein oder mehrere Modell-Nuggets angestellt. Analyseknoten können außerdem zum Vergleich von Vorhersagemodellen mit anderen Vorhersagemodellen dienen.

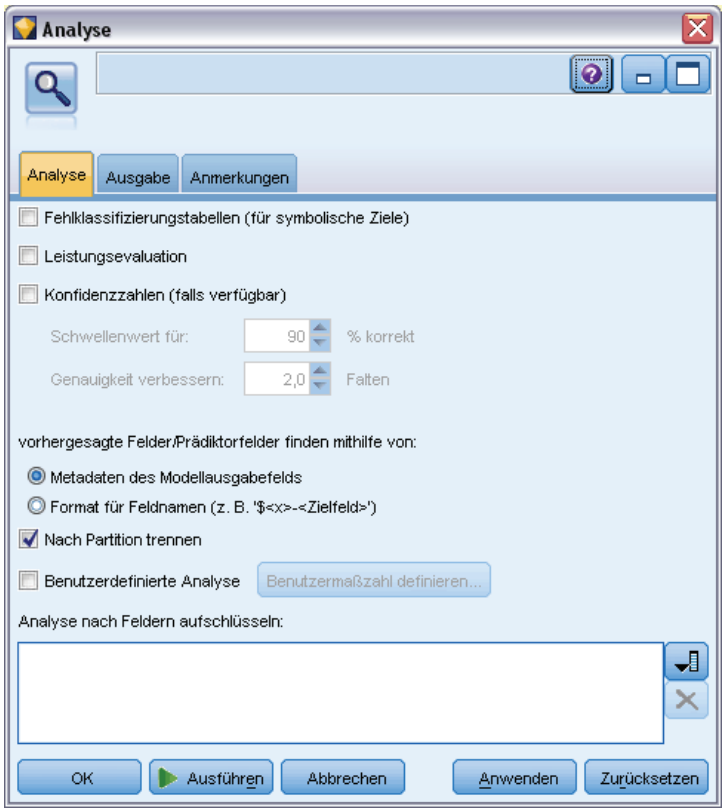
Wenn Sie einen Analyseknoten ausführen, wird auf der Registerkarte "Übersicht" unter "Analyse" automatisch eine Zusammenfassung der Analyseergebnisse für jedes Modell-Nugget im ausgeführten Stream eingetragen. Die ausführlichen Analyseergebnisse werden auf der Registerkarte "Ausgabe" des Manager-Fensters angezeigt oder können direkt in eine Datei geschrieben werden.

Anmerkung: Da Analyseknoten vorhergesagte Felder mit tatsächlichen Werten vergleichen, sind sie nur in überwachten Modellen (Modelle, die ein Zielfeld erfordern) sinnvoll. Bei nicht überwachten Modellen wie beispielsweise Cluster-Algorithmen, stehen keine tatsächlichen Ergebnisse als Grundlage für einen Vergleich zur Verfügung.

Registerkarte "Analyse" beim Analyseknoten

Auf der Registerkarte "Analyse" können Sie die Details für die Analyse angeben.

Abbildung 6-12
Analyseknoten: Analyse, Registerkarte



Fehlklassifizierungstabellen (für symbolische bzw. kategoriale Ziele). Zeigt für kategoriale Ziele (Flag, nominal oder ordinal) das Muster der Übereinstimmungen zwischen jedem erzeugten (vorhergesagten) Feld und dem zugehörigen Zielfeld an. Es wird eine Tabelle eingeblendet, bei der die Zeilen durch tatsächliche Werte definiert sind und die Spalten durch vorhergesagte Werte; hierbei wird die Anzahl der Datensätze ersichtlich, die dieses Muster in den einzelnen Zellen aufweisen. Mit dieser Option können systematische Fehler bei der Vorhersage erkannt werden. Falls mehrere erzeugte Felder zu einem bestimmten Ausgabefeld gehören, jedoch durch unterschiedliche Modelle erstellt wurden, werden die Fälle gezählt, in denen diese Felder übereinstimmen bzw. nicht übereinstimmen, und die Gesamtsummen werden angezeigt. In den Fällen mit Übereinstimmung wird eine weitere Gruppe mit Richtig/Falsch-Statistiken eingeblendet.

Leistungsauswertung. Zeigt die Leistungsauswertungsstatistik für Modelle mit kategorialen Ausgaben. Diese Statistik wird für jede Kategorie des oder der Ausgabefelder erstellt und ist ein Maß für den durchschnittlichen Informationsgehalt (in Bit) des Modells bei der Vorhersage für Datensätze, die zu der betreffenden Kategorie gehören. Hierbei wird die Schwierigkeit des Klassifizierungsproblems berücksichtigt; genaue Vorhersagen für seltene Kategorien erhalten somit einen höheren Leistungsauswertungsindex als genaue Vorhersagen für häufig auftretende Kategorien. Liefert das Modell quasi nur “geratene” Werte für eine Kategorie, erhält diese Kategorie den Leistungsauswertungsindex 0.

Konfidenzzahlen (falls verfügbar). Bei Modellen, bei denen ein Konfidenzfeld erzeugt wird, lassen Sie mit dieser Option die Statistik zu den Konfidenzwerten und deren Beziehung zu den Vorhersagen zusammenstellen. Für diese Option stehen zwei Einstellungen zur Auswahl:

- **Schwellenwert für.** Meldet das Konfidenzniveau, ab dem die Genauigkeit den angegebenen Prozentsatz erreicht.
- **Genauigkeit verbessern.** Meldet das Konfidenzniveau, ab dem die Genauigkeit um den angegebenen Faktor verbessert wird. Beispiel: Die Gesamtgenauigkeit liegt bei 90 % und für diese Option wurde der Wert 2,0 angegeben. Der gemeldete Wert entspricht somit der Konfidenz für eine Genauigkeit von 95 %.

Vorhergesagte Felder/Prädiktorfelder finden mithilfe von. Bestimmt, wie vorhergesagte Felder dem ursprünglichen Zielfeld zugeordnet werden sollen.

- **Metadaten des Modellausgabefelds.** Ordnet vorhergesagte Felder auf der Grundlage der Informationen zum Modell-Feld dem Ziel zu und ergibt auch dann eine Übereinstimmung, wenn ein vorhergesagtes Ziel umbenannt wurde. Die Informationen zum Modell-Feld können für jedes vorhergesagte Feld mithilfe eines Typknotens über das Dialogfeld "Werte" aufgerufen werden. Für weitere Informationen siehe Thema Verwenden des Dialogfelds "Werte" in Kapitel 4 auf S. 144.
- **Format für Feldnamen.** Ordnet Felder anhand der Namensgebungskonventionen zu. So müssen sich beispielsweise vorhergesagte Werte, die von einem C5.0-Modell-Nugget für das Ziel *Antwort* erstellt wurden, in einem Feld mit der Bezeichnung *$C-Antwort* befinden.

Trennen nach Partition. Wenn Datensätze mithilfe eines Partitionsfelds in Trainings-, Test- und Validierungsstichproben aufgeteilt werden, lassen Sie mit dieser Option die Ergebnisse für die einzelnen Partitionen separat anzeigen. Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 auf S. 208.

Anmerkung: Beim Aufteilen nach Partition werden Datensätze mit Nullwerten im Partitionsfeld von der Analyse ausgeschlossen. Dieses Problem tritt nicht auf, wenn Sie einen Partitionsknoten verwenden, weil diese Knoten keine Nullwerte erzeugen.

Benutzerdefinierte Analyse. Sie können eine eigene Analyseberechnung angeben, mit der die Modelle ausgewertet werden sollen. Legen Sie die zu berechnenden Werte für die einzelnen Datensätze mithilfe von CLEM-Ausdrücken fest und geben Sie an, wie die Scores auf der Ebene der Datensätze zu einem Gesamtwert zusammengefasst werden sollen. Mit den Funktionen @TARGET und @PREDICTED verweisen Sie auf den Zielwert (tatsächliche Ausgabe) bzw. auf den vorhergesagten Wert.

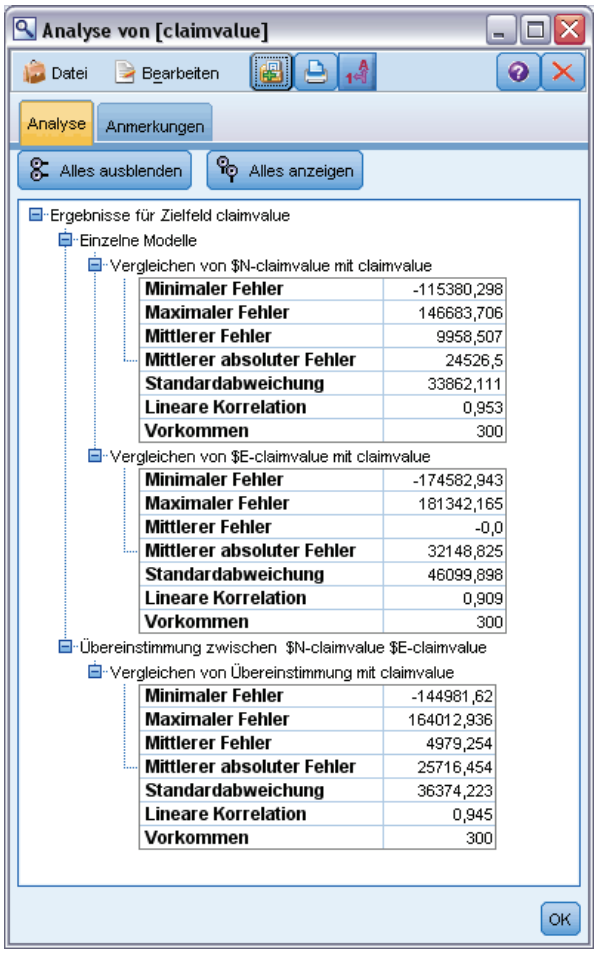
- **Wenn.** Legen Sie einen bedingten Ausdruck fest, wenn die auszuführende Berechnung von einer bestimmten Bedingungen abhängt.
- **Dann.** Geben Sie die Berechnung an, die ausgeführt werden soll, wenn die Wenn-Bedingung wahr ist.
- **Sonst.** Geben Sie die Berechnung an, die ausgeführt werden soll, wenn die Wenn-Bedingung falsch ist.
- **Verwenden.** Wählen Sie eine Statistik aus, mit der ein Gesamt-Score aus den Einzel-Scores berechnet werden soll.

Analyse nach Feldern aufschlüsseln. Zeigt die kategorialen Felder, die für die Aufschlüsselung der Analyse zur Verfügung stehen. Neben der Gesamtanalyse wird je eine separate Analyse für die einzelnen Kategorien in jedem Aufschlüsselungsfeld erstellt.

Analyseausgabe-Browser

Im Analyseausgabe-Browser werden die Ergebnisse aus der Ausführung des Analyseknotens angezeigt. Das Menü "Datei" enthält die üblichen Befehle zum Speichern, Exportieren und Drucken. Für weitere Informationen siehe Thema Anzeigen der Ausgabe auf S. 399.

Abbildung 6-13
Analyseausgabe-Browser



Beim Öffnen des Analyseausgabe-Browsers werden die Ergebnisse erweitert. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement links neben dem gewünschten Element die Ergebnisse reduzieren. Alternativ können Sie mit der Schaltfläche Alles ausblenden alle Ergebnisse ausblenden. Um die für Sie relevanten Ergebnisse nach dem Reduzieren wieder anzeigen zu lassen, erweitern Sie die gewünschten Ergebnisse

mithilfe des Erweiterungssteuerelements auf der linken Seite oder klicken Sie auf die Schaltfläche Alles anzeigen, um alle Ergebnisse anzuzeigen.

Ergebnisse für Zielfeld. Die Analyseausgabe enthält je einen Abschnitt für die einzelnen Ausgabefelder, für die ein zugehöriges Vorhersagefeld vorliegt, das durch ein erzeugtes Modell erstellt wurde.

Vergleich. Der Abschnitt mit den Ausgabefeldern enthält je einen Unterabschnitt für jedes Vorhersagefeld, das mit dem betreffenden Ausgabefeld verknüpft ist. Bei kategorialen Ausgabefeldern wird im oberen Bereich dieses Abschnitts eine Tabelle angezeigt, aus der die Anzahl und der Prozentsatz der richtigen und falschen Vorhersagen sowie die Gesamtanzahl der Datensätze im Stream hervorgeht. Bei numerischen Ausgabefeldern enthält dieser Abschnitt die folgenden Informationen:

- **Minimaler Fehler.** Zeigt den minimalen Fehler (Differenz zwischen beobachteten und vorhergesagten Werten).
- **Maximaler Fehler.** Zeigt den maximalen Fehler.
- **Mittlerer Fehler.** Zeigt den Durchschnitt (Mittelwert) der Fehler über alle Datensätze hinweg. Hieraus geht hervor, ob ein systematischer **Fehler** (eine stärkere Tendenz zu Überbewertungen als zu Unterbewertungen und umgekehrt) im Modell vorliegt.
- **Mittlerer absoluter Fehler.** Zeigt den Durchschnitt der absoluten Werte der Fehler über alle Datensätze hinweg. Weist auf die durchschnittliche Fehlergröße hin, unabhängig von der Richtung.
- **Standardabweichung.** Zeigt die Standardabweichung der Fehler.
- **Lineare Korrelation.** Zeigt die lineare Korrelation zwischen den vorhergesagten und den tatsächlichen Werten. Diese Statistik reicht von –1,0 bis 1,0. Werte nahe +1,0 weisen auf eine starke positive Assoziation hin; dies bedeutet, dass hohe vorhergesagte Werte mit hohen tatsächlichen Werten verknüpft sind und entsprechend niedrige vorhergesagte Werte mit niedrigen tatsächlichen Werten. Werte nahe –1,0 weisen auf eine starke negative Assoziation hin; dies bedeutet, dass hohe vorhergesagte Werte mit niedrigen tatsächlichen Werten verknüpft sind und umgekehrt. Werte nahe 0,0 weisen auf eine schwache Assoziation hin; dies bedeutet, dass die vorhergesagten Werte relativ unabhängig von den tatsächlichen Werten sind. *Hinweis*: Ein leerer Eintrag hier gibt an, dass eine lineare Korrelation in diesem Fall nicht berechnet werden kann, da entweder die tatsächlichen oder die vorhergesagten Werte Konstanten sind.
- **Vorkommen.** Zeigt die Anzahl der Datensätze, die für die Analyse herangezogen wurden.

Fehlklassifizierungstabelle. Bei kategorialen Ausgabefeldern wird hier ein Unterabschnitt mit der Matrix angezeigt, wenn Sie in den Analyseoptionen eine Fehlklassifizierungstabelle angefordert haben. Die Zeilen stehen für tatsächlich beobachtete Werte, die Spalten für vorhergesagte Werte. Die Zelle in der Tabelle gibt die Anzahl der Datensätze für jede Kombination aus vorhergesagten und tatsächlichen Werten an.

Leistungsauswertung. Bei kategorialen Ausgabefeldern werden hier die Ergebnisse der Leistungsauswertung angezeigt, wenn Sie die Leistungsauswertungsstatistik in den Analyseoptionen angefordert haben. Jede Ausgabekategorie wird gemeinsam mit der zugehörigen Leistungsauswertungsstatistik aufgeführt.

Übersicht Konfidenzwerte. Bei kategorialen Ausgabefeldern werden hier die Konfidenzwerte angezeigt, wenn Sie diese Werte in den Analyseoptionen angefordert haben. Für Modellkonfidenzwerte werden die folgenden Statistiken zusammengestellt:

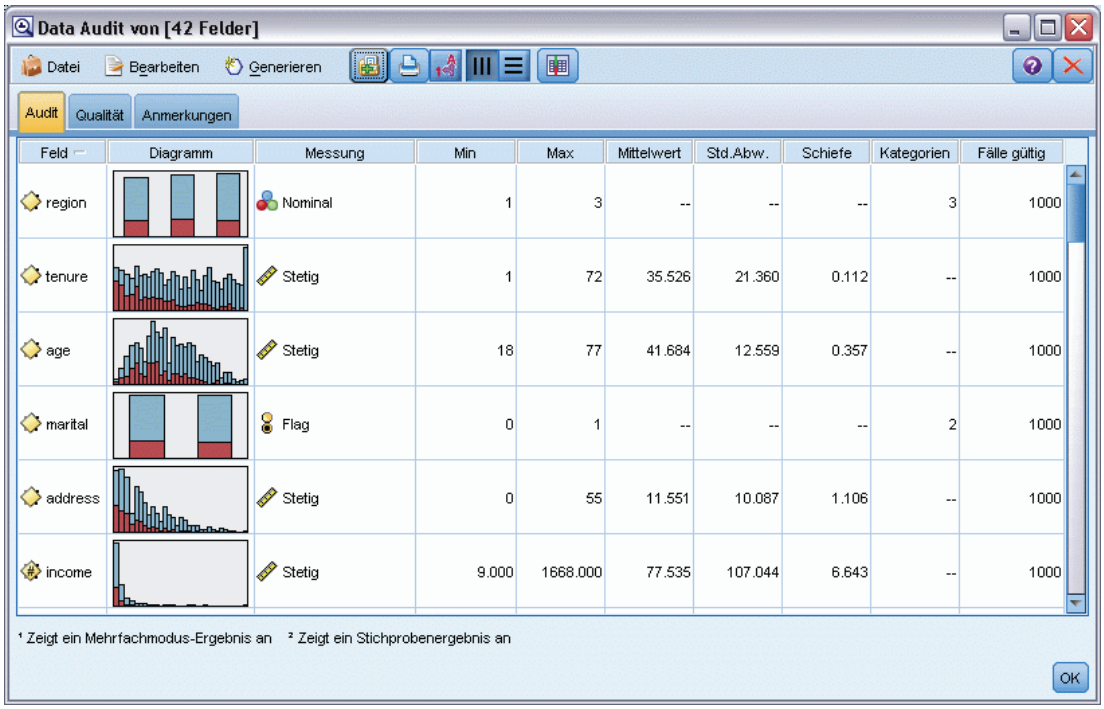
- **Bereich.** Zeigt den Bereich (kleinster und größter Wert) der Konfidenzwerte für die Datensätze in den Stream-Daten.
- **Mittelwert korrekt.** Zeigt die durchschnittliche Konfidenz für Datensätze, die als korrekt klassifiziert wurden.
- **Mittelwert inkorrekt.** Zeigt die durchschnittliche Konfidenz für Datensätze, die als inkorrekt klassifiziert wurden.
- **Immer korrekt über.** Zeigt den Schwellenwert für die Konfidenz, ab dem die Vorhersagen immer korrekt sind, sowie den Prozentsatz der Fälle, die dieses Kriterium erfüllen.
- **Immer korrekt unter.** Zeigt den Schwellenwert für die Konfidenz, bis zu dem die Vorhersagen immer korrekt sind, sowie den Prozentsatz der Fälle, die dieses Kriterium erfüllen.
- **x % Genauigkeit über.** Zeigt das Konfidenzniveau, bei dem die Genauigkeit bei *x %* liegt. *x* ist hierbei etwa gleich dem Wert, den Sie in den Analyseoptionen unter Schwellenwert für angegeben haben. Bei bestimmten Modellen und Daten-Sets ist es nicht möglich, einen Konfidenzwert auszuwählen, der genau gleich dem in den Optionen angegebenen Schwellenwert ist (in der Regel aufgrund von Clustern ähnlicher Fälle, die denselben Konfidenzwert nahe dem Schwellenwert aufweisen). Der angezeigte Schwellenwert ist der bestmögliche Näherungswert an das angegebene Kriterium für die Genauigkeit, der mit einem einzigen Schwellenwert für den Konfidenzwert erzielt werden kann.
- **x-fach korrekt über.** Zeigt den Konfidenzwert, bei dem die Genauigkeit *x*-mal besser ist als für das gesamte Daten-Set. *x* ist hierbei etwa gleich dem Wert, den Sie in den Analyseoptionen unter Genauigkeit verbessern angegeben haben.

Übereinstimmung zwischen. Enthält der Stream zwei oder mehr erzeugte Modelle, die eine Vorhersage für dasselbe Ausgabefeld enthalten, wird auch eine Statistik über die **Übereinstimmung** zwischen den durch die Modelle abgegebenen Vorhersagen angezeigt. Hierzu gehören die Anzahl und der Prozentsatz der Datensätze, bei denen die Vorhersagen übereinstimmen (bei kategorialen Ausgabefeldern), bzw. die Fehlerübersichtsstatistik (für stetige Ausgabefelder). Bei kategorialen Feldern wird eine Analyse der Vorhersagen im Vergleich zu den tatsächlichen Werten für die Untergruppe der Datensätze angezeigt, bei denen die Modelle miteinander übereinstimmen (also denselben vorhergesagten Wert erzeugen).

Data Audit-Knoten

Mit dem Data Audit-Knoten können Sie einen umfassenden ersten Blick auf die Daten werfen, die Sie in IBM® SPSS® Modeler einbringen. Die Ausgabe erfolgt in einer leicht lesbaren Matrix, die sortiert und zur Erstellung von normal großen Diagrammen und einer Vielzahl von Datenvorbereitungsknoten verwendet werden kann.

Abbildung 6-14
Data Audit-Browser

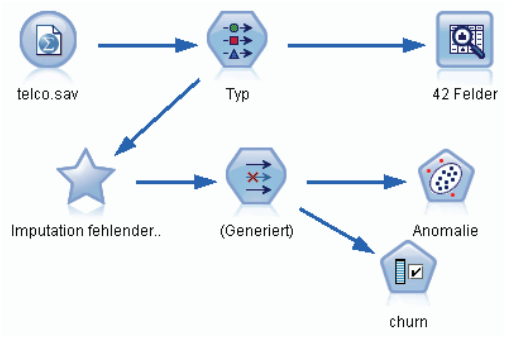


- Die Registerkarte “Audit” enthält einen Bericht mit Übersichtsstatistiken, Histogrammen und Verteilungsdiagrammen, die dabei helfen können, einen ersten Einblick in die Daten zu gewinnen. Der Bericht zeigt außerdem vor dem Feldnamen ein Symbol für den Speichertyp an.
- Auf der Registerkarte “Qualität” das Audit-Berichts finden Sie Informationen zu Ausreißern, Extremwerten und fehlenden Werten sowie Tools für den Umgang mit diesen Werten.

Verwenden des Data Audit-Knotens

Der Data Audit-Knoten kann direkt an einen Quellenknoten angehängt oder unterhalb eines instanziierten Typknotens eingefügt werden. Außerdem können Sie auf der Grundlage der Ergebnisse eine Reihe von Datenvorbereitungsknoten generieren. Beispielsweise können Sie einen Filterknoten generieren, der Felder ausschließt, die so viele fehlende Werte aufweisen, dass sie bei der Modellierung nicht mehr sinnvoll einsetzbar sind, und einen Superknoten generieren, der fehlende Werte für bestimmte oder alle verbleibenden Felder imputiert. Hier zeigt sich die wahre Stärke des Audit: Sie können nicht nur den aktuellen Status Ihrer Daten einschätzen, sondern auch aufgrund dieser Einschätzung aktiv werden.

Abbildung 6-15
Stream mit Superknoten für fehlende Werte

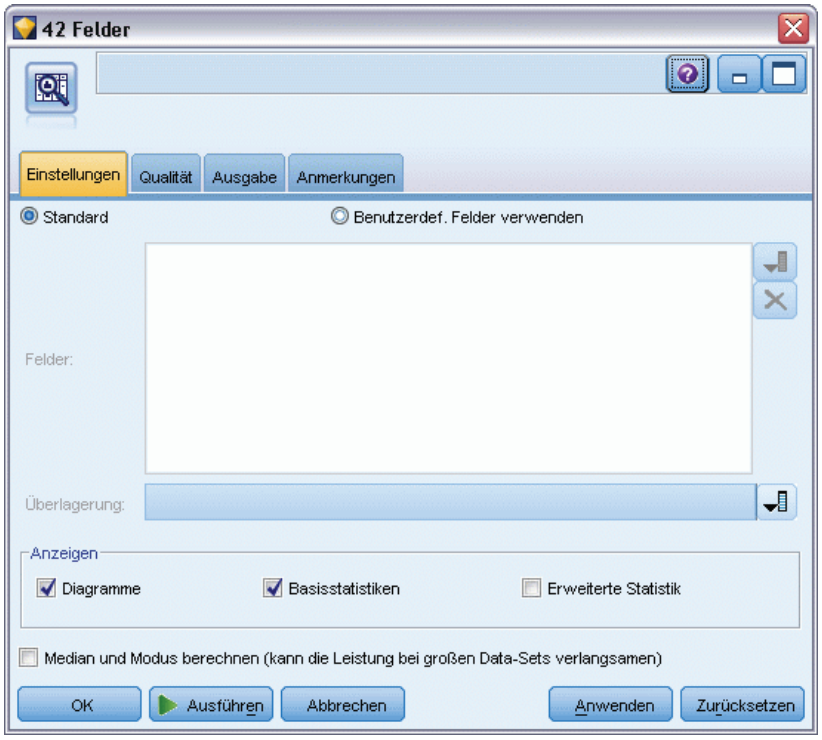


Sichten oder Stichprobennahme der Daten. Das anfängliche Audit ist besonders wirkungsvoll, wenn umfangreiche Daten anstehen. Aus diesem Grund kann ein Stichprobenknoten verwendet werden, mit dem Sie die Verarbeitungszeit während dieser ersten Untersuchung verkürzen können, indem Sie nur eine Untergruppe der Datensätze auswählen. Der Data Audit-Knoten kann auch in Verbindung mit Knoten wie "Merkmalsauswahl" und "Anomalieerkennung" in den Erkundungsphasen der Analyse verwendet werden.

Registerkarte "Einstellungen" beim Data Audit-Knoten

Auf der Registerkarte "Einstellungen" können Sie grundlegende Parameter für das Audit angeben.

Abbildung 6-16
Data Audit-Knoten: Registerkarte "Einstellungen"



Standard. Sie können beispielsweise den Knoten einfach an den Stream anhängen und auf Ausführen klicken. Auf diese Weise wird wie folgt ein Audit-Bericht für alle Felder erzeugt, der auf Standardeinstellungen beruht.

- Enthält der Typknoten keine Einstellungen, werden alle Felder in den Bericht aufgenommen.
- Falls Typeinstellungen vorhanden sind (unabhängig davon, ob diese instanziiert wurden oder nicht), werden alle Felder vom Typ *Eingabe*, *Ziel* und *Beides* in der Anzeige dargestellt. Liegt ein einzelnes Feld *Ziel* vor, wird dieses als Überlagerungsfeld herangezogen. Bei mehreren Feldern vom Typ *Ziel* wird keine Standardüberlagerung festgelegt.

Benutzerdef. Felder verwenden. Mit dieser Option können Sie manuell Felder auswählen. Wählen Sie die Felder mit der Feldauswahl-Schaltfläche rechts einzeln oder nach Typ aus.

Überlagerungsfeld Das Überlagerungsfeld dient zum Zeichnen der Miniaturdiagramme, die im Audit-Bericht angezeigt werden. Bei einem stetigen Feld (numerischer Bereich) werden außerdem bivariate Statistiken (Kovarianz und Korrelation) berechnet. Wenn ein einzelnes Feld vom Typ *Ziel* vorliegt, das auf den Einstellungen für den Knotentyp beruht, wird dieses, wie oben beschrieben, als Standard-Überlagerungsfeld verwendet. Alternativ können Sie Benutzerdef. Felder verwenden auswählen, um eine Überlagerung anzugeben.

Anzeigen. Ermöglicht die Angabe, ob Diagramme in der Ausgabe verfügbar sein sollen, sowie die Auswahl der standardmäßig anzuzeigenden Statistiken.

- **Diagramme.** Zeigt ein Diagramm für jedes ausgewählte Feld an. Dabei kann es sich um ein Verteilungsdiagramm (Balkendiagramm), ein Histogramm oder ein Streudiagramm handeln, je nachdem, was für die Daten geeignet ist. Die Diagramme werden im ursprünglichen Bericht in Miniaturform angezeigt, es können jedoch auch Diagramme in normaler Größe sowie Diagrammknoten generiert werden. Für weitere Informationen siehe Thema Data Audit-Ausgabe-Browser auf S. 425.
- **Basisstatistiken/Erweiterte Statistiken.** Gibt an in welchem Detailliertheitsgrad die Statistiken standardmäßig in der Ausgabe angezeigt werden sollen. Diese Einstellung legt zwar die ursprüngliche Anzeige fest, es sind jedoch unabhängig von dieser Einstellung alle Statistiken in der Ausgabe verfügbar. Für weitere Informationen siehe Thema Statistik anzeigen auf S. 428.

Median und Modus. Berechnet Median und Modus für alle Felder im Bericht. Beachten Sie, dass diese Statistiken bei großen Daten-Sets die Verarbeitungszeit erhöhen können, da ihre Berechnung länger dauert als die anderer Statistiken. Beim Median (und nur dort) kann der gemeldete Wert in einigen Fällen auf einer Stichprobe von 2.000 Datensätzen (anstelle des vollständigen Daten-Sets) beruhen. Diese Stichprobennahme erfolgt in Fällen, bei denen ansonsten die Arbeitsspeichergrenzen überschritten würden, auf der Grundlage einzelner Felder. Wenn die Stichprobennahme aktiviert wird, werden die Ergebnisse auch so in der Ausgabe beschriftet (*Median Stichpr.* anstatt *Median*). Alle anderen Statistiken als der Median werden immer mit dem vollständigen Daten-Set berechnet.

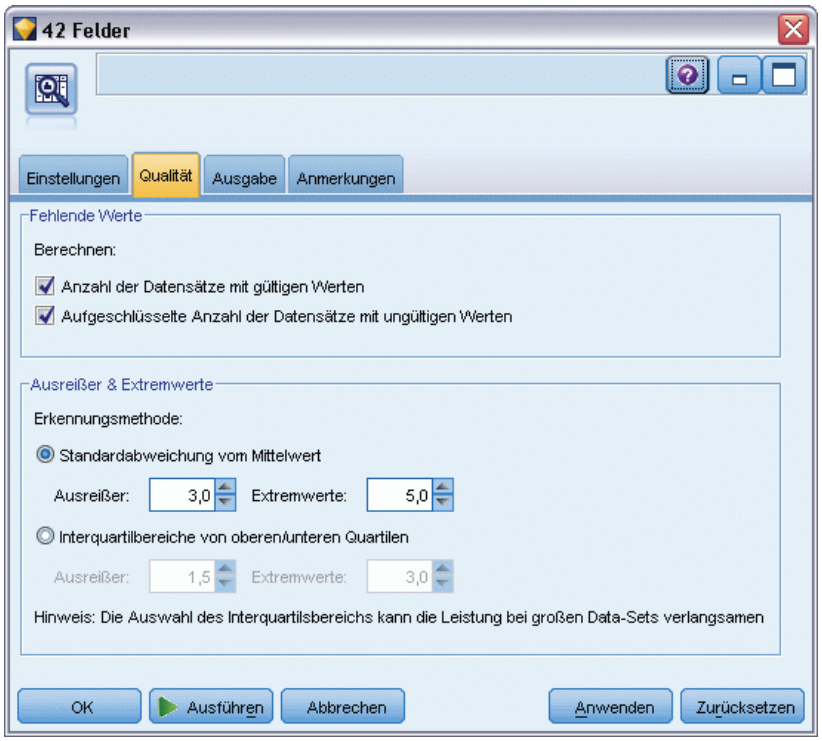
Leere Felder bzw. Felder ohne Typ. Bei Verwendung mit instanziierten Daten werden Felder ohne Typ nicht in den Audit-Bericht aufgenommen. Um Felder ohne Typ (einschließlich leerer Felder) aufzunehmen, wählen Sie in allen weiter oben im Stream liegenden Typknoten die Option Alle Werte löschen. Dadurch wird sichergestellt, dass keine Daten instanziiert und somit alle Felder in den Bericht aufgenommen werden. Dies kann beispielsweise dann nützlich sein, wenn Sie eine vollständige Liste aller Felder benötigen oder einen Filterknoten generieren möchten, der

alle leeren Felder ausschließt. Für weitere Informationen siehe Thema Filtern von Feldern mit fehlenden Daten auf S. 433.

Data Audit – Registerkarte "Qualität"

Die Registerkarte "Qualität" im Data Audit-Knoten bietet Optionen für den Umgang mit fehlenden Werten, Ausreißern und Extremwerten.

Abbildung 6-17
Registerkarte "Qualität" beim Data Audit-Knoten



Fehlende Werte definieren

- **Anzahl der Datensätze mit gültigen Werten.** Mit dieser Option rufen Sie die Anzahl der Datensätze ab, die gültige Werte für alle ausgewählten Felder enthalten. Beachten Sie, dass Nullwerte (nicht definierte Werte), Leerwerte, leere Bereiche und leere Zeichenketten stets als ungültige Werte behandelt werden.
- **Aufgeschlüsselte Anzahl der Datensätze mit ungültigen Werten.** Mit dieser Option ermitteln Sie die Anzahl der Datensätze sowie die verschiedenen Typen der ungültigen Werte für jedes Feld ab.

Ausreißer und Extremwerte

Erkennungsmethode für Ausreißer und Extremwerte. Es werden zwei Methoden unterstützt:

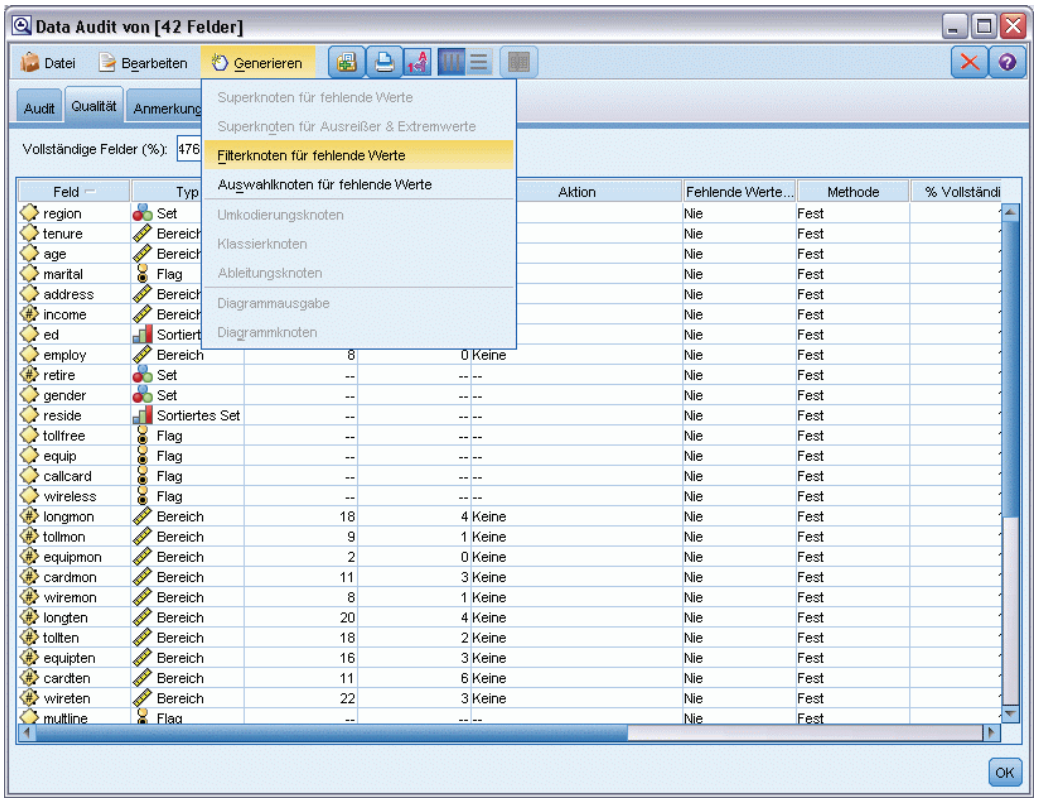
Standardabweichung vom Mittelwert. Erkennt Ausreißer und Extremwerte anhand der Anzahl an Standardabweichungen vom Mittelwert. Nehmen wir beispielsweise an, Sie haben ein Feld mit einem Mittelwert von 100 und einer Standardabweichung von 10. In diesem Fall könnten Sie 3,0 angeben, um festzulegen, dass jeder Wert unter 70 und über 130 als Ausreißer behandelt werden soll.

Interquartilbereich. Erkennt Ausreißer und Extremwerte anhand des Interquartilbereichs, also des Bereichs, in den die beiden mittleren Quartile fallen (zwischen dem 25. und dem 75. Perzentil). Auf der Grundlage der Standardeinstellung 1,5 beispielsweise wäre der untere Schwellenwert für Ausreißer Q1 – 1.5 * IQB und der obere Schwellenwert wäre Q3 + 1,5 * IQB. Beachten Sie, dass diese Option bei großen Daten-Sets zu Geschwindigkeitseinbußen führen kann.

Data Audit-Ausgabe-Browser

Der Data Audit-Browser ist ein leistungsstarkes Tool, mit dem Sie einen Überblick über Ihre Daten gewinnen können. Auf der Registerkarte “Audit” werden Miniaturdiagramme, Symbole für den Speichertyp und Statistiken für alle Felder angezeigt, auf der Registerkarte “Qualität” finden Sie Informationen zu Ausreißern, Extremwerten und fehlenden Werten. Auf der Grundlage der anfänglichen Diagramme und der Übersichtsstatistik können Sie beispielsweise ein numerisches Feld neu kodieren, ein neues Feld ableiten oder auch die Werte eines nominalen Felds umkodieren. Des Weiteren können Sie die Untersuchung mithilfe einer ausgereiften Visualisierung fortsetzen. Dies können Sie über den Browser für Audit-Berichte erreichen, indem Sie mithilfe des Menüs “Generieren” eine Reihe von Knoten erzeugen, die zum Transformieren bzw. Visualisieren der Daten verwendet werden können.

Abbildung 6-18
Erstellen eines Filterknotens für fehlende Werte

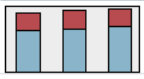
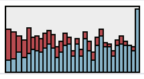


- Sortieren Sie die Spalten durch Klicken auf die gewünschte Spaltenüberschrift oder ordnen Sie die Spalten durch Ziehen und Ablegen neu. Außerdem werden die meisten Standard-Ausgabeoperationen unterstützt. Für weitere Informationen siehe Thema Anzeigen der Ausgabe auf S. 399.
- Werte und Bereiche für Felder abrufen: Doppelklicken Sie auf ein Feld in der Spalte “Messung” oder “Eindeutig”.
- Verwenden Sie die Symbolleiste oder das Menü “Bearbeiten”, um Wertelabels ein- bzw. auszublenden bzw. um die anzuzeigenden Statistiken auszuwählen. Für weitere Informationen siehe Thema Statistik anzeigen auf S. 428.
- Überprüfen Sie die Speichertypsymbole links neben den Feldnamen. Der Speichertyp beschreibt die Art und Weise, wie Daten in einem Feld gespeichert werden. Beispiel: Ein Feld mit den Werten 1 und 0 speichert ganzzahlige Daten. Dies ist vom Messniveau zu unterscheiden, das die Verwendung der Daten beschreibt und sich nicht auf den Speichertyp auswirkt. Für weitere Informationen siehe Thema Festlegen von Feldspeichertyp und Formatierung in Kapitel 2 auf S. 32.

Anzeigen und Generieren von Diagrammen

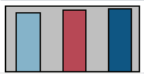
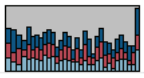
Ist keine Überlagerung ausgewählt, werden auf der Registerkarte “Audit” entweder Balkendiagramme (für nominale oder Flag-Felder) oder Histogramme (stetige Felder) angezeigt.

Abbildung 6-19
Auszug aus Audit-Ergebnissen ohne Überlagerungsfeld

Feld	Diagramm	Typ	Min	Max	Mittelwert	Std.Abw.	Schiefe	Kategorien	Fälle gültig
region		Set	1	3	--	--	--	3	1000
tenure		Bereich	1	72	35.526	21.360	0.112	--	1000

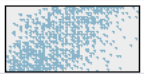
Bei Überlagerung mit einem nominalen oder Flag-Feld werden die Diagramme gemäß den Werten der Überlagerung farbig gekennzeichnet.

Abbildung 6-20
Auszug aus Audit-Ergebnissen mit Überlagerung durch ein nominales Feld

Feld	Diagramm	Typ	Min	Max	Mittelwert	Std.Abw.	Schiefe	Kategorien	Fälle gültig
region		Set	1	3	--	--	--	3	1000
tenure		Bereich	1	72	35.526	21.360	0.112	--	1000

Bei Überlagerung mit einem stetigen Feld werden keine eindimensionalen Balkendiagramme und Histogramme erzeugt, sondern zweidimensionale Streudiagramme. In diesem Fall wird die *x*-Achse dem Überlagerungsfeld zugeordnet, sodass bei allen *x*-Achsen in der gesamten Tabelle jeweils dieselbe Skala verwendet wird.

Abbildung 6-21
Auszug aus Audit-Ergebnissen mit Überlagerung durch ein stetiges Feld

Feld	Diagramm	Typ	Min	Max	Mittelwert	Korrelation	Korrelation T	Korrelation T df.
region		Set	1	3	--	--	--	--
tenure		Bereich	1	72	35.526	0.490	17.768	998.000

- Halten Sie bei Flag- bzw. nominalen Feldern den Mauscursor über einen Balken, um den zugrunde liegenden Wert bzw. die Beschriftung in einer QuickInfo anzuzeigen.
- Verwenden Sie bei Flag- bzw. nominalen Feldern die Symbolleiste, um die Ausrichtung der Miniaturdiagramme von horizontal zu vertikal zu ändern.
- Um ein normal großes Diagramm aus einer Miniaturansicht zu generieren, doppelklicken Sie auf die Miniaturansicht oder wählen Sie eine Miniaturansicht aus und wählen Sie im Menü "Generieren" die Option Diagrammausgabe. *Hinweis*: Beruht ein Miniaturdiagramm auf Daten in einer Stichprobe, werden alle Fälle in das erzeugte Diagramm aufgenommen, wenn der ursprüngliche Daten-Stream noch geöffnet ist.

  Sie können auch ein Diagramm erzeugen, wenn der Data-Audit-Knoten, der die Ausgabe erzeugt hat, mit dem Stream verbunden ist.

- Um einen passenden Diagrammknoten zu generieren, wählen Sie auf der Registerkarte “Audit” mindestens ein Feld aus und wählen Sie im Menü “Generieren” die Option Diagrammknoten. Der so entstehende Knoten wird dem Stream-Zeichenbereich hinzugefügt und kann verwendet werden, um das Diagramm bei jeder Ausführung des Streams neu zu erstellen.
- Enthält eine Überlagerung mehr als 100 Werte, wird eine Warnmeldung eingeblendet und die Überlagerung wird nicht berücksichtigt.

Statistik anzeigen

Im Dialogfeld” Statistik anzeigen” können Sie die auf der Registerkarte “Audit” anzuzeigenden Statistiken auswählen. Die ursprünglichen Einstellungen werden im Data Audit-Knoten angegeben. Für weitere Informationen siehe Thema Registerkarte “Einstellungen” beim Data Audit-Knoten auf S. 422.

Abbildung 6-22
Statistik anzeigen

Minimum. Der kleinste Wert einer numerischen Variablen.

Maximum. Der größte Wert einer numerischen Variablen.

Summe. Die Summe der Werte über alle Fälle mit nichtfehlenden Werten.

Spannweite. Die Differenz zwischen den größten und kleinsten Werten einer numerischen Variablen; Maximalwert minus Minimalwert.

Mittelwert. Ein Lagemaß. Die Summe der Ränge, geteilt durch die Zahl der Fälle.

Standardfehler des Mittelwerts. Ein Maß für die mögliche Variation des Mittelwerts zwischen aus derselben Verteilung stammenden Stichproben. Dieser Wert kann für einen ungefähren Vergleich des beobachteten Mittelwerts mit einem hypothetischen Wert verwendet werden. (Es kann geschlossen werden, dass die beiden Werte unterschiedlich sind, wenn das Verhältnis der Differenz zum Standardfehler kleiner als -2 oder größer als +2 ist.)

Standardabweichung. Ein Maß für die Streuung um den Mittelwert, definiert als positive Wurzel aus der Varianz. Die Standardabweichung wird in denselben Einheiten gemessen wie die ursprüngliche Variable.

Varianz. Ein Maß der Streuung um den Mittelwert. Es ist gleich dem Quotienten aus der Summe der quadrierten Abweichung vom Mittelwert und der um 1 verringerten Fallanzahl. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.

Schiefe. Ein Maß für die Asymmetrie einer Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefe-Wert, der mehr als doppelt so groß ist wie sein Standardfehler, für eine Abweichung von der Symmetrie spricht.

Standardfehler der Schiefe. Das Verhältnis der Schiefe zu ihrem Standardfehler kann für einen Test auf Normalverteilung verwendet werden (d. h. die Annahme, dass Normalverteilung vorliegt, kann abgelehnt werden, wenn das Verhältnis kleiner als -2 oder größer als +2 ist). Ein großer positiver Wert für die Schiefe bedeutet, dass die Verteilung eine lange rechte Flanke hat; ein extremer negativer Wert bedeutet, dass sie eine lange linke Flanke hat.

Kurtosis. Ein Maß dafür, wie sich die Beobachtungen um einen zentralen Punkt gruppieren. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng gruppiert und haben dickere Flanken bis hin zu den Extremverten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.

Standardfehler der Kurtosis. Das Verhältnis der Kurtosis zu ihrem Standardfehler kann für einen Test auf Normalverteilung verwendet werden (d. h. die Annahme, dass Normalverteilung vorliegt, kann abgelehnt werden, wenn das Verhältnis kleiner als -2 oder größer als +2 ist). Ein großer positiver Wert für die Kurtosis deutet darauf hin, dass die Flanken der Verteilung länger sind als bei einer Normalverteilung; ein negativer Wert bedeutet, dass sie kürzer sind (etwa wie bei einer kastenförmigen, gleichförmigen Verteilung).

Eindeutig. Bewertet alle Effekte gleichzeitig; damit werden alle Effekte an alle sonstigen Effekte jedweden Typs angepasst.

Gültig. Gültige Fälle, d. h. solche, die weder den systemdefiniert fehlenden Wert noch einen benutzerdefiniert fehlenden Wert aufweisen.

Median. Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).

Modalwert. Der am häufigsten auftretende Wert. Wenn mehrere Werte gleichermaßen die größte Häufigkeit aufweisen, ist jeder von ihnen ein Modalwert.

Beachten Sie, dass Median und Modus standardmäßig unterdrückt sind, um die Leistungsfähigkeit zu erhöhen. Diese Statistiken können jedoch im Data Audit-Knoten auf der Registerkarte “Einstellungen” ausgewählt werden. Für weitere Informationen siehe Thema Registerkarte “Einstellungen” beim Data Audit-Knoten auf S. 422.

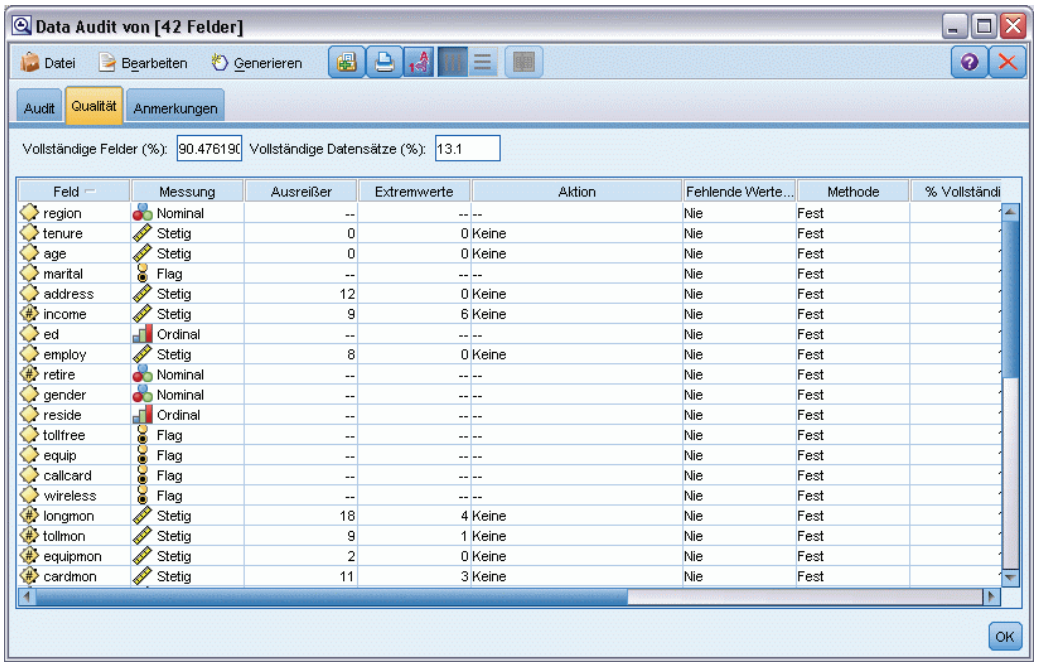
Statistiken für Überlagerungen

Wenn ein stetiges Überlagerungsfeld (numerischer Bereich) verwendet wird, stehen außerdem folgende Statistiken zur Verfügung:

Kovarianz. Ein nicht standardisiertes Maß für den linearen Zusammenhang zwischen zwei Variablen. Es ist gleich der Kreuzproduktsumme der Abweichungen geteilt durch N-1.

Registerkarte "Qualität" beim Data Audit-Browser

Abbildung 6-23
Qualitätsbericht im Data Audit-Browser



Feld	Messung	Ausreißer	Extremwerte	Aktion	Fehlende Werte...	Methode	% Vollständi
region	Nominal	--	--	--	Nie	Fest	
tenure	Stetig	0	0	Keine	Nie	Fest	
age	Stetig	0	0	Keine	Nie	Fest	
marital	Flag	--	--	--	Nie	Fest	
address	Stetig	12	0	Keine	Nie	Fest	
income	Stetig	9	6	Keine	Nie	Fest	
ed	Ordinal	--	--	--	Nie	Fest	
employ	Stetig	8	0	Keine	Nie	Fest	
retire	Nominal	--	--	--	Nie	Fest	
gender	Nominal	--	--	--	Nie	Fest	
reside	Ordinal	--	--	--	Nie	Fest	
tollfree	Flag	--	--	--	Nie	Fest	
equip	Flag	--	--	--	Nie	Fest	
callcard	Flag	--	--	--	Nie	Fest	
wireless	Flag	--	--	--	Nie	Fest	
longmon	Stetig	18	4	Keine	Nie	Fest	
tollmon	Stetig	9	1	Keine	Nie	Fest	
equipmon	Stetig	2	0	Keine	Nie	Fest	
cardmon	Stetig	11	3	Keine	Nie	Fest	

Auf der Registerkarte "Qualität" des Data Audit-Browsers werden die Ergebnisse der Datenqualitätsanalyse angezeigt. Außerdem können Sie hier angeben, wie Ausreißer, Extremwerte und fehlende Werte behandelt werden sollen.

Fehlende Werte imputieren

Der Audit-Bericht listet den Prozentsatz vollständiger Datensätze für die einzelnen Felder auf, dazu die Anzahl der gültigen Werte, der Nullwerte und der leeren Werte. Sie können je nach Bedarf fehlende Werte für bestimmte Felder imputieren und anschließend einen Superknoten generieren, um die Transformationen anzuwenden.

- Geben Sie in der Spalte Fehlende Werte imputieren die zu imputierenden Wertetypen an, sofern vorhanden. Sie können festlegen, dass Leerstellen, Nullen oder beides imputiert werden sollen, oder eine benutzerdefinierte Bedingung bzw. einen benutzerdefinierten Ausdruck angeben, der die zu imputierenden Werte auswählt.

In Clementine gibt es mehrere Arten von fehlenden Werten, die von IBM® SPSS® Modeler erkannt werden:

- **Nullwerte oder systemdefiniert fehlende Werte.** Bei diesen Werten handelt es sich um Nicht-Zeichenketten-Werte, die in der Datenbank bzw. der Quelldatei leer gelassen und nicht speziell in einem Quellen- oder Typknoten als "fehlend" definiert wurden. Systemdefiniert fehlende Werte werden als $null$ angezeigt. Beachten Sie, dass leere Zeichenketten in SPSS Modeler nicht als Nullen betrachtet werden, auch wenn sie von bestimmten Datenbanken (siehe unten) als Nullen behandelt werden können.
- **Leere Zeichenketten und leere Bereiche.** Leere Zeichenkettenwerte und leere Bereiche (Zeichenketten ohne sichtbare Zeichen) werden anders als Nullwerte behandelt. Leere Zeichenketten werden in den meisten Fällen als äquivalent mit leeren Bereichen (Leerzeichen) behandelt. Beispiel: Wenn Sie die Option auswählen, dass leere Bereiche in einem Quellen- oder Typknoten als Leerstellen behandelt werden sollen, gilt diese Einstellung auch für leere Zeichenketten.
- **Leere oder benutzerdefiniert fehlende Werte.** Es handelt sich hierbei um Werte wie unbekannt, 99 oder –1, die in einem Quellen- oder Typknoten ausdrücklich als fehlend definiert sind. Optional können Sie auch auswählen dass Nullen und leere Bereiche als Leerzeichen behandelt werden sollen. Dadurch können sie mit Flags für eine spezielle Behandlung versehen und aus den meisten Berechnungen ausgeschlossen werden. Beispielsweise können Sie die Funktion @BLANK verwenden, um diese Werte gemeinsam mit anderen Arten von fehlenden Werten als Leerstellen zu behandeln. Für weitere Informationen siehe Thema Verwenden des Dialogfelds "Werte" in Kapitel 4 auf S. 144.

▶ Geben Sie in der Spalte Methode die zu verwendende Methode an.

Folgende Methoden stehen zum Imputieren fehlender Werte zur Verfügung:

Fest. Ersetzt einen festen Wert (Feldmittelwert, Mittelpunkt des Bereichs oder eine von Ihnen angegebene Konstante).

Zufällig. Ersetzt einen Zufallswert auf der Grundlage einer Normal- oder Gleichverteilung.

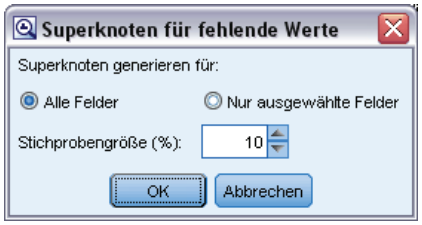
Ausdruck. Ermöglicht die Angabe eines benutzerdefinierten Ausdrucks. Beispielsweise könnten Sie Werte durch eine globale Variable ersetzen, die vom Globalwerteknoten erstellt wurde.

Algorithmus. Ersetzt einen von einem Modell vorhergesagten Wert auf der Grundlage eines C&RT-Algorithmus. Für jedes Feld, das unter Verwendung dieser Methode imputiert wurde, gibt es ein separates C&RT-Modell sowie einen Füllerknoten, der Leerstellen und Nullen durch den vom Modell vorhergesagten Wert ersetzt. Anschließend werden die vom Modell generierten Vorhersagefelder mithilfe eines Filterknotens entfernt.

▶ Um einen Superknoten für fehlende Werte zu generieren, wählen Sie folgende Optionsfolge aus den Menüs aus:

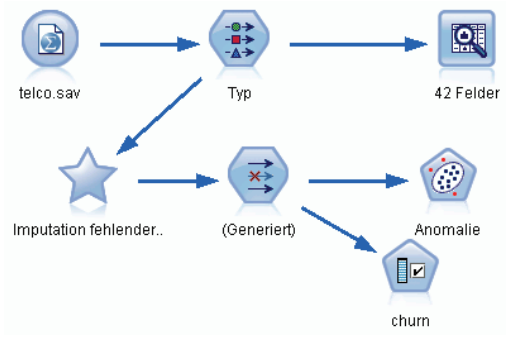
Erzeugen > Superknoten für fehlende Werte

Abbildung 6-24
Dialogfeld "Superknoten für fehlende Werte"



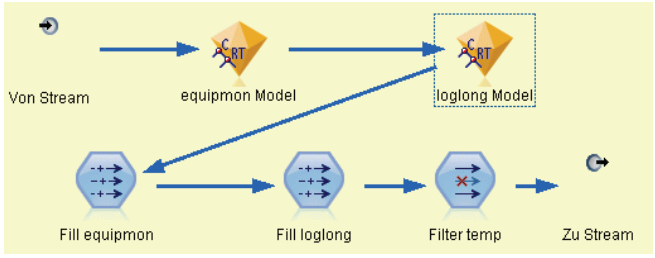
- Wählen Sie Alle Felder oder Nur ausgewählte Felder aus und geben Sie bei Bedarf einen Stichprobenumfang an. (Die Stichprobe wird als Prozentsatz angegeben; standardmäßig werden 10 % aller Datensätze in die Stichprobe aufgenommen.)
- Klicken Sie auf OK, um den generierten Superknoten zum Stream-Zeichenbereich hinzuzufügen.
- Fügen Sie den Superknoten zum Stream hinzu, um die Transformationen anzuwenden.

Abbildung 6-25
Hinzufügen des Superknotens zum aktuellen Stream



Innerhalb des Superknotens wird je nach Bedarf eine Kombination aus Modell-Nugget, Füllerknoten und Filterknoten verwendet. Um Einblicke in die Funktionsweise zu gewinnen, können Sie den Superknoten bearbeiten und auf Vergrößern klicken. Außerdem können Sie einzelne Knoten im Superknoten hinzufügen, bearbeiten bzw. entfernen, um eine Feineinstellung des Verhaltens vorzunehmen.

Abbildung 6-26
Vergrößern des Superknotens.



Umgang mit Ausreißern und Extremwerten

Im Audit-Bericht wird die Anzahl der Ausreißer und Extremwerte für die einzelnen Felder auf der Grundlage der im Data Audit-Knoten angegebenen Erkennungsoptionen aufgeführt. Für weitere Informationen siehe Thema Data Audit – Registerkarte "Qualität" auf S. 424. Sie können je nach Bedarf diese Werte für bestimmte Felder erzwingen, verwerfen oder auf null setzen und anschließend einen Superknoten generieren, um die Transformationen anzuwenden.

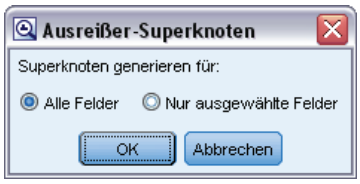
▶ Geben Sie in der Spalte Aktion den gewünschten Umgang mit Ausreißern und Extremwerten für bestimmte Felder an.

Für den Umgang mit Ausreißern und Extremwerten stehen folgende Aktionen zur Auswahl:

- **Erzwingen.** Ersetzt Ausreißer und Extremwerte durch den nächsten Wert, der nicht als Extremwert betrachtet würde. Wenn beispielsweise alle Werte als Ausreißer gelten, die den Bereich von drei Standardabweichungen über- bzw. unterschreiten, werden alle Ausreißer mit dem höchsten bzw. niedrigsten Wert innerhalb dieses Bereichs ersetzt.
- **Verwerfen.** Verwirft Datensätze mit Ausreißern bzw. Extremwerten für das angegebene Feld.
- **Auf null setzen.** Ersetzt Ausreißer und Extremwerte durch den Nullwert bzw. den systemdefiniert fehlenden Wert.
- **Ausreißer erzwingen/Extremwerte verwerfen.** Verwirft nur Extremwerte.
- **Ausreißer erzwingen/Extremwerte auf null setzen.** Setzt nur Extremwerte auf null.

▶ Um den Superknoten zu generieren, wählen Sie folgende Optionsfolge aus den Menüs aus:
Erzeugen > Superknoten für Ausreißer & Extremwerte

Abbildung 6-27
Dialogfeld "Ausreißer-Superknoten"



▶ Wählen Sie Alle Felder oder Nur ausgewählte Felder aus und klicken Sie dann auf OK, um den generierten Superknoten zum Stream-Zeichenbereich hinzuzufügen.

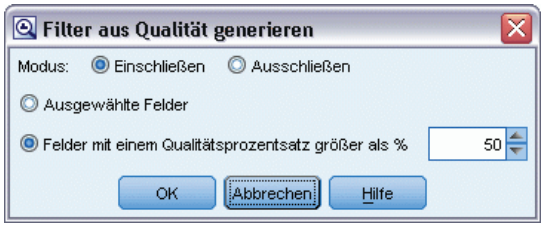
▶ Fügen Sie den Superknoten zum Stream hinzu, um die Transformationen anzuwenden.

Optional können Sie den Superknoten bearbeiten und vergrößern, um ihn zu durchsuchen oder Änderungen vorzunehmen. Innerhalb des Superknotens werden Werte mithilfe einer Reihe von Auswahl- und/oder Füllerknoten verworfen, erzwungen oder auf null gesetzt.

Filtern von Feldern mit fehlenden Daten

Ausgehend vom Data Audit-Browser können Sie einen neuen Filterknoten auf der Grundlage der Ergebnisse aus der Qualitätsanalyse erstellen.

Abbildung 6-28
Dialogfeld "Filter aus Qualität generieren"



Modalwert. Wählen Sie die gewünschte Funktion für die ausgewählten Felder aus (Einschließen oder Ausschließen).

- **Ausgewählte Felder.** Mit dem Filterknoten werden die Felder eingeschlossen bzw. ausgeschlossen, die in der Qualitätstabelle ausgewählt wurden. Beispielsweise können Sie die Tabelle anhand der Spalte % Vollständig sortieren, durch Klicken bei gedrückter Umschalttaste die am wenigsten vollständigen Felder auswählen und anschließend einen Filterknoten generieren, der diese Felder ausschließt.
- **Felder mit einem Qualitätsprozentsatz größer als.** Mit dem Filterknoten werden die Felder eingeschlossen bzw. ausgeschlossen, bei dem der Prozentsatz der abgeschlossenen Datensätze höher ist als der angegebene Schwellenwert. Der Standard-Schwellenwert beträgt 50 %.

Filtern von leeren Feldern bzw. Feldern ohne Typ

Beachten Sie: Nach dem Instanziieren von Datenwerten werden Felder ohne Typ bzw. leere Felder aus den Audit-Ergebnissen und den meisten anderen Ausgaben in IBM® SPSS® Modeler ausgeschlossen. Diese Felder werden zum Zweck der Modellierung ignoriert, können die Daten jedoch aufblähen bzw. unübersichtlich werden lassen. In diesem Fall können Sie mit dem Data Audit-Browser einen Filterknoten generieren, der diese Felder aus dem Stream entfernt.

▶ Um sicherzustellen, dass alle Felder in das Audit aufgenommen werden, einschließlich leerer Felder und Feldern ohne Typen, wählen Sie im weiter oben im Stream liegenden Quellen- oder Typknoten die Option Alle Werte löschen oder setzen Sie für alle Felder "Werte" auf <*Übergeben*>.

▶ Sortieren Sie die Daten im Data Audit-Browser anhand der Spalte % Vollständig, wählen Sie die Felder mit 0 (oder anderer Schwellenwert) gültigen Werten aus und erstellen Sie über das Menü "Generieren" einen Filterknoten, der zum Stream hinzugefügt werden kann.

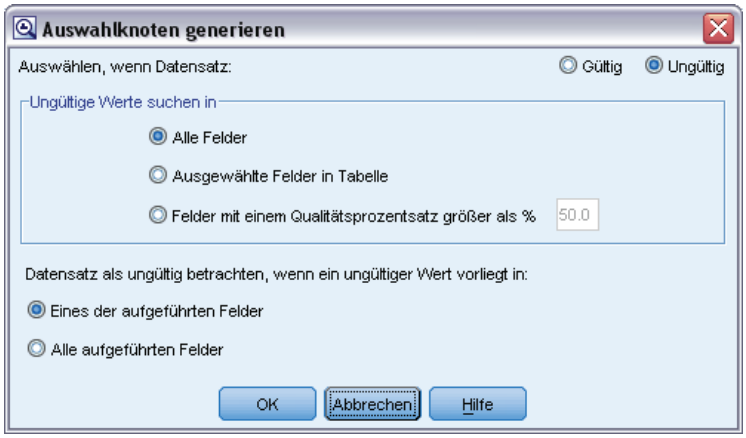
Auswählen von Datensätzen mit fehlenden Daten

Ausgehend vom Data Audit-Browser können Sie einen neuen Auswahlknoten auf der Grundlage der Ergebnisse aus der Qualitätsanalyse erstellen.

▶ Wählen Sie im Data Audit-Browser die Registerkarte "Qualität".

▶ Wählen Sie die folgenden Befehle aus dem Menü aus:
Erzeugen > Fehlende Werte - Auswahlknoten

Abbildung 6-29
Dialogfeld "Auswahlknoten generieren"



Auswählen, wenn Datensatz. Geben Sie an, ob die Datensätze beibehalten werden sollen, wenn diese Gültig oder Ungültig sind.

Ungültige Werte suchen in. Geben Sie an, wo nach ungültigen Werten gesucht werden soll.

- **Alle Felder.** Mit dem Auswahlknoten werden alle Felder auf ungültige Werte geprüft.
- **Ausgewählte Felder in Tabelle.** Mit dem Auswahlknoten werden nur die Felder geprüft, die derzeit in der Qualitätsausgabetabelle ausgewählt sind.
- **Felder mit einem Qualitätsprozentsatz größer als.** Mit dem Auswahlknoten werden alle Felder geprüft, bei dem der Prozentsatz der abgeschlossenen Datensätze höher ist als der angegebene Schwellenwert. Der Standard-Schwellenwert beträgt 50 %.

Datensatz als ungültig betrachten, wenn ein ungültiger Wert vorliegt in. Legen Sie die Bedingung fest, unter der ein Datensatz als ungültig betrachtet wird.

- **Eines der aufgeführten Felder.** Mit dem Auswahlknoten wird ein Datensatz als ungültig betrachtet, wenn *eines* der oben angegebenen Felder einen ungültigen Wert für diesen Datensatz enthält.
- **Alle aufgeführten Felder.** Mit dem Auswahlknoten wird ein Datensatz als ungültig betrachtet, wenn *alle* oben angegebenen Felder einen ungültigen Wert für diesen Datensatz enthalten.

Erzeugen von anderen Knoten zur Datenvorbereitung

Zahlreiche Knoten für die Datenvorbereitung können direkt aus dem Audit-Bericht-Browser heraus erzeugt werden, beispielsweise Umkodierungs-, Klassier- und Ableitungsknoten. Beispiel:

- Soll ein neues Feld auf der Grundlage der Werte für *Schadensersatzforderung* und *Grundwert* erzeugt werden, wählen Sie beide Felder im Audit-Bericht aus und wählen Sie dann im Menü "Generieren" den Befehl Ableiten. Der neue Knoten wird in den Stream-Zeichenbereich aufgenommen.
- Unter Umständen stellen Sie auf der Grundlage der Audit-Ergebnisse fest, dass eine Umkodierung von *Grundwert* in perzentilbasierte Klassen eine stärker zielgerichtete Analyse ergäbe. Um einen Klassierknoten zu erzeugen, wählen Sie die Feldzeile in der Anzeige aus und wählen Sie dann im Menü "Generieren" den Befehl Klassieren.

Sobald Sie einen Knoten erzeugt und zum Stream-Zeichenbereich hinzugefügt haben, muss dieser Knoten an den Stream angehängt und geöffnet werden; legen Sie dann die Optionen für das oder die ausgewählten Felder fest.

Transformationsknoten

Die Normalisierung der Eingabefelder ist ein wichtiger Schritt vor der Anwendung herkömmlicher Scoring-Verfahren wie Regression, logistische Regression und Diskriminanzanalyse. Bei diesen Verfahren wird von Annahmen über die Normalverteilung von Daten ausgegangen, die für viele Rohdatendateien möglicherweise nicht gelten. Ein Ansatz für den Umgang mit realen Daten besteht in der Anwendung von Transformationen, die ein Rohdatenelement mehr in Richtung einer Normalverteilung verschieben. Außerdem können normalisierte Felder leicht miteinander verglichen werden. So befinden sich Einkommen und Alter in einer Rohdatendatei auf vollständig unterschiedlichen Skalen, nach einer Normalisierung jedoch lassen sich die relativen Auswirkungen der beiden Datenarten leicht interpretieren.

Der Transformationsknoten enthält einen Ausgabe-Viewer, mit dem Sie eine schnelle Sichtprüfung zur Ermittlung der besten Transformation durchführen können Sie sehen auf einen Blick, ob die Variablen normal verteilt sind und können, falls erforderlich, die gewünschte Transformation auswählen und anwenden. Sie können mehrere Felder auswählen und pro Feld jeweils eine Transformation durchführen.

Nach Auswahl der bevorzugten Transformationen für die Felder können Sie Ableitungs- oder Füllerknoten generieren, die die Transformationen durchführen, und diese Knoten zum Stream hinzufügen. Der Ableitungsknoten erstellt neue Felder, während der Füllerknoten die bestehenden transformiert. Für weitere Informationen siehe Thema Erzeugen von Diagrammen auf S. 441.

Registerkarte “Felder” beim Transformationsknoten

Auf der Registerkarte “Felder” können Sie angeben, welche Datenfelder zur Anzeige und Anwendung möglicher Transformationen verwendet werden sollen. Eine Transformation ist nur bei numerischen Feldern möglich. Klicken Sie auf die Feldauswahlschaltfläche und wählen Sie mindestens ein numerisches Feld aus der angezeigten Liste aus.

Abbildung 6-30
Transformationsknoten: Registerkarte "Felder"

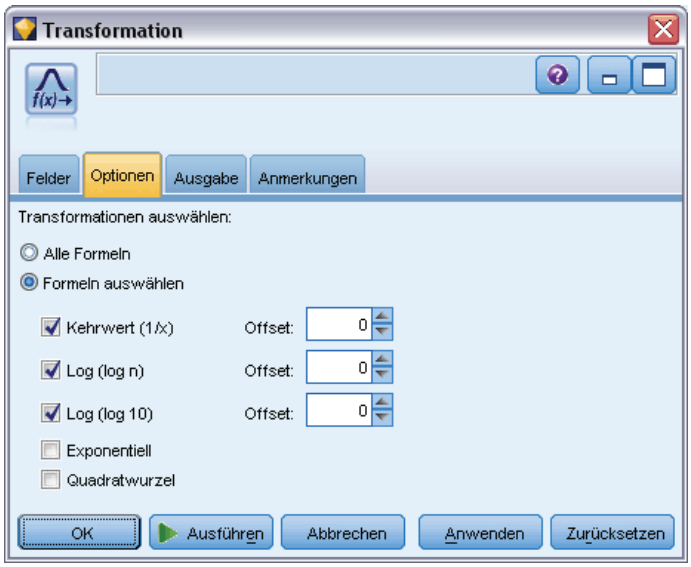
Registerkarte "Optionen" beim Transformationsknoten

Auf der Registerkarte "Optionen" können Sie den Typ der einzuschließenden Transformationen angeben. Sie können auswählen, dass alle verfügbaren Transformationen eingeschlossen werden sollen, oder alle Transformationen einzeln auswählen.

Im letzteren Fall können Sie außerdem einen Wert für den Offset der Daten für Kehrwert- und Logarithmustransformationen eingeben. Dies ist nützlich in Situationen, bei denen ein großer Anteil von Nullen in den Daten die Ergebnisse für Mittelwert und Standardabweichung verzerren würde.

Beispiel: Angenommen, Sie haben ein Feld namens *BALANCE* mit einigen 0-Werten und möchten die Kehrwerttransformation darauf anwenden. Um unerwünschte Verzerrungen zu vermeiden, wählen Sie Kehrwert (1/x) aus und geben im Feld Daten-Offset verwenden den Wert 1 ein. (Beachten Sie, dass dieses Offset nichts mit dem Offset zu tun hat, das durch die Sequenzfunktion @OFFSET in IBM® SPSS® Modeler ausgeführt wird.)

Abbildung 6-31
Transformationsknoten: Registerkarte "Optionen"



Alle Formeln. Gibt an, dass alle verfügbaren Transformationen berechnet und in der Ausgabe angezeigt werden sollen.

Formeln auswählen. Ermöglicht die Auswahl anderer Transformationen für Berechnung und Anzeige in der Ausgabe.

- **Kehrwert (1/x).** Gibt an, dass die Kehrwerttransformation berechnet und in der Ausgabe angezeigt werden soll.
- **Log (log n).** Gibt an, dass die Transformation $\log_n$ berechnet und in der Ausgabe angezeigt werden soll.
- **Log (log 10).** Gibt an, dass die Transformation $\log_{10}$ berechnet und in der Ausgabe angezeigt werden soll.
- **Exponentialverteilung** Gibt an, dass die exponentielle Transformation (e^x) berechnet und in der Ausgabe angezeigt werden soll.
- **Quadratwurzel.** Gibt an, dass die Quadratwurzeltransformation berechnet und in der Ausgabe angezeigt werden soll.

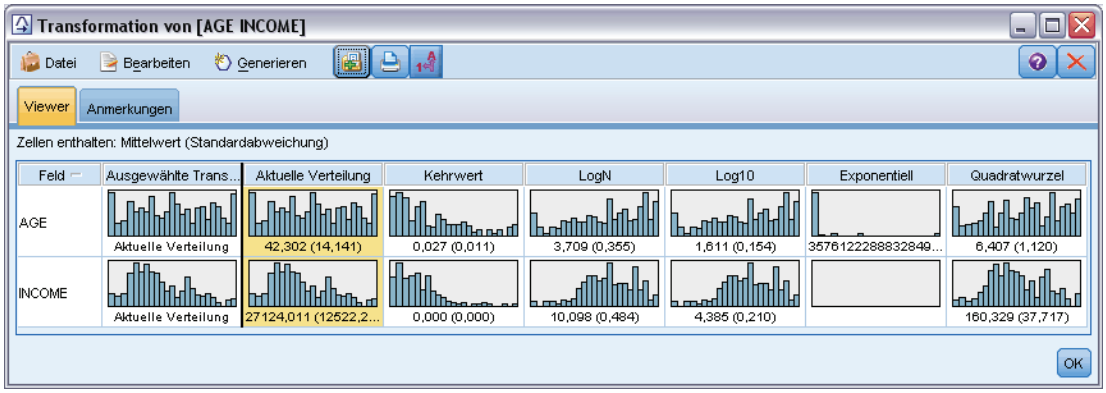
Registerkarte "Ausgabe" beim Transformationsknoten

Auf der Registerkarte "Ausgabe" legen Sie Format und Position der Ausgabe fest. Sie können auswählen, dass die Ergebnisse auf dem Bildschirm angezeigt werden sollen, oder sie an einen der Standarddateitypen senden. Für weitere Informationen siehe Thema Registerkarte "Ausgabe" beim Ausgabeknoten auf S. 406.

Transformationsknoten – Ausgabe-Viewer

Im Ausgabe-Viewer werden die Ergebnisse aus der Ausführung des Transformationsknotens angezeigt. Der Viewer ist ein leistungsstarkes Tool, das mehrere Transformationen pro Feld in Miniaturansichten der Transformation anzeigt, sodass ein schneller Vergleich der Felder möglich ist. Mit den Optionen im zugehörigen Menü "Datei" können Sie die Ausgaben speichern, exportieren bzw. drucken. Für weitere Informationen siehe Thema Anzeigen der Ausgabe auf S. 399.

Abbildung 6-32
Anzeige der verfügbaren Transformationen pro Feld



Für jede Transformation (mit Ausnahme von "Ausgewählte Transformation") wird unterhalb eine Legende mit folgendem Format angezeigt:

Mean (Standard deviation)

Generieren von Knoten für die Transformationen

Der Ausgabe-Viewer bildet einen soliden Ausgangspunkt für die Vorbereitung der Daten. Beispielsweise können Sie das Feld *ALTER* normalisieren, um die Möglichkeit zu erhalten, ein Scoring-Verfahren (z. B. logistische Regression oder Diskriminanzanalyse) anzuwenden, das eine Normalverteilung voraussetzt. Auf der Grundlage der ursprünglichen Diagramme und Übersichtsstatistiken könnten Sie sich entscheiden, das Feld *ALTER* gemäß einer bestimmten Verteilung zu transformieren (z. B. .log). Nach Auswahl der bevorzugten Verteilung können Sie anschließend einen Ableitungsknoten mit einer standardisierten Transformation für die Verwendung beim Scoring generieren.

Sie können folgende Feldoperationsknoten aus dem Ausgabe-Viewer generieren:

- Ableiten
- Füller

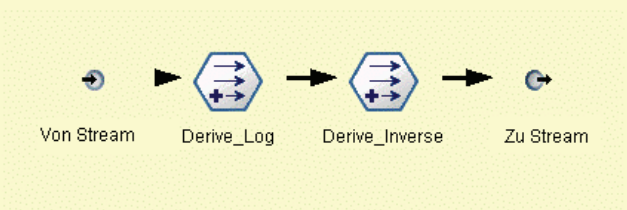
Ein Ableitungsknoten erstellt neue Felder mit den gewünschten Transformationen, während der Füllerknoten bestehende Felder transformiert. Die Knoten werden in Form eines Superknotens im Zeichenbereich platziert.

Wenn Sie dieselbe Transformation für verschiedene Felder auswählen, enthält ein Ableitungs- bzw. Füllerknoten die Formeln für den betreffenden Transformationstyp für alle Felder, für die die Transformation gilt. Nehmen wir beispielsweise an, dass Sie folgende Felder und Transformationen ausgewählt haben, um einen Ableitungsknoten zu generieren:

Feld	Transformation
ALTER	Aktuelle Verteilung
EINKOMMEN	Log
OPEN_BAL	Inverse
BALANCE	Inverse

Folgende Knoten sind im Superknoten enthalten:

Abbildung 6-33
Superknoten im Zeichenbereich



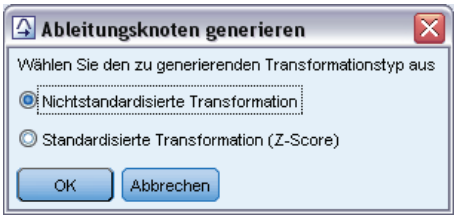
In diesem Beispiel weist der Knoten Derive_Log die log-Formel für das Feld *EINKOMMEN* und der Knoten Derive_Inverse weist die Kehrwertformeln für die Felder *OPEN_BAL* und *BALANCE* auf.

So generieren Sie einen Knoten:

- Wählen Sie für jedes Feld im Ausgabe-Viewer die gewünschte Transformation.
- Wählen Sie im Menü “Generieren” die Option Ableitungsknoten bzw. Füllerknoten.

Auf diese Weise wird das Dialogfeld “Ableitungsknoten generieren” bzw. “Füllerknoten generieren” angezeigt.

Abbildung 6-34
Auswählen einer standardisierten bzw. nichtstandardisierten Transformation



Wählen Sie Nichtstandardisierte Transformation bzw. Standardisierte Transformation (Z-Score). Die zweite Option wendet einen *z*-Score auf die Transformation an; *z*-Scores stellen Werte als Funktion der Distanz vom Mittelwert der Variablen in Standardabweichungen dar. Wenn Sie beispielsweise

die logarithmische Transformation auf das Feld *ALTER* anwenden und eine standardisierte Transformation auswählen, lautet die endgültige Gleichung für den generierten Knoten wie folgt:

```
(log(AGE)-Mean)/SD
```

Gehen Sie wie folgt vor, sobald ein Knoten generiert wurde und im Stream-Zeichenbereich angezeigt wird:

- Fügen Sie ihn zum Stream hinzu.
- Bei einem Superknoten können Sie optional auf den Knoten doppelklicken, um seinen Inhalt anzuzeigen.
- Optional können Sie auf einen Ableitungs- oder Füllerknoten doppelklicken, um die Optionen für die ausgewählten Felder zu ändern.

Erzeugen von Diagrammen

Sie können aus einem Miniaturhistogramm im Ausgabe-Viewer Histogrammausgaben in normaler Größe generieren.

So generieren Sie ein Diagramm:

- Doppelklicken Sie auf ein Miniaturdiagramm im Ausgabe-Viewer.

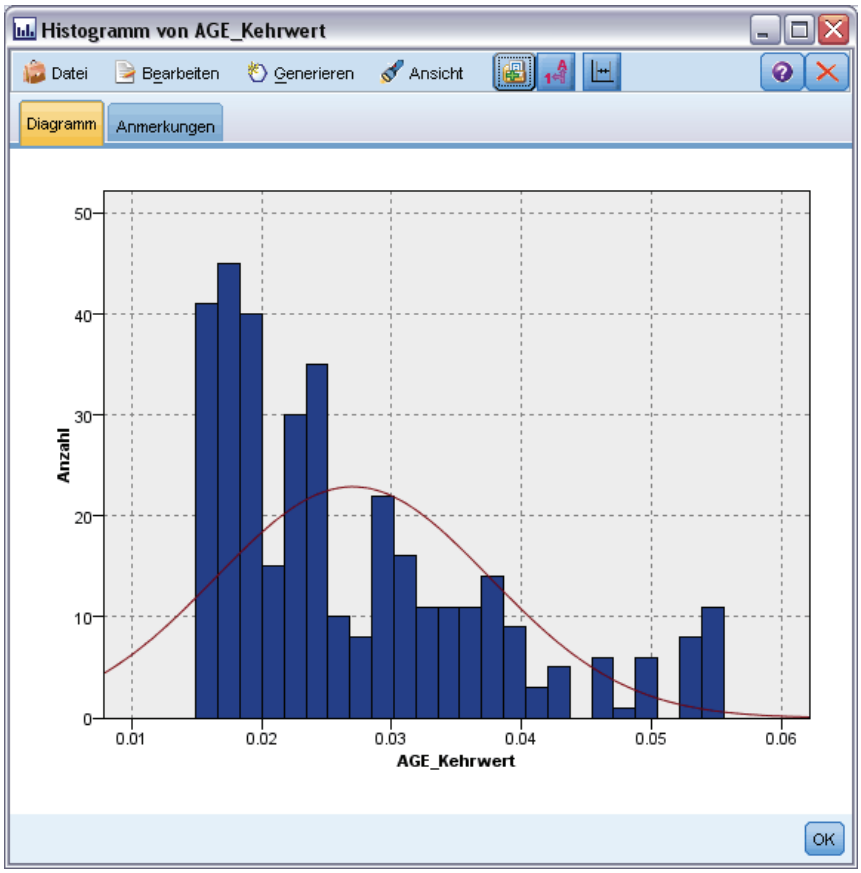
  oder

- Wählen Sie ein Miniaturdiagramm im Ausgabe-Viewer.
- Wählen Sie im Menü “Generieren” den Befehl Diagrammausgabe.

Dadurch wird das Histogramm mit einer überlagerten Normalverteilungskurve angezeigt. So können Sie vergleichen, wie eng die einzelnen Transformationen mit einer Normalverteilung übereinstimmen.

Hinweis: Sie können auch ein Diagramm erzeugen, wenn der Transformationsknoten, der die Ausgabe erzeugt hat, mit dem Stream verbunden ist.

Abbildung 6-35
Transformationshistogramm mit überlagerter Normalverteilungskurve



Andere Operationen

Im Ausgabe-Viewer haben Sie außerdem folgende Möglichkeiten:

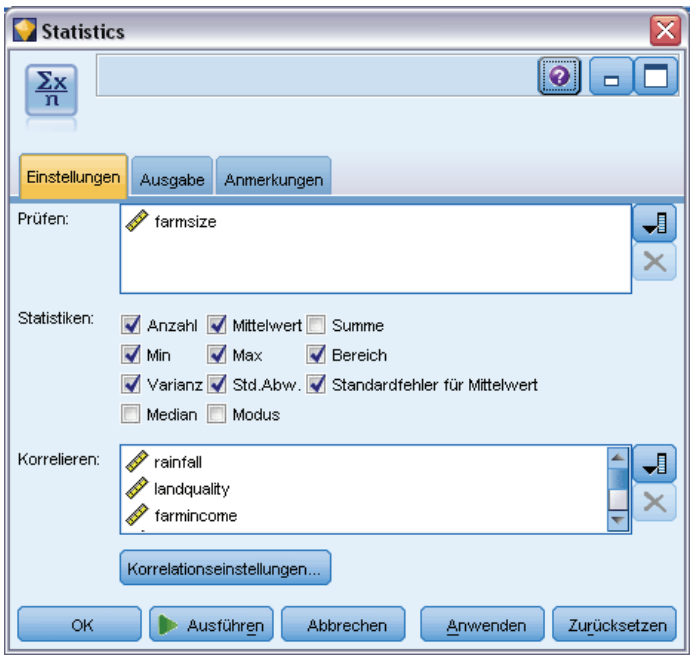
- Sortieren des Ausgabegitters nach der Spalte “Feld”.
- Exportieren der Ausgabe in eine HTML-Datei. Für weitere Informationen siehe Thema Exportieren von Ausgaben auf S. 403.

Statistikknoten

Der Statistikknoten liefert grundlegende Übersichtsdaten zu numerischen Feldern. Die Übersichtsstatistiken können für einzelne Felder und für die Korrelationen zwischen den Feldern abgerufen werden.

Registerkarte "Einstellungen" beim Statistikknoten

Abbildung 6-36
Statistikknoten: Registerkarte "Einstellungen"



Prüfen. Wählen Sie das oder die Felder aus, für die eine separate Übersichtsstatistik zusammengestellt werden soll. Sie können mehrere Felder auswählen.

Statistiken. Wählen Sie die zu bildenden Statistiken aus. Die folgenden Optionen stehen zur Auswahl: Anzahl, Mittelwert, Summe, Min, Max, Bereich, Varianz, Std.Abw., Standardfehler für Mittelwert, Median und Modus.

Korrelieren. Wählen Sie das oder die zu korrelierenden Felder aus. Sie können mehrere Felder auswählen. Wenn Sie Korrelationsfelder auswählen, wird die Korrelation zwischen jedem Untersuchungsfeld und den Korrelationsfeldern in der Ausgabe aufgeführt.

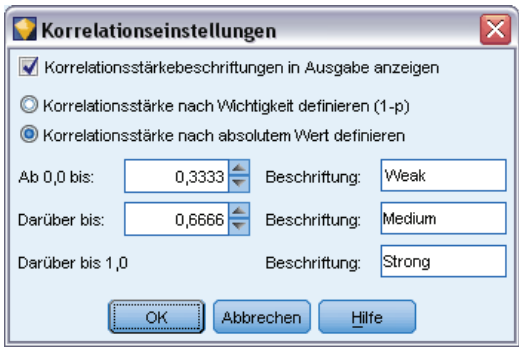
Korrelationseinstellungen. Sie können Optionen zur Anzeige der Korrelationsstärke in der Ausgabe angeben.

Korrelationseinstellungen

Bei IBM® SPSS® Modeler können Korrelationen mit deskriptiven Beschriftungen versehen werden, um so wichtige Beziehungen hervorzuheben. Die **Korrelation** bezeichnet die Stärke der Beziehung zwischen zwei stetigen Feldern (numerischer Bereich). Zulässige Werte: –1,0 bis 1,0. Werte nahe +1,0 weisen auf eine starke positive Assoziation hin; dies bedeutet, dass hohe Werte in einem Feld mit hohen Werten im zweiten Feld verknüpft sind und entsprechend niedrige Werte mit niedrigen Werten. Werte nahe –1,0 weisen auf eine starke negative Assoziation hin; dies bedeutet, dass hohe Werte in einem Feld mit niedrigen Werten im zweiten Feld verknüpft sind und umgekehrt. Werte nahe 0,0 weisen auf eine schwache Assoziation hin; dies bedeutet, dass die Werte der beiden Felder relativ unabhängig voneinander sind.

Sie können die Anzeige der Korrelationsbeschriftungen festlegen, die Schwellenwerte ändern, die die Kategorien definieren, und die für die einzelnen Bereiche verwendeten Beschriftungen ändern. Die Charakterisierung der Korrelationswerte ist stark abhängig von der Problemdomäne. Aus diesem Grund sollten Sie die Bereiche und Beschriftungen an die jeweils gegebene Situation anpassen.

Abbildung 6-37
Dialogfeld "Korrelationseinstellungen"



Korrelationsstärkebeschriftungen in Ausgabe anzeigen. Diese Option ist standardmäßig aktiviert. Wenn die deskriptiven Beschriftungen nicht in die Ausgabe aufgenommen werden sollen, deaktivieren Sie diese Option.

Korrelationsstärke. Es gibt zwei Optionen zur Definition und Beschriftung der Korrelationsstärke:

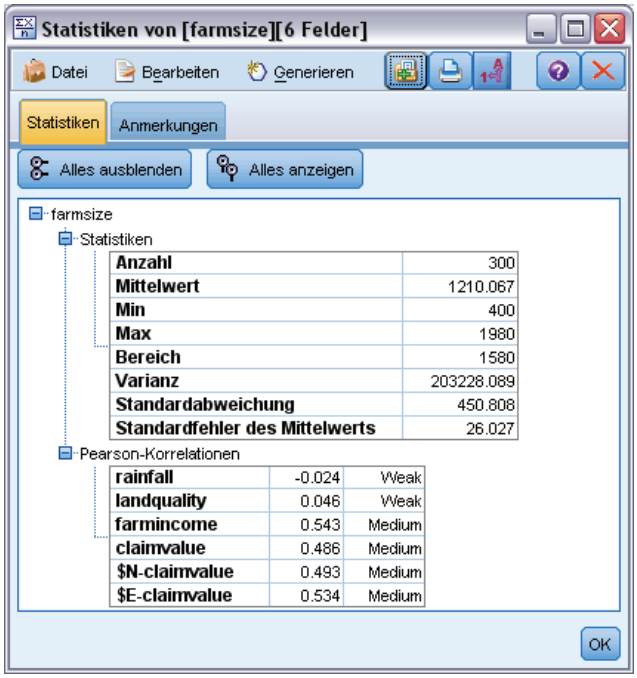
- **Korrelationsstärke nach Wichtigkeit definieren (1-p).** Gibt die Korrelationen auf der Grundlage der Wichtigkeit an, definiert als 1 minus Signifikanz oder 1 minus die Wahrscheinlichkeit, dass sich die Differenz bei den Mittelwerten durch den Zufall allein erklären lässt. Je näher dieser Wert bei 1 liegt, desto größer ist die Wahrscheinlichkeit, dass die beiden Felder *nicht* unabhängig sind, dass sie also in einer gewissen Beziehung zueinander stehen. Die Angabe der Korrelationen auf der Grundlage der Wichtigkeit wird normalerweise gegenüber dem absoluten Wert empfohlen, da hierbei die Variabilität in den Daten berücksichtigt wird. So kann es beispielsweise sein, das ein Koeffizient von 0,6 in einem Daten-Set extrem signifikant ist, in einem anderen jedoch ganz und gar nicht. Standardmäßig werden Wichtigkeitswerte zwischen 0,0 und 0,9 als *Schwach* beschriftet, Wichtigkeitswerte zwischen 0,9 und 0,95 als *Mittel* und Wichtigkeitswerte zwischen 0,95 und 1,0 entsprechend als *Stark*.
- **Korrelationsstärke nach absolutem Wert definieren.** Beschriftet Korrelationen auf der Grundlage des absoluten Werts des Korrelationskoeffizienten nach Pearson, der zwischen –1 und 1 liegt, wie oben beschrieben. Je näher der absolute Wert dieses Maßes bei 1 liegt, desto stärker ist die Korrelation. Standardmäßig werden Korrelationen zwischen 0,0 und 0,3333 (absoluter Wert) als *Schwach* beschriftet, Korrelationen zwischen 0,3333 und 0,6666 als *Mittel* und Korrelationen zwischen 0,6666 und 1,0 entsprechend als *Stark*. Beachten Sie jedoch, dass sich die Signifikanz eines Werts schlecht über mehrere Daten-Sets hinweg generalisieren lässt; aus diesem Grund wird in den meisten Fällen empfohlen, Korrelationen auf der Grundlage der Wahrscheinlichkeit und nicht auf der Grundlage ihres absoluten Werts zu definieren.

Statistikausgabe-Browser

Der Ausgabe-Browser des Statistik-Knotens enthält die Ergebnisse der statistischen Analyse und ermöglicht die Ausführung verschiedener Funktionen, z. B. Felder auswählen, neue Knoten auf der Grundlage der Auswahl erzeugen sowie die Ergebnisse speichern und drucken. Das Menü "Datei" enthält die üblichen Befehle zum Speichern, Exportieren und Drucken, das Menü "Bearbeiten" die üblichen Bearbeitungsfunktionen. Für weitere Informationen siehe Thema Anzeigen der Ausgabe auf S. 399.

Beim Öffnen des Statistikausgabe-Browsers werden die Ergebnisse erweitert. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement links neben dem gewünschten Element die Ergebnisse reduzieren. Alternativ können Sie mit der Schaltfläche Alles ausblenden alle Ergebnisse ausblenden. Um die für Sie relevanten Ergebnisse nach dem Reduzieren wieder anzeigen zu lassen, erweitern Sie die gewünschten Ergebnisse mithilfe des Erweiterungssteuerelements auf der linken Seite oder klicken Sie auf die Schaltfläche Alles anzeigen, um alle Ergebnisse anzuzeigen.

Abbildung 6-38
Statistikausgabe-Browser



Die Ausgabe umfasst je einen Abschnitt für die einzelnen Felder *Prüfen* mit einer Tabelle der angeforderten Statistiken.

- **Häufigkeiten.** Anzahl der Datensätze mit gültigen Werten für das Feld.
- **Mittelwert.** Durchschnittlicher Wert (Mittelwert) für das Feld über alle Datensätze hinweg.
- **Summe.** Summe der Werte für das Feld über alle Datensätze hinweg.
- **Min.** Mindestwert für das Feld.
- **Max.** Höchstwert für das Feld.

- **Bereich.** Differenz zwischen Mindest- und Höchstwert.
- **Varianz.** Maß für die Schwankungen der Werte eines Felds. Dieser Wert wird wie folgt berechnet: Zunächst wird die Differenz zwischen jedem Wert und dem Gesamtmittelwert ermittelt. Diese Differenz wird quadriert, über alle Werte summiert und dann durch die Anzahl der Datensätze dividiert.
- **Standardabweichung.** Ein weiteres Maß für die Schwankungen der Werte eines Felds, berechnet als die Quadratwurzel der Varianz.
- **Standardfehler des Mittelwerts.** Maß für die Unsicherheit bei der Schätzung des Mittelwerts für das Feld, falls der Mittelwert vermutungsgemäß für neue Daten gilt.
- **Median.** "Mittlerer" Wert für das Feld, also der Wert, der die obere Hälfte der Daten von der unteren Hälfte trennt (auf der Grundlage der Werte des Felds).
- **Modalwert.** Am häufigsten auftretender einzelner Wert in den Daten.

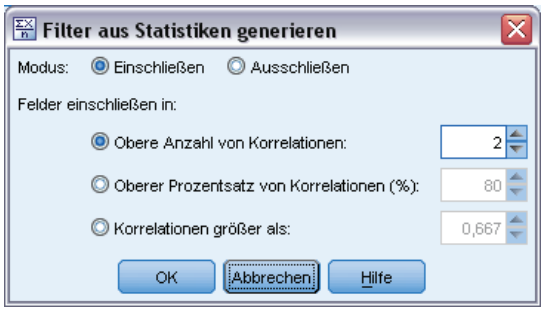
Korrelationen. Wenn Sie Korrelationsfelder angegeben haben, enthält die Ausgabe außerdem einen Abschnitt, in dem die Korrelation nach Pearson zwischen dem Prüffeld und jedem Korrelationsfeld aufgeführt wird und auch die optionalen deskriptiven Beschriftungen für die Korrelationswerte genannt werden. Für weitere Informationen siehe Thema Korrelationseinstellungen auf S. 443.

Menü "Generieren". Das Menü "Generieren" enthält Funktionen zum Erzeugen von Knoten.

- **Filter.** Erzeugt einen Filterknoten, mit dem die Felder herausgefiltert werden, für die keine oder nur eine schwache Korrelation mit anderen Feldern besteht.

Erzeugen eines Filterknotens aus den Statistiken heraus

Abbildung 6-39
Dialogfeld "Filter aus Statistiken generieren"



Mit dem aus einem Statistikausgabe-Browser heraus erzeugten Filterknoten werden die Felder auf der Grundlage ihrer Korrelation mit anderen Feldern gefiltert. Hierzu werden die Korrelationen nach dem absoluten Wert sortiert. Anschließend werden die größten Korrelationen ermittelt (gemäß dem im Dialogfeld festgelegten Kriterium); dann wird ein Filter erstellt, der alle Felder passieren lässt, die in einer dieser großen Korrelationen auftreten.

Modalwert. Legen Sie fest, auf welche Weise die Korrelationen ausgewählt werden sollen. Mit der Option Einschließen werden alle Felder beibehalten, die in den angegebenen Korrelationen auftreten. Mit Ausschließen werden die Felder gefiltert.

Felder einschließen in/Felder ausschließen in: Definieren Sie das Kriterium zum Auswählen der Korrelationen.

- **Obere Anzahl von Korrelationen.** Die angegebene Anzahl an Korrelationen wird ausgewählt; Felder, die in einer dieser Korrelationen auftreten, werden dabei eingeschlossen bzw. ausgeschlossen.
- **Oberer Prozentsatz von Korrelationen (%).** Der angegebene Prozentsatz (*n %*) an Korrelationen wird ausgewählt; Felder, die in einer dieser Korrelationen auftreten, werden dabei eingeschlossen bzw. ausgeschlossen.
- **Korrelationen größer als.** Hiermit werden Korrelationen ausgewählt, deren absoluter Wert größer ist als der angegebene Schwellenwert.

Mittelwertknoten

Der Mittelwertknoten vergleicht die Mittelwerte zwischen unabhängigen Gruppen oder zwischen Paaren von in Bezug stehenden Feldern, um zu testen, ob ein signifikanter Unterschied vorliegt. So können Sie beispielsweise die Einnahmen vor und nach der Durchführung einer Werbeaktion vergleichen oder die Einnahmen, die von Kunden stammen, die keine Werbezettel erhielten, mit den Einnahmen von Kunden vergleichen, die von der Werbeaktion erreicht wurden.

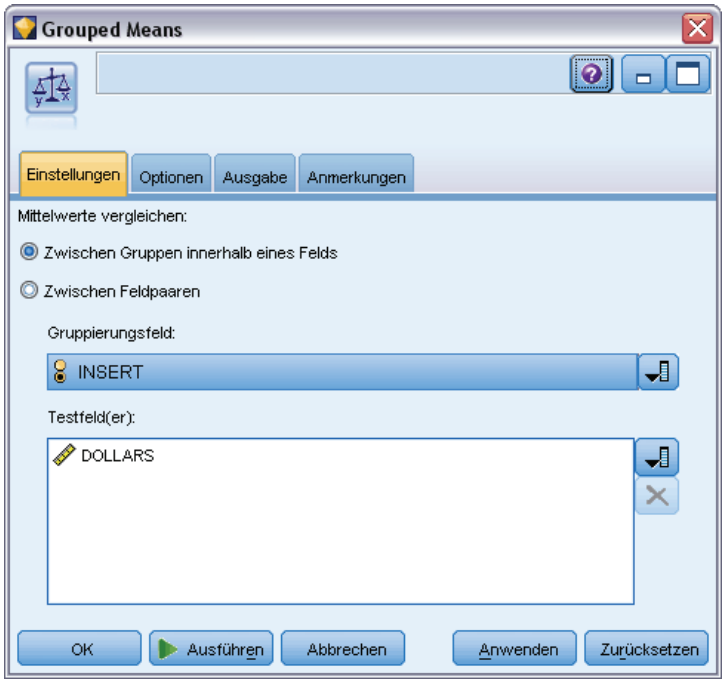
Sie können Mittelwerte auf zwei verschiedene Weisen vergleichen, je nach den verwendeten Daten:

- **Zwischen Gruppen innerhalb eines Felds.** Um unabhängige Gruppen vergleichen zu können, wählen Sie ein Testfeld und ein Gruppierungsfeld aus. Sie können beispielsweise eine Stichprobe von “Standhalte”-Kunden beim Versenden von Werbesendungen ausschließen und die durchschnittlichen Einnahmen durch die Standhalte-Gruppe mit den anderen Kunden vergleichen. In diesem Fall geben Sie ein einzelnes Testfeld an, das die Einnahmen für jeden Kunden anzeigt, sowie ein Flag- bzw. nominales Feld, das angibt, ob der jeweilige Kunde das Angebot erhalten hat. Die Stichproben sind unabhängig in dem Sinn, dass jeder Datensatz entweder der einen oder anderen Gruppe zugewiesen wird und es nicht möglich ist, ein bestimmtes Mitglied einer Gruppe einem bestimmten Mitglied einer anderen Gruppe zuzuweisen. Sie können auch ein nominales Feld mit mehr als zwei Werten angeben, um die Mittelwerte für mehrere Gruppen zu vergleichen. Bei der Ausführung berechnet der Knoten einen einfachen ANOVA-Test für die ausgewählten Felder. In Fällen, in denen es nur zwei Feldgruppen gibt, stimmen die Ergebnisse der einfachen ANOVA im Wesentlichen mit denen eines *t*-Tests mit unabhängigen Stichproben überein. Für weitere Informationen siehe Thema Vergleich der Mittelwerte für unabhängige Gruppen auf S. 448.
- **Zwischen Feldpaaren.** Beim Vergleich der Mittelwerte für verwandte Felder müssen die Gruppen gepaart werden, um aussagekräftige Ergebnisse zu erhalten. Sie könnten beispielsweise den Mittelwert der Einnahmen aus derselben Gruppe von Kunden vor und nach der Durchführung einer Werbeaktion vergleichen oder die Nutzungsquoten für einen Service zwischen Paaren aus Ehemann und Ehefrau vergleichen, um zu sehen, ob Unterschiede vorliegen. Jeder Datensatz enthält zwei verschiedene, jedoch miteinander verwandte Maße, bei denen ein aussagekräftiger Vergleich möglich ist. Bei der Ausführung berechnet der Knoten einen *t*-Test mit gepaarten Stichproben für jedes ausgewählte Feldpaar. Für weitere Informationen siehe Thema Vergleich der Mittelwerte zwischen gepaarten Feldern auf S. 448.

Vergleich der Mittelwerte für unabhängige Gruppen

Wählen Sie Zwischen Gruppen innerhalb eines Felds im Mittelwertknoten, um die Mittelwerte für zwei oder mehrere unabhängige Gruppen miteinander zu vergleichen.

Abbildung 6-40
Vergleichen der Mittelwerte zwischen Gruppen innerhalb eines Felds



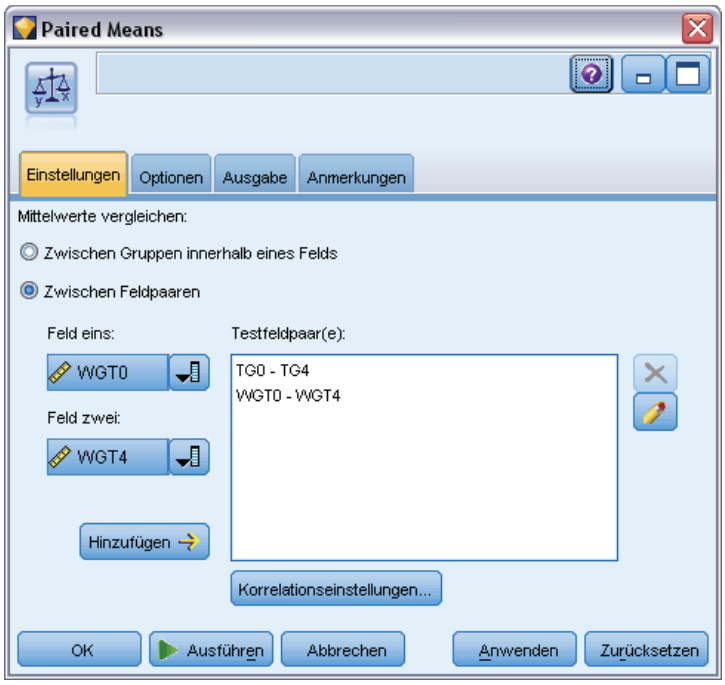
Gruppierungsfeld. Wählen Sie ein numerisches Flag- oder nominales Feld mit zwei oder mehr verschiedenen Werten aus, das Datensätze in die Gruppen einteilt, die verglichen werden sollen, beispielsweise in die Gruppe der Personen, die ein Angebot erhalten haben, und die Gruppe von Personen, bei denen dies nicht der Fall ist. Unabhängig von der Anzahl der Testfelder kann nur ein Gruppierungsfeld ausgewählt werden.

Testfelder. Wählen Sie mindestens ein numerisches Feld aus, das die zu testenden Maße enthält. Für jedes ausgewählte Feld wird ein separater Test ausgeführt. Sie können beispielsweise die Auswirkungen einer bestimmten Werbeaktion auf Nutzung, Ertrag und Abwanderung testen.

Vergleich der Mittelwerte zwischen gepaarten Feldern

Wählen Sie im Mittelwertknoten die Option Zwischen Feldpaaren, um die Mittelwerte zwischen unterschiedlichen Feldern zu vergleichen. Die Felder müssen in einem bestimmten Bezug zueinander stehen, damit die Ergebnisse aussagekräftig sind, beispielsweise die Einnahmen vor und nach einer Werbeaktion. Es können auch mehrere Feldpaare ausgewählt werden.

Abbildung 6-41
Vergleich der Mittelwerte zwischen gepaarten Feldern



Feld eins. Wählen Sie ein numerisches Feld aus, das das erste Maß enthält, das verglichen werden soll. In einer Vorher-Nachher-Studie wäre dies das Feld "Vorher".

Feld zwei. Wählen Sie das zweite Feld für den Vergleich aus.

Hinzufügen. Fügt das ausgewählte Paar zur Liste der Testfeldpaare hinzu.

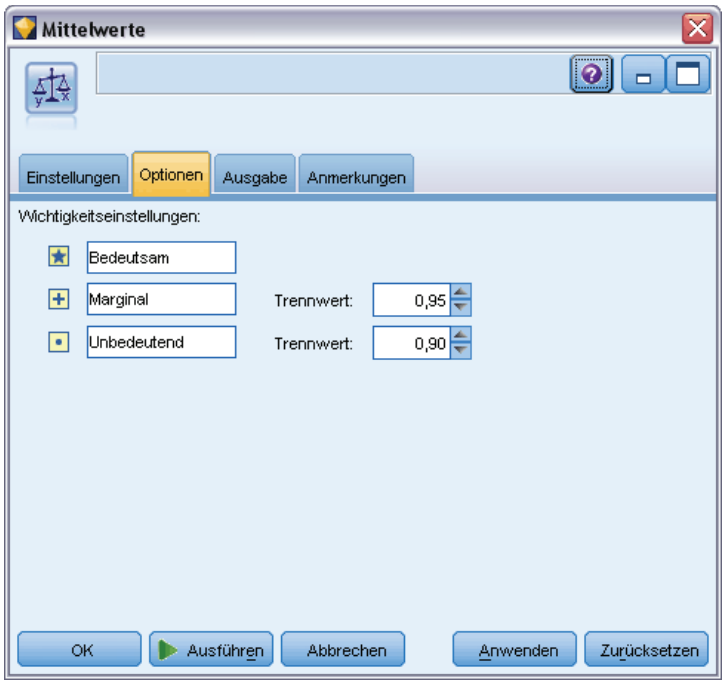
Wiederholen Sie die Feldauswahl nach Bedarf, um mehrere Paare zur Liste hinzuzufügen.

Korrelationseinstellungen. Ermöglicht die Angabe von Optionen zur Angabe der Korrelationsstärke. Für weitere Informationen siehe Thema Korrelationseinstellungen auf S. 443.

Mittelwertknoten – Optionen

Auf der Registerkarte "Optionen" können Sie Schwellenwert-*p*-Werte festlegen, die verwendet werden, um Ergebnisse als "Bedeutsam", "Marginal" oder "Unbedeutend" zu beschriften. Außerdem können Sie die Beschriftung für jede Einstufung bearbeiten. Die Wichtigkeit wird auf einer Prozentskala gemessen und lässt sich grob wie folgt definieren: 1 minus die Wahrscheinlichkeit, ein Ergebnis (beispielsweise die Differenz der Mittelwerte zwischen zwei Feldern) zu erhalten, das nur allein Zufall mindestens so extrem ist wie das beobachtete Ergebnis. Ein *p*-Wert größer als 0,95 beispielsweise zeigt an, dass eine Wahrscheinlichkeit von weniger als 5 % besteht, dass sich das Ergebnis allein durch Zufall erklären lässt.

Abbildung 6-42
Wichtigkeitseinstellungen



Wichtigkeitsbeschriftungen. Sie können die Beschriftungen für die einzelnen Feldpaare bzw. -gruppen in der Ausgabe bearbeiten. Die Standardbeschriftungen lauten *Bedeutsam*, *Marginal* und *Unbedeutend*.

Cutoff-Werte. Geben den Schwellenwert für jeden Rang an. Üblicherweise werden *p*-Werte über 0,95 als bedeutsam eingestuft, Werte unter 0,9 als unbedeutend. Diese Schwellenwerte lassen sich jedoch nach Bedarf anpassen.

Anmerkung: Wichtigkeitsmaße sind in einer Reihe von Knoten verfügbar. Die speziellen Berechnungen hängen vom Knoten und vom verwendeten Ziel- und Eingabefeldtyp ab, die Werte können jedoch weiterhin verglichen werden, da sie auf einer Prozentskala gemessen werden.

Mittelwertknoten – Ausgabe-Browser

Der Mittelwert-Ausgabe-Browser zeigt Daten als Kreuztabellen an und ermöglicht die Ausführung von Standardoperationen wie Auswählen und Kopieren der Tabelle Zeile für Zeile, Sortieren nach einer beliebigen Spalte sowie Speichern und Drucken der Tabelle. Für weitere Informationen siehe Thema Anzeigen der Ausgabe auf S. 399.

Die spezifischen Informationen in der Tabelle hängen vom Vergleichstyp (Gruppen innerhalb eines Felds oder gesonderte Felder) ab.

Sortieren nach. Ermöglicht die Sortierung der Ausgabe nach einer bestimmten Spalte. Klicken Sie auf den nach oben bzw. nach unten weisenden Pfeil, um die Sortierrichtung zu ändern. Alternativ können Sie auf die Überschrift einer Spalte klicken, um eine Sortierung nach dieser

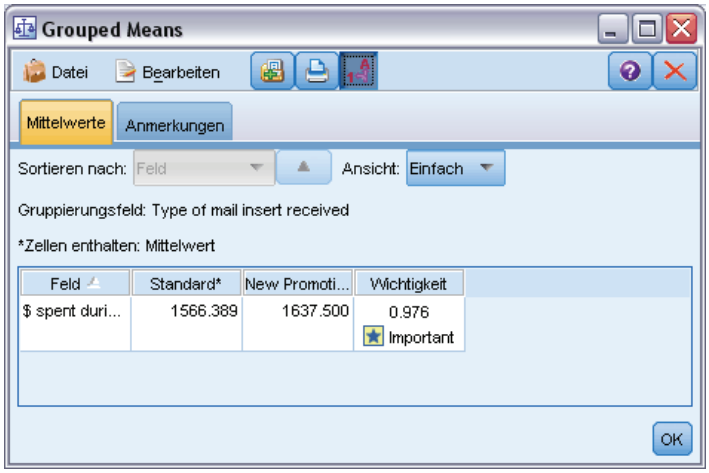
Spalte vorzunehmen. (Wenn Sie die Sortierrichtung innerhalb der Spalte ändern möchten, klicken Sie noch einmal.)

Ansicht. Sie haben die Wahl zwischen Einfach und Erweitert, um den Detailliertheitsgrad der Anzeige zu steuern. Die erweiterte Ansicht enthält alle Informationen der einfachen Ansicht sowie weitere Einzelheiten.

Mittelwertausgabe zum Vergleich von Gruppen innerhalb eines Felds

Beim Vergleich von Gruppen innerhalb eines Felds wird der Name des Gruppierungsfelds oberhalb der Ausgabetabelle angezeigt; außerdem werden Mittelwerte und verwandte Statistiken separat für die einzelnen Gruppen gemeldet. Die Tabelle enthält eine gesonderte Zeile für jedes Testfeld.

Abbildung 6-43
Vergleichen von Gruppen innerhalb eines Felds



Folgende Spalten werden angezeigt:

- **Feld.**Hier werden die Namen der ausgewählten Testfelder angegeben.
- **Mittelwerte nach Gruppe.** Zeigt den Mittelwert für die einzelnen Kategorien des Gruppierungsfelds an. Sie könnten beispielsweise die Personen, die ein Sonderangebot (*New Promotion*) erhalten haben, mit denen vergleichen, die keines erhalten haben (*Standard*). In der erweiterten Ansicht werden außerdem Standardabweichung, Standardfehler und Anzahl angezeigt.
- **Wichtigkeit.** Zeigt Wert und Beschriftung für die Wichtigkeit an. Für weitere Informationen siehe Thema Mittelwertknoten – Optionen auf S. 449.

Erweiterte Ausgabe

In der erweiterten Ansicht werden folgende zusätzlichen Spalten angezeigt.

- **F-Test.** Dieser Test beruht auf dem Quotienten aus der Varianz zwischen den Gruppen und der Varianz innerhalb der einzelnen Gruppen. Wenn die Mittelwerte für alle Gruppen gleich sind, ist zu erwarten, dass das *F*-Verhältnis nahe bei 1 liegt, da beides Schätzungen derselben

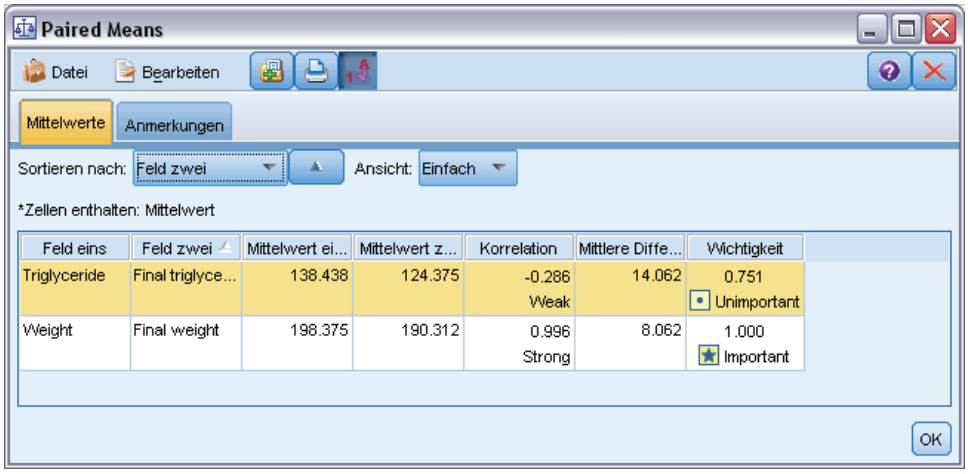
Populationsvarianz sind. Je größer dieser Quotient, desto größer ist die Variation zwischen den Gruppen und desto größer ist die Wahrscheinlichkeit, dass eine signifikante Differenz vorliegt.

- **df.** Zeigt die Freiheitsgrade an.

Mittelwertausgabe zum Vergleich von Feldpaaren

Beim Vergleich zwischen verschiedenen Feldern enthält die Ausgabetabelle eine Zeile für jedes ausgewählte Feldpaar.

Abbildung 6-44
Vergleich zwischen Feldpaaren



- **Feld eins/zwei.** Zeigt den Namen des ersten und zweiten Felds in jedem Paar an. In der erweiterten Ansicht werden außerdem Standardabweichung, Standardfehler und Anzahl angezeigt.
- **Mittelwert eins/zwei.** Zeigt den Mittelwert für das jeweilige Feld an.
- **Korrelation.** Misst die Stärke der Beziehung zwischen zwei stetigen Feldern (numerischer Bereich). Werte in der Nähe von +1,0 deuten auf eine starke positive Assoziation hin und Werte in der Nähe von –1,0 deuten auf eine starke negative Assoziation hin. Für weitere Informationen siehe Thema Korrelationseinstellungen auf S. 443.
- **Mittlere Differenz.** Zeigt die Differenz zwischen den beiden Feldmittelwerten an.
- **Wichtigkeit.** Zeigt Wert und Beschriftung für die Wichtigkeit an. Für weitere Informationen siehe Thema Mittelwertknoten – Optionen auf S. 449.

Erweiterte Ausgabe

Bei der erweiterten Ausgabe sind folgende Spalten hinzugefügt:

95 %-Konfidenzintervall. Unter- und Obergrenze des Bereichs, in dem der wahre Mittelwert statistisch gesehen in 95 % aller möglichen Stichproben dieser Größe aus dieser Grundgesamtheit fällt.

T-Test. Die *t*-Statistik wird berechnet, indem die mittlere Differenz durch ihren Standardfehler dividiert wird. Je größer der absolute Wert dieser Statistik, desto größer ist die Wahrscheinlichkeit, dass die Mittelwerte nicht identisch sind.

df. Zeigt die Freiheitsgrade für die Statistik an.

Berichtknoten

Mit dem Berichtknoten erstellen Sie formatierte Berichte, die sowohl festen Text als auch Daten und andere aus den Daten abgeleitete Ausdrücke enthält. Das Format des Berichts wird mithilfe von Textvorlagen festgelegt, mit denen der feste Text und die Datenausgabekonstruktionen definiert werden. Sie können eine benutzerdefinierte Textformatierung angeben; hierzu stehen HTML-Tags in der Vorlage sowie Optionen auf der Registerkarte "Ausgabe" zur Verfügung. Datenwerte und andere bedingte Ausgaben werden mithilfe von CLEM-Ausdrücken in der Vorlage in den Bericht aufgenommen.

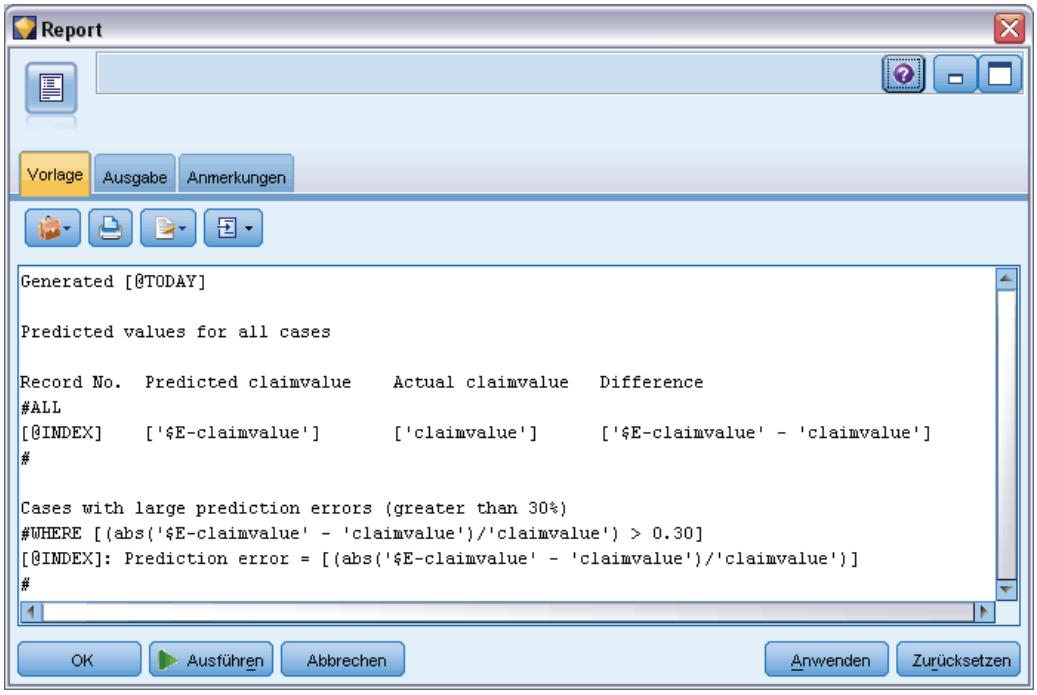
Alternativen zum Berichtknoten

Der Berichtknoten wird normalerweise verwendet, um ausgegebene Datensätze oder Fälle aus einem Stream aufzulisten, beispielsweise alle Datensätze, die eine bestimmte Bedingung erfüllen. In dieser Hinsicht kann er als weniger strukturierte Alternative zum Tabellenknoten betrachtet werden.

- Wenn Sie einen Bericht wünschen, der Feldinformationen oder andere Elemente auflistet, die im Stream definiert wurden, und nicht die Daten selbst (beispielsweise die in einem Typknoten angegebenen Felddefinitionen), kann stattdessen ein Skript verwendet werden.
- Um einen Bericht zu generieren, der mehrere Ausgabeobjekte enthält (z. B. eine Sammlung von Modellen, Tabellen und Diagrammen, die von einem oder mehreren Streams generiert wurden) und in mehreren Formaten (z. B. Textformat, HTML und Microsoft Word/Office) ausgegeben werden kann, können Sie ein IBM® SPSS® Modeler-Projekt verwenden.
- Um eine Liste von Feldnamen ohne Verwendung von Skripts zu erstellen, können Sie einen Tabellenknoten verwenden, dem ein Stichprobenknoten vorangeht, der alle Datensätze verwirft. Auf diese Weise wird eine Tabelle ohne Zeilen erstellt, die beim Export transponiert werden kann, um eine Liste von Feldnamen in einer einzelnen Spalte zu erzeugen. (Wählen Sie dazu im Tabellenknoten auf der Registerkarte "Ausgabe" die Option Daten transponieren.)

Registerkarte "Vorlage" beim Berichtknoten

Abbildung 6-45
Berichtknoten: Vorlage, Registerkarte



Erstellen einer Vorlage. Um den Inhalt des Berichts zu definieren, erstellen Sie eine Vorlage auf der Registerkarte "Vorlage" im Berichtknoten. Die Vorlage besteht aus Textzeilen, die jeweils Angaben zum Inhalt des Berichts enthalten, sowie aus einigen Zeilen mit Sonder-Tags, aus denen der Bereich der Inhaltszeilen hervorgeht. In jeder Inhaltszeile werden zunächst CLEM-Ausdrücke in eckigen Klammern ([]) ausgewertet, bevor die betreffende Zeile an den Bericht gesendet wird. Für die Zeilen in der Vorlage stehen jeweils drei Bereiche zur Auswahl:

Fest. Zeilen, die nicht anderweitig gekennzeichnet sind, werden als fest betrachtet. Feste Zeilen werden nur einmal in den Bericht kopiert, sobald alle in diesen Zeilen enthaltenen Ausdrücke ausgewertet wurden. Beispiel: Mit der Zeile

```
Dies ist mein Bericht, gedruckt am [@TODAY]
```

wird eine einzelne Zeile in den Bericht kopiert, die den angegebenen Text und das aktuelle Datum enthält.

Global (Alles iterieren). Die Zeilen zwischen den Sonder-Tags #ALL und # werden je einmal für jeden Datensatz mit Eingabedaten in den Bericht kopiert. Die CLEM-Ausdrücke (in Klammern) werden auf der Grundlage des jeweils aktuellen Datensatzes für jede Ausgabezeile ausgewertet. Beispiel: Mit den Zeilen

```
#ALL
Beim Datensatz [@INDEX] ist der Wert für ALTER gleich [ALTER]
#
```

wird je eine Zeile für jeden Datensatz eingefügt, aus der die Nummer des Datensatzes und das Alter hervorgeht.

So generieren Sie eine Liste aller Datensätze:

```
#ALL
[Alter]   [Geschl]   [Cholesterol]   [BP]
#
```

Bedingt (Iterieren, wenn). Die Zeilen zwischen den Sonder-Tags #WHERE <*Bedingung*> und # werden je einmal für jeden Datensatz, bei dem die angegebene Bedingung wahr ist, in den Bericht kopiert. Die Bedingung besteht aus einem CLEM-Ausdruck. (Die Klammern bei der WHERE-(Wenn-)Bedingung sind optional.) Beispiel: Mit den Zeilen

```
#WHERE [GESCHL = 'M']
Der Mann in Datensatz Nr. [@INDEX] ist [ALTER] Jahre alt.
#
```

wird je eine Zeile für jeden Datensatz in die Datei geschrieben, bei dem der Wert *M* für das Geschlecht vorliegt. Der vollständige Datensatz enthält die festen, globalen und bedingten Zeilen, die durch Anwendung der Vorlage auf die Eingabedaten definiert wurden.

Auf der Registerkarte “Ausgabe” können Sie Optionen für das Anzeigen und Speichern der Ergebnisse festlegen, die verschiedenen Arten von Ausgabeknoten gemeinsam sind. Für weitere Informationen siehe Thema Registerkarte “Ausgabe” beim Ausgabeknoten auf S. 406.

Ausgabe von Daten im HTML- oder XML-Format

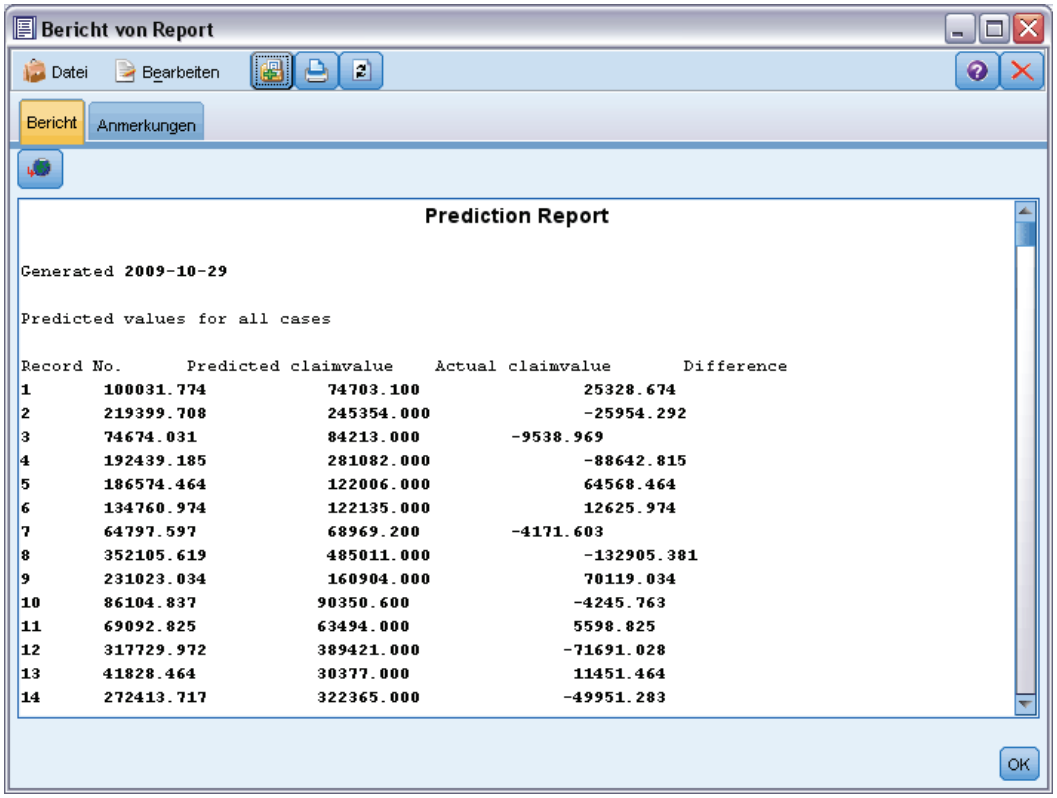
Sie können HTML- oder XML-Tags direkt in die Vorlage einfügen, um Berichte in einem dieser Formate zu schreiben. Die folgende Vorlage beispielsweise führt zu einer HTML-Tabelle.

```
Dieser Bericht wurde in HTML geschrieben.
Dabei werden nur Fälle eingeschlossen, bei denen die Variable "Age" (Alter) größer 60 ist.

<HTML>
 <TABLE border="2">
   <TR>
    <TD>Alter</TD>
    <TD>BP</TD>
    <TD>Cholesterol</TD>
    <TD>Drug</TD>
   </TR>

 #WHERE Alter > 60
  <TR>
     <TD>[Alter]</TD>
     <TD>[BP]</TD>
     <TD>[Cholesterol]</TD>
     <TD>[Drug]</TD>
    </TR>
 #
 </TABLE>
</HTML>
```

Berichtknoten-Ausgabe-Browser

Der Bericht-Browser zeigt den Inhalt des erzeugten Berichts. Das Menü “Datei” enthält die üblichen Befehle zum Speichern, Exportieren und Drucken, das Menü “Bearbeiten” die üblichen Bearbeitungsfunktionen. Für weitere Informationen siehe Thema Anzeigen der Ausgabe auf S. 399.

Abbildung 6-46
Bericht-Browser



Globalwerteknoten

Mit dem Globalwerteknoten werden die Daten gescannt und Übersichtswerte berechnet, die in CLEM-Ausdrücken herangezogen werden können. Mit einem Globalwerteknoten können Sie beispielsweise die Statistiken für das Feld *age* (Alter) berechnen und dann den Gesamtmittelwert für *age* (Alter) in CLEM-Ausdrücken verwenden. Fügen Sie hierzu die Funktion @GLOBAL_MEAN(age) ein.

Registerkarte "Einstellungen" beim Globalwerteknoten

Abbildung 6-47
Globalwerteknoten: Registerkarte "Einstellungen"

Zu erstellende Globalwerte. Wählen Sie das oder die Felder aus, für die Globalwerte verfügbar sein sollen. Sie können mehrere Felder auswählen. Geben Sie die zu berechnenden Statistiken für jedes Feld an. Wählen Sie hierzu die gewünschten Statistiken in den Spalten neben dem Feldnamen aus.

- **Mittelwert.** Durchschnittlicher Wert (Mittelwert) für das Feld über alle Datensätze hinweg.
- **Summe.** Summe der Werte für das Feld über alle Datensätze hinweg.
- **Min.** Mindestwert für das Feld.
- **Max.** Höchstwert für das Feld.
- **Std.Abw.** Die Standardabweichung, ein Maß für die Variabilität der Werte eines Felds, berechnet als die Quadratwurzel der Varianz.

Standardoperation(en). Die hier ausgewählten Optionen werden verwendet, wenn Sie weitere Felder zur obigen Liste der Globalwerte hinzufügen. Um die Standardgruppe der Statistiken zu ändern, wählen Sie die gewünschten Statistiken aus oder heben Sie die Auswahl bestimmter Statistiken wieder auf. Mit der Schaltfläche Anwenden können Sie zudem die Standardoperationen auf alle Felder in der Liste gleichzeitig anwenden.

Alle Globalwerte vor Ausführung löschen. Vor der Berechnung neuer Globalwerte werden alle vorhandenen Globalwerte gelöscht. Ist diese Option nicht ausgewählt, ersetzen die neuen berechneten Werte zwar die bisherigen Werte, die nicht neu berechneten Globalwerte bleiben jedoch weiterhin verfügbar.

Vorschau der Globalwerte anzeigen, die nach der Ausführung erstellt wurden. Mit dieser Option wird nach der Ausführung das Dialogfeld "Stream-Eigenschaften" mit der Registerkarte "Globalwerte" geöffnet; hier werden die berechneten Globalwerte angezeigt.

IBM SPSS Statistics-Hilfsprogramme

Wenn eine kompatible Version von IBM® SPSS® Statistics auf dem Computer installiert und lizenziert ist, können Sie IBM® SPSS® Modeler so konfigurieren, dass Daten mit SPSS Statistics-Funktionen über den Statistics-Transformations-, den Statistics-Modell-, den Statistics-Ausgabe- oder den Statistics-Exportknoten verarbeitet werden.

▶ Um SPSS Modeler für die Zusammenarbeit mit SPSS Statistics und anderen Anwendungen zu konfigurieren, wählen Sie:
Werkzeuge > Optionen > Hilfsprogramme

IBM SPSS Statistics Interactive. Geben Sie den vollständigen Pfad und Namen des Befehls (z. B. *C:\Programme\IBM\SPSS\Statistics\<nn>\stats.exe*) ein, der verwendet werden soll, wenn SPSS Statistics direkt für eine vom Statistikexportknoten erzeugte Datendatei gestartet wird. Für weitere Informationen siehe Thema Statistikexportknoten in Kapitel 8 auf S. 514.

Verbindung. Wenn sich SPSS Statistics Server auf demselben Host befindet wie IBM® SPSS® Modeler Server, können Sie eine Verbindung zwischen diesen beiden Anwendungen aktivieren, mit der die Effizienz gesteigert wird, weil Daten während der Analyse auf dem Server belassen werden. Mit Server aktivieren Sie unten die Option Port. Die Standardeinstellung lautet Lokal.

Port. Bestimmen Sie den Server-Port für SPSS Statistics Server.

IBM SPSS Statistics-Dienstprogramm für Lizenzstandort. Damit SPSS Modeler den Statistiktransformations-, den Statistikmodell- und den Statistikausgabeknoten verwenden kann, muss auf dem Computer, auf dem der Stream ausgeführt wird, eine Kopie von SPSS Statistics installiert und lizenziert sein. Bei der Ausführung im verteilten Modus unter einer entfernten SPSS Modeler Server-Instanz muss sich auf Ihrem SPSS Modeler-Client-Computer außerdem eine lizenzierte Kopie von SPSS Statistics befinden.

- Wenn Sie SPSS Modeler im lokalen Modus (Standalone-Modus) ausführen, muss sich die lizenzierte Kopie von SPSS Statistics auf dem lokalen Computer befinden. Klicken Sie auf diese Schaltfläche, um den Standort der lokalen SPSS Statistics-Installation anzugeben, die für die Lizenzierung verwendet werden soll.
- Bei einer Ausführung im verteilten Modus unter einer entfernten SPSS Modeler Server-Instanz muss sich außerdem die lizenzierte Version von SPSS Statistics auf dem Server-Computer befinden und die Lizenzkonfiguration muss auf dem Server erfolgen. Wechseln Sie hierzu über die Befehlszeilen-Eingabeaufforderung zum *Bin*-Verzeichnis von SPSS Modeler Server und führen Sie unter Windows folgenden Befehl aus:

```
statisticsutility -location=<Pfad zur IBM SPSS Statistics-Server-Lizenzdatei>/bin
```

Alternativ führen Sie unter UNIX folgenden Befehl aus:

```
./statisticsutility -location=<Pfad zur IBM SPSS Statistics-Server-Lizenzdatei>/bin
```

Dabei ist *<Pfad zur SPSS Statistics Server-Lizenzdatei>* das Installationsverzeichnis eines lizenzierten SPSS Statistics-Servers.

Wenn sich keine lizenzierte Kopie von SPSS Statistics auf Ihrem lokalen Rechner befindet, können Sie den Statistikdateiknoten trotzdem mithilfe eines lizenzierten SPSS Statistics-Server ausführen, das Ausführen anderer SPSS Statistics-Knoten wird jedoch zu Fehlermeldungen führen.

Kommentare

Wenn Schwierigkeiten beim Ausführen von SPSS Statistics-Prozedurknoten auftreten, beachten Sie die folgenden Tipps:

- Falls die Feldnamen in SPSS Modeler länger als acht (bei Versionen vor SPSS Statistics) bzw. 64 Zeichen sind (bei SPSS Statistics 12.0 und späteren Versionen) oder ungültige Zeichen enthalten, müssen diese Namen vor dem Einlesen in SPSS Statistics geändert oder gekürzt werden. Für weitere Informationen siehe Thema Umbenennen oder Filtern von Feldern für IBM SPSS Statistics in Kapitel 8 auf S. 516.
- Wenn SPSS Statistics nach SPSS Modeler installiert wurde, müssen Sie möglicherweise den Standort der SPSS Statistics-Lizenz angeben, wie oben erläutert.

Kapitel 7

Exportknoten

Überblick über Exportknoten

Exportknoten bieten einen Mechanismus zum Exportieren von Daten in verschiedenen Formaten, sodass Sie diese Daten auch mit anderen Software-Tools nutzen können.

Folgende Exportknoten stehen zur Verfügung:

Der Datenbankexportknoten schreibt Daten in eine ODBC-kompatible relationale Datenquelle. Um Daten in eine ODBC-Datenquelle schreiben zu können, muss die betreffende Datenquelle bereits vorhanden sein und Sie benötigen Schreibzugriff dafür. Für weitere Informationen siehe Thema Datenbankexportknoten auf S. 461.

Der Textdatei-Export gibt Daten in einer Textdatei mit Trennzeichen aus. Diese Vorgehensweise eignet sich für das Exportieren von Daten, die von anderen Analyse- oder Tabellenkalkulationsprogrammen gelesen werden sollen. Für weitere Informationen siehe Thema Textdatei-Exportknoten auf S. 482.

Der Statistikexportknoten gibt Daten im Format IBM® SPSS® Statistics.*sav* aus. Die *.sav*-Dateien können von SPSS Statistics Base und anderen Produkten gelesen werden. Dieses Format wird auch für Cache-Dateien in IBM® SPSS® Modeler verwendet. Für weitere Informationen siehe Thema Statistikexportknoten in Kapitel 8 auf S. 514.

Der IBM® SPSS® Data Collection-Exportknoten gibt Daten in dem von der Marktforschungssoftware Data Collection verwendeten Format aus. Um diesen Knoten verwenden zu können, muss die Data Collection Data Library installiert sein. Für weitere Informationen siehe Thema IBM SPSS Data Collection-Exportknoten auf S. 484.

Mit dem SAS-Exportknoten werden Daten in das SAS-Format ausgegeben, die dann in SAS oder in SAS-kompatible Softwarepakete eingelesen werden können. Es stehen drei SAS-Dateiformate zur Verfügung: SAS für Windows/OS2, SAS für UNIX sowie SAS Version 7/8 Für weitere Informationen siehe Thema SAS-Exportknoten auf S. 490.

Der Excel-Exportknoten gibt Daten im Microsoft Excel-Format (*.xls*) aus. Optional können Sie auswählen, dass bei der Ausführung des Knotens Excel automatisch gestartet und die exportierte Datei geöffnet werden soll. Für weitere Informationen siehe Thema Excel-Exportknoten auf S. 491.

Der XML-Exportknoten gibt Daten an eine Datei im XML-Format aus. Optional können Sie einen XML-Quellenknoten erstellen, um die exportierten Daten wieder in den Stream einzulesen. Für weitere Informationen siehe Thema XML-Exportknoten auf S. 493.

Datenbankexportknoten

Mithilfe des Knotens “Datenbank” können Sie Daten in ODBC-konforme relationale Datenquellen schreiben, die in der Beschreibung des Quellenknotens “Datenbank” erläutert werden. Für weitere Informationen siehe Thema Datenbankquellenknoten in Kapitel 2 auf S. 15.

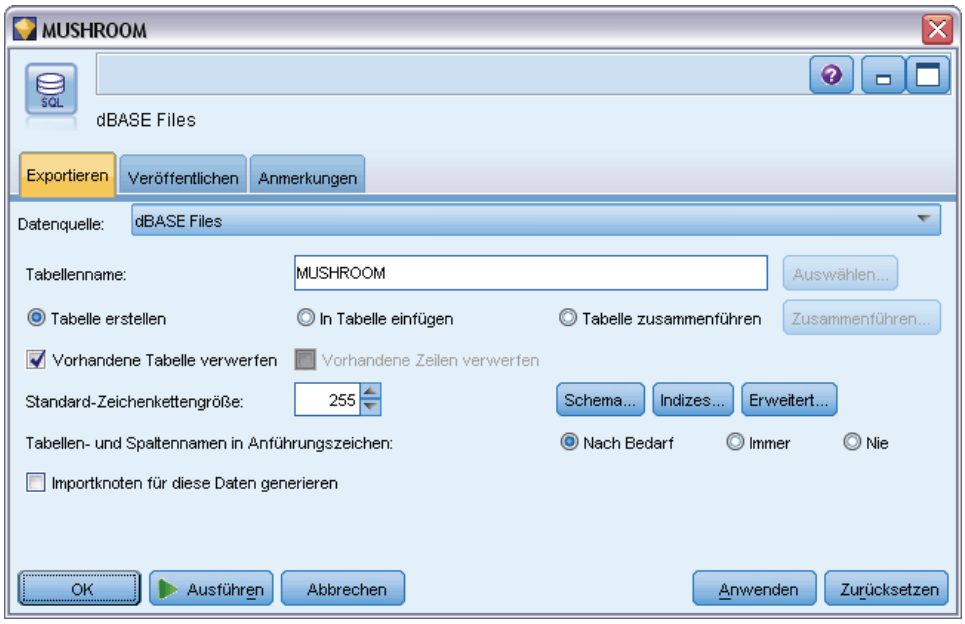
Führen Sie die folgenden allgemeinen Schritte aus, um Daten in eine Datenbank zu schreiben:

- Installieren Sie einen ODBC-Treiber und konfigurieren Sie eine Datenquelle für die zu verwendende Datenbank.
- Geben Sie auf der Registerkarte “Exportieren” des Datenbankknotens die Datenquelle und die Tabelle an, in die geschrieben werden soll. Sie können eine neue Tabelle erstellen oder Daten in eine bestehende Tabelle einfügen.
- Geben Sie nach Bedarf weitere Optionen an.

Diese Schritte werden in den nächsten Themenabschnitten ausführlicher beschrieben.

Registerkarte “Exportieren” beim Datenbankknoten

Abbildung 7-1
Registerkarte “Exportieren” beim Datenbankexportknoten



Datenquelle. Zeigt die ausgewählte Datenquelle. Geben Sie den Namen ein oder wählen Sie einen Eintrag in der Dropdown-Liste aus. Wird die gewünschte Datenbank nicht in der Liste aufgeführt, wählen Sie Neue Datenbankverbindung hinzufügen und wechseln Sie im Dialogfeld “Datenbankverbindungen” zu dieser Datenbank. Für weitere Informationen siehe Thema Hinzufügen einer Datenbankverbindung in Kapitel 2 auf S. 18.

Tabellenname. Geben Sie den Namen der Tabelle ein, an die die Daten gesendet werden sollen. Bei der Option In Tabelle einfügen können Sie eine vorhandene Tabelle in der Datenbank auswählen, indem Sie auf die Schaltfläche Auswählen klicken.

Tabelle erstellen. Mit dieser Option können Sie eine neue Datenbanktabelle anlegen oder eine vorhandene Datenbanktabelle überschreiben.

In Tabelle einfügen. Mit dieser Option fügen Sie die Daten als neue Zeilen in eine vorhandene Datenbanktabelle ein.

Tabelle einlesen. (Wenn verfügbar) Aktivieren Sie diese Option, um ausgewählte Datenbankspalten mit Werten aus entsprechenden Quellendatenfeldern zu aktualisieren. Wenn Sie diese Option auswählen, wird die Schaltfläche Zusammenführen aktiviert, die ein Dialogfeld öffnet, in dem Sie Quellendatenfelder zu Datenbankspalten zuordnen können.

Vorhandene Tabelle verwerfen. Wenn Sie eine neue Tabelle erstellen, lassen Sie mit dieser Option alle vorhandenen Tabellen löschen, die denselben Namen besitzen wie die neu zu erstellende Tabelle.

Vorhandene Zeilen verwerfen. Wenn Sie Daten in eine Tabelle einfügen, lassen Sie mit dieser Option vorhandene Zeilen vor dem Exportieren aus der Tabelle löschen.

Hinweis: Bei den oben genannten beiden Optionen wird eine Überschreibungswarnung eingeblendet, sobald Sie den Knoten ausführen. Sollen diese Warnungen unterdrückt werden, deaktivieren Sie im Dialogfeld “Benutzeroptionen” auf der Registerkarte “Benachrichtigungen” die Option Warnen, wenn eine Datenbanktabelle durch einen Knoten überschrieben wird.

Standard-Zeichenkettengröße Felder, die Sie als “Ohne Typ” in einem aufwärts liegenden Typknoten gekennzeichnet haben, werden als Zeichenkettenfelder in die Datenbank geschrieben. Geben Sie die Größe der Zeichenketten an, die für Felder ohne Typ verwendet werden sollen.

Klicken Sie auf Schema, um ein Dialogfeld zu öffnen, in dem Sie verschiedene Exportoptionen festlegen können (für Datenbanken, die diese Funktion unterstützen), und geben Sie den Primärschlüssel für die Datenbankindizierung an. Für weitere Informationen siehe Thema Schemaoptionen für den Datenbankexport auf S. 465.

Klicken Sie auf Indizes, um Optionen für die Indizierung der exportierten Tabelle anzugeben und damit die Datenbankleistung zu verbessern. Für weitere Informationen siehe Thema Indexoptionen für den Datenbankexport auf S. 469.

Mit der Schaltfläche Erweitert können Sie Optionen für das Massenladen und die Datenbankübertragung festlegen. Für weitere Informationen siehe Thema Erweiterte Optionen für den Datenbankexport auf S. 472.

Tabellen- und Spaltennamen in Anführungszeichen. Wählen Sie die Optionen aus, die beim Senden der Anweisung CREATE TABLE an die Datenbank verwendet werden sollen. Enthält der Name von Tabellen und Spalten ein Leerzeichen oder ein Sonderzeichen, muss der Name in Anführungszeichen gesetzt werden.

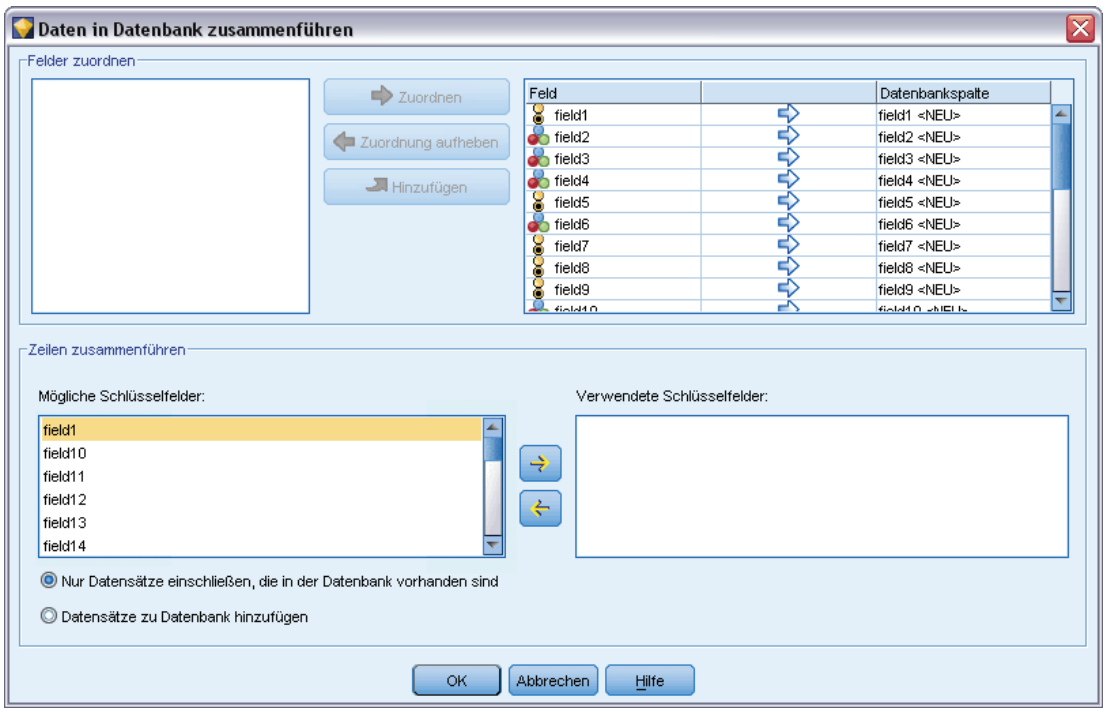
- **Nach Bedarf.** Hiermit lassen Sie automatisch von Fall zu Fall durch IBM® SPSS® Modeler feststellen, ob Anführungszeichen erforderlich sind oder nicht.
- **Immer.** Die Tabellen- und Spaltennamen werden immer in Anführungszeichen eingeschlossen.
- **Nie.** Es werden keine Anführungszeichen verwendet.

Importknoten für diese Daten generieren. Es wird ein Datenbankquellenknoten für die Daten erzeugt, die in die angegebene Datenquelle und Tabelle exportiert wurden. Beim Ausführen wird dieser Knoten in den Stream-Zeichenbereich aufgenommen.

Zusammenführungsoptionen für den Datenbankexport

Dieses Dialogfeld ermöglicht Ihnen, Felder aus den Quelldaten zu Spalten in der Zieldatenbanktabelle zuzuordnen Beim Zuordnen eines Quelldatenfelds zu einer Datenbankspalte wird der Spaltenwert durch den Quelldatenwert ersetzt, wenn der Stream ausgeführt wird. Nicht zugeordnete Quellenfelder bleiben in der Datenbank unverändert.

Abbildung 7-2
Zuordnen von Quelldatenfeldern zu Datenbankspalten



Felder zuordnen. Hier geben Sie die Zuordnung zwischen Quelldatenfeldern und Datenbankspalten an. Quelldatenfelder mit demselben Namen wie Spalten in der Datenbank werden automatisch zugeordnet.

- **Zuordnen.** Ordnet ein Quelldatenfeld, das in der Feldliste links neben der Schaltfläche ausgewählt wurde, einer Datenbankspalte zu, die in der Liste rechts ausgewählt wurde. Sie können mehrere Felder gleichzeitig zuordnen, aber die Anzahl der ausgewählten Einträge muss in beiden Listen gleich sein.

- **Zuordnung aufheben.**Entfernt die Zuordnung für ein oder mehrere ausgewählte Datenbankspalten. Diese Schaltfläche wird aktiviert, wenn Sie ein Feld oder eine Datenbankspalte in der Tabelle im rechten Bereich des Dialogfelds auswählen.
- **Hinzufügen.** Fügt ein oder mehr Quelldatenfelder, die in der Feldliste links neben der Schaltfläche ausgewählt wurden, der Liste rechts zu, die bereit für die Zuordnung ist. Diese Schaltfläche wird aktiviert, wenn Sie ein Feld in der Liste im linken Bereich auswählen und in der Liste im rechten Bereich kein Feld mit diesem Namen vorhanden ist. Wenn Sie auf die Schaltfläche klicken, wird das ausgewählte Feld einer neuen Datenbankspalte mit dem gleichen Namen zugeordnet. Das Wort <NEU> wird hinter dem Namen der Datenbankspalte angezeigt, um anzuzeigen, dass es sich um ein neues Feld handelt.

Zeilen zusammenführen. Sie verwenden ein Schlüsselfeld wie *Transaktions-ID*, um Datensätze mit demselben Wert im Schlüsselfeld zusammenzuführen. Dies entspricht einem Datenbank-"Equi-Join". Schlüsselwerte müssen zu Primärschlüsseln gehören, das heißt, sie müssen eindeutig sein und dürfen keine Nullwerte enthalten.

- **Mögliche Felder.** Listet alle Felder auf, die in allen Eingabedatenquellen gefunden wurden. Wählen Sie ein oder mehrere Felder aus dieser Liste und fügen Sie sie mithilfe der Pfeilschaltfläche als Schlüsselfeld für die Zusammenführung von Datensätzen hinzu. Jedes Zuordnungsfeld mit einer entsprechenden zugeordneten Datenbankspalte ist als Schlüssel verfügbar, lediglich Felder, die als neue Datenbankspalten hinzugefügt wurden (durch <NEU> nach dem Namen gekennzeichnet) sind nicht verfügbar.
- **Verwendete Schlüsselfelder.** Listet alle Felder auf, die für die Zusammenführung der Datensätze aus allen Eingabedatenquellen auf der Grundlage der Schlüsselfeldwerte verwendet werden. Um einen Schlüssel aus der Liste zu entfernen, wählen Sie ihn aus und verschieben Sie ihn mithilfe der Pfeilschaltfläche zurück in die Liste "Mögliche Schlüsselfelder". Bei Auswahl mehrerer Schlüssel wird die unten stehende Option aktiviert.
- **Nur Datensätze einschließen, die in der Datenbank existieren.** Führt einen partiellen Join aus. Wenn sich der Datensatz in der Datenbank und im Stream befindet, werden die zugeordneten Felder aktualisiert.
- **Datensätze zur Datenbank hinzufügen.** Führt einen Outer Join aus. Alle Datensätze im Stream werden zusammengeführt (wenn derselbe Datensatz in der Datenbank vorhanden ist) oder hinzugefügt (wenn der Datensatz noch nicht in der Datenbank existiert).

So ordnen Sie ein Quelldatenfeld einer neuen Datenbankspalte zu:

▶ Klicken Sie auf den Quellenfeldnamen in der Liste links unter Felder zuordnen.

▶ Klicken Sie auf die Schaltfläche Hinzufügen, um die Zuordnung abzuschließen.

So ordnen Sie ein Quelldatenfeld einer vorhandenen Datenbankspalte zu:

▶ Klicken Sie auf den Quellenfeldnamen in der Liste links unter Felder zuordnen.

▶ Klicken Sie rechts unter Datenbankspalte auf den Spaltennamen.

▶ Klicken Sie auf die Schaltfläche Zuordnen, um die Zuordnung abzuschließen.

So entfernen Sie eine Zuordnung:

- Klicken Sie in der Liste rechts unter “Feld” auf den Namen des Felds, für das Sie die Zuordnung entfernen möchten.
- Klicken Sie auf die Schaltfläche Zuordnung aufheben.

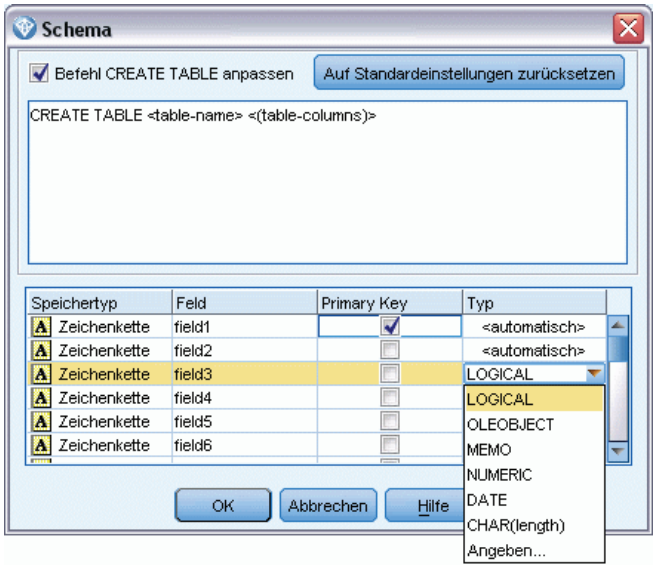
So wählen Sie ein Feld in einer beliebigen Liste aus

- Halten Sie die STRG-Taste gedrückt und klicken Sie auf den Feldnamen.

Schemaoptionen für den Datenbankexport

Im Dialogfeld “Schema” für den Datenbankexport können Sie Optionen für den Datenbankexport festlegen (für Datenbanken, die diese Optionen unterstützen), die SQL-Datentypen für die Felder bestimmen, angeben, bei welchen Feldern es sich um Primärschlüssel handelt, sowie die beim Exportieren erstellte Anweisung CREATE TABLE anpassen.

Abbildung 7-3
Beispiel für das Dialogfeld “Schema” für den Datenbankexport



Das Dialogfeld besteht aus verschiedenen Teilen:

- Der Abschnitt oben (sofern angezeigt) enthält Optionen für den Export in eine Datenbank, die diese Optionen unterstützt. Dieser Bereich wird nicht angezeigt, wenn Sie nicht mit einer solchen Datenbank verbunden sind.
- Das Textfeld in der Mitte zeigt die zur Generierung des Befehls CREATE TABLE verwendete Vorlage an, die standardmäßig folgendes Format aufweist:

  CREATE TABLE <table-name> <(table columns)>
- Mit der Tabelle im unteren Bereich können Sie den SQL-Datentyp für die einzelnen Felder festlegen und angeben, bei welchen Feldern es sich um Primärschlüssel handelt (siehe unten). Das Dialogfeld generiert automatisch die Werte der Parameter <table-name> und <(table columns)> auf der Grundlage der Spezifikationen in der Tabelle.

Festlegen der Optionen für den Datenbankexport

Wenn dieser Abschnitt angezeigt wird, können Sie eine Reihe von Einstellungen für den Export in die Datenbank angeben. Diese Funktion wird von folgenden Datenbanktypen unterstützt:

- IBM InfoSphere Warehouse unter DB2 9.1 oder höher. Für weitere Informationen siehe Thema Optionen für IBM DB2 InfoSphere Warehouse auf S. 467.
- SQL Server 2008 oder höher, Enterprise und Developer Edition. Für weitere Informationen siehe Thema Optionen für SQL Server auf S. 467.
- Oracle 10g und 11gR1 oder höher, Enterprise oder Personal Edition. Für weitere Informationen siehe Thema Optionen für Oracle auf S. 468.

Anpassen der Anweisung CREATE TABLE

Im Textfeldbereich dieses Dialogfelds können Sie zusätzliche datenbankspezifische Optionen in die Anweisung CREATE TABLE aufnehmen.

▶ Aktivieren Sie das Kontrollkästchen Befehl CREATE TABLE anpassen, damit das Textfenster zur Verfügung gestellt wird.

▶ Ergänzen Sie die Anweisung mit den gewünschten datenbankspezifischen Optionen. Die Textparameter <table-name> und (<table-columns>) müssen beibehalten werden, weil diese Parameter von IBM® SPSS® Modeler durch den tatsächlichen Tabellennamen und die entsprechende Spaltendefinition ersetzt werden.

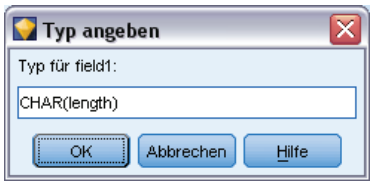
Festlegen von SQL-Datentypen

Standardmäßig ist es bei SPSS Modeler möglich, dass die SQL-Datentypen automatisch durch den Datenbankserver zugewiesen werden. Soll der automatisch festgelegte Typ für ein Feld überschrieben werden, wechseln Sie zur zugehörigen Zeile für das Feld und wählen Sie den gewünschten Typ in der Schematabelle in der Dropdown-Liste in der Spalte *Type* aus. Durch Drücken der Umschalttaste und Klicken können Sie mehr als eine Reihe auswählen.

Bei Typen, für die ein Längen-, Genauigkeits- oder Skalenargument erforderlich ist (BINARY, VARBINARY, CHAR, VARCHAR, NUMERIC und NUMBER), sollten Sie selbst eine Länge zuweisen, also nicht die automatische Längenzuweisung durch den Datenbankserver nutzen. Wenn Sie

beispielsweise einen angemessenen Wert für die Länge festlegen, z. B. VARCHAR(25), ist sichergestellt, dass der Speichertyp in SPSS Modeler überschrieben wird (falls Sie dies wünschen). Soll die automatische Zuweisung überschrieben werden, wählen Sie in der Dropdown-Liste "Typ" den Eintrag Angeben und ersetzen Sie die Typdefinition durch die gewünschte Anweisung für die SQL-Typdefinition.

Abbildung 7-4
Dialogfeld "Typ angeben" bei der Datenbankausgabe



Die einfachste Methode besteht darin, zunächst den Typ auszuwählen, der der gewünschten Typdefinition am nächsten kommt, dann die Option Angeben zu wählen und schließlich die zugehörige Definition zu bearbeiten. Um den SQL-Datentyp beispielsweise auf VARCHAR(25) zu setzen, wählen Sie zunächst in der Dropdown-Liste "Typ" den Eintrag VARCHAR(length) aus. Wählen Sie dann Angeben und ersetzen Sie die Textlänge durch den Wert 25.

Primärschlüssel

Wenn eine oder mehrere Spalten in der exportierten Tabelle einem eindeutigen Wert bzw. eine eindeutige Wertekombination für jede Zeile aufweisen muss, können Sie dies angeben, indem Sie für jedes betroffene Feld das Kontrollkästchen Primärschlüssel aktivieren. Bei den meisten Datenbanken ist es nicht zulässig, die Tabelle auf eine Weise zu ändern, die eine Primärschlüsselbeschränkung ungültig macht. Bei diesen Datenbanken wird zur Durchsetzung dieser Einschränkung automatisch ein Index über dem Primärschlüssel erstellt. (Optional können Sie im Dialogfeld "Indizes" Indizes für andere Felder erstellen). Für weitere Informationen siehe Thema Indexoptionen für den Datenbankexport auf S. 469.)

Optionen für IBM DB2 InfoSphere Warehouse

Tabellenbereich. Der Tabellenbereich, der für den Export verwendet wird. Datenbankadministratoren können Tabellenbereiche partioniert erstellen oder konfigurieren. Wir empfehlen, einen dieser Tabellenbereiche (anstelle des standardmäßig eingestellten) für den Datenbankexport zu verwenden.

Daten nach Feld partitionieren. Legt das Eingabefeld für die Partionierung fest.

Komprimierung verwenden. Wenn ausgewählt, werden Tabellen für den komprimierten Export erstellt (entspricht z. B. CREATE TABLE MYTABLE(…) COMPRESS YES; in SQL).

Optionen für SQL Server

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Zeile.** Aktiviert Komprimierung auf der Zeilenebene (entspricht z. B. CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = ROW); in SQL).
- **Seite.** Aktiviert Komprimierung auf der Seitenebene (z. B. CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = PAGE); in SQL).

Optionen für Oracle

Oracle 10g-Einstellungen

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt. Für diese Version der Datenbank steht nur einfache Komprimierung zur Verfügung (beispielsweise CREATE TABLE MYTABLE(…) COMPRESS; in SQL).

Oracle 11gR1-Einstellungen

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. CREATE TABLE MYTABLE(…) COMPRESS; in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option Direkte Ladevorgänge.
- **Direkte Ladevorgänge.** Aktiviert Komprimierung ausschließlich für Masseneinfügevorgänge (direkter Pfad) (z. B. CREATE TABLE MYTABLE(…)COMPRESS FOR DIRECT_LOAD OPERATIONS; in SQL).
- **Alle Vorgänge.** Aktiviert Komprimierung für alle Vorgänge (z. B. CREATE TABLE MYTABLE(…)COMPRESS FOR ALL OPERATIONS; in SQL).

Oracle 11gR2-Einstellungen – Option "Basic" (Einfach)

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. CREATE TABLE MYTABLE(…) COMPRESS; in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option Einfach.
- **Einfach.** Aktiviert einfache Komprimierung (z. B. CREATE TABLE MYTABLE(…) COMPRESS BASIC; in SQL).

Oracle 11gR2-Einstellungen – Option "Advanced" (Erweitert)

Komprimierung verwenden. Wenn diese Option ausgewählt ist, werden Tabellen für den Export mit Komprimierung erstellt.

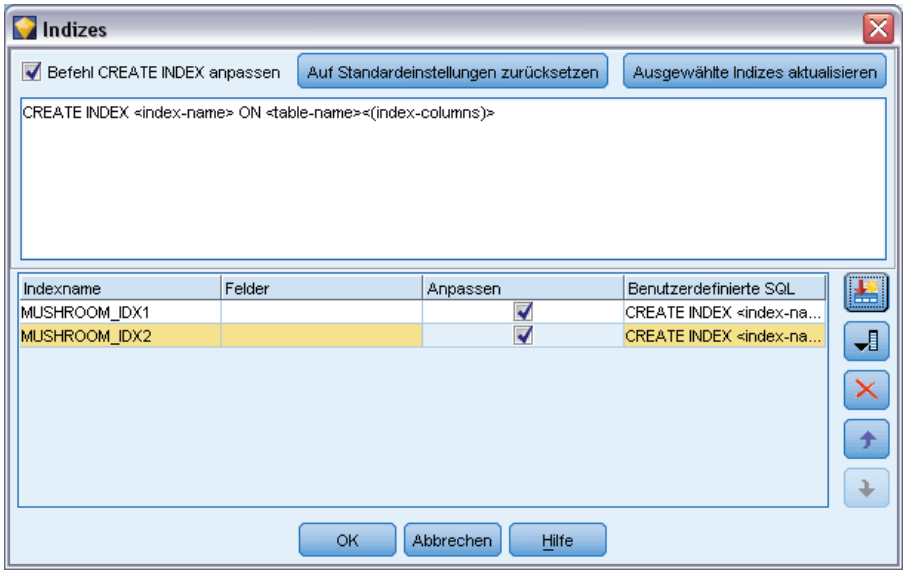
Komprimierung für. Wählen Sie die Komprimierungsstufe aus.

- **Standard.** Aktiviert Standardkomprimierung (z. B. CREATE TABLE MYTABLE(…) COMPRESS; in SQL). In diesem Fall hat sie dieselbe Wirkung wie die Option Einfach.
- **Einfach.** Aktiviert einfache Komprimierung (z. B. CREATE TABLE MYTABLE(…) COMPRESS BASIC; in SQL).
- **OLTP.** Aktiviert OLTP-Komprimierung (z. B. CREATE TABLE MYTABLE(…)COMPRESS FOR OLTP; in SQL).
- **Abfrage niedrig/hoch.** (nur Exadata-Server) Aktiviert Hybrid Columnar Compression für Abfrage (z. B. CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY LOW; oder CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY HIGH; in SQL). Komprimierung für Abfrage ist in Data Warehousing-Umgebungen sinnvoll; HIGH bietet ein höheres Komprimierungsverhältnis als LOW.
- **Archiv niedrig/hoch.** (nur Exadata-Server) Aktiviert Hybrid Columnar Compression für Archiv (z. B. CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE LOW; oder CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE HIGH; in SQL). Komprimierung für Archiv ist sinnvoll zur Komprimierung von Daten, die lange Zeit gespeichert werden sollen; HIGH bietet ein höheres Komprimierungsverhältnis als LOW.

Indexoptionen für den Datenbankexport

Mit dem Dialogfeld “Indizes” können Sie Indizes für aus IBM® SPSS® Modeler exportierte Datenbanktabellen erstellen. Sie können die einzuschließenden Feld-Sets angeben und den Befehl CREATE INDEX nach Bedarf anpassen.

Abbildung 7-5
Dialogfeld “Indizes” bei der Datenbankausgabe



Das Dialogfeld besteht aus zwei Teilen:

- Das Textfeld im oberen Teil zeigt eine Vorlage an, die zur Generierung eines oder mehrerer Befehle vom Typ CREATE INDEX verwendet werden kann. Diese Vorlage weist standardmäßig folgendes Format auf:

  CREATE INDEX <index-name> ON <table-name>

- Die Tabelle im unteren Bereich des Dialogfelds ermöglicht die Angabe von Spezifikationen für jeden Index, der erstellt werden soll. Geben Sie für jeden Index den Indexnamen und die einzuschließenden Felder bzw. Spalten an. Das Dialogfeld generiert automatisch die Werte der Parameter <index-name> und <table-name> entsprechend den Angaben.

  Beispielsweise kann die generierte SQL für einen einzelnen Index für die Felder *empid* und *deptid* wie folgt aussehen:

  CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)

  Sie können mehrere Zeilen hinzufügen, um mehrere Indizes zu erstellen. Für jede Zeile wird ein gesonderter CREATE INDEX-Befehl generiert.

Anpassen des Befehls CREATE INDEX

Optional können Sie den Befehl CREATE INDEX für alle Indizes oder nur für einen bestimmten Index anpassen. Dadurch haben Sie die Flexibilität, spezielle Datenbankanforderungen oder -optionen zu berücksichtigen und Anpassungen nach Bedarf auf alle oder nur auf bestimmte Indizes anzuwenden.

- Wählen Sie Befehl CREATE INDEX anpassen oben im Dialogfeld, um die Vorlage, die für alle danach hinzugefügten Indizes verwendet wird, anzupassen. Beachten Sie, dass die Änderungen nicht automatisch auf Indizes angewendet werden, die bereits zur Tabelle hinzugefügt wurden.
- Wählen Sie mindestens eine Zeile in der Tabelle aus und klicken Sie oben im Dialogfeld auf Ausgewählte Indizes aktualisieren, um die aktuellen Anpassungen auf alle ausgewählten Zeilen anzuwenden.
- Aktivieren Sie das Kontrollkästchen Anpassen in den einzelnen Zeilen, um die Befehlsvorlage nur für den betreffenden Index zu ändern.

Beachten Sie, dass die Werte der Parameter <index-name> und <table-name> vom Dialogfeld automatisch auf der Grundlage der Tabellenspezifikationen generiert werden und nicht direkt bearbeitet werden können.

BITMAP KEYWORD. Bei Verwendung einer Oracle-Datenbank können Sie die Vorlage so anpassen, dass statt eines Standard-Index ein Bitmap-Index erstellt wird. Dies geschieht wie folgt:

CREATE BITMAP INDEX <index-name> ON <table-name>

Bitmap-Indizes sind nützlich für die Indizierung von Spalten mit einer kleinen Anzahl unterschiedlicher Werte. Die entstehende SQL sieht etwa folgendermaßen aus:

CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)

Schlüsselwort UNIQUE. Die meisten Datenbanken unterstützen das Schlüsselwort UNIQUE im Befehl CREATE INDEX. Dadurch wird eine Eindeutigkeitsbeschränkung ähnlich einer Primärschlüsselbeschränkung in der zugrunde liegenden Tabelle erzwungen.

```
CREATE UNIQUE INDEX <index-name> ON <table-name>
```

Beachten Sie, dass diese Angabe für Felder, die tatsächlich als Primärschlüssel angegeben sind, nicht erforderlich ist. Die meisten Datenbanken erstellen automatisch einen Index für alle Felder, die im Befehl CREATE TABLE als Primärschlüsselfelder festgelegt wurden. Eine explizite Erstellung von Indizes für diese Felder ist also nicht erforderlich. Für weitere Informationen siehe Thema Schemaoptionen für den Datenbankexport auf S. 465.

Schlüsselwort FILLFACTOR. Für einige physische Parameter des Index können Feineinstellungen vorgenommen werden. Beispielsweise ermöglicht SQL Server dem Benutzer, die Indexgröße (nach der ursprünglichen Erstellung) gegen die Wartungskosten bei zukünftigen Änderungen an der Tabelle abzuwägen.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

Weitere Kommentare

- Wenn bereits ein Index mit dem angegebenen Namen vorhanden ist, schlägt die Indexerstellung fehl. Alle Fehlschläge werden zunächst als Warnungen behandelt, sodass die nachfolgenden Indizes erstellt werden können. Nachdem die Erstellung aller Indizes versucht wurde, werden diese Fehlschläge dann im Meldungsprotokoll als Fehler gemeldet.
- Um eine bestmögliche Leistung zu erzielen, sollten die Indizes erstellt werden, nachdem Daten in die Tabelle geladen wurden. Indizes müssen mindestens eine Spalte enthalten.
- Vor der Ausführung des Knotens können Sie die generierte SQL im Meldungsprotokoll anzeigen.
- Für temporäre Tabellen, die in die Datenbank geschrieben wurden (d. h. wenn der Knoten-Cache aktiviert ist) sind die Optionen zur Angabe von Primärschlüsseln und Indizes nicht verfügbar. Das System kann jedoch nach Bedarf Indizes in der temporären Tabelle erstellen, je nachdem, wie die Daten in abwärtsgelegenen Knoten verwendet werden sollen. Wenn die Daten im Cache beispielsweise anschließend durch eine *DEPT*-Spalte verbunden werden, ist es sinnvoll, die im Cache gespeicherte Tabelle auf dieser Spalte zu indizieren.

Indizes und Abfrageoptimierung

In einigen Datenbankverwaltungssystemen ist nach dem Erstellen, Laden und Indizieren einer Datenbanktabelle ein weiterer Schritt erforderlich, bevor der Optimizer die Indizes zur Beschleunigung der Query-Ausführung in der neuen Tabelle nutzen kann. In Oracle beispielsweise erfordert der kostenbasierte Query-Optimizer, dass eine Tabelle analysiert wird, bevor ihre Indizes für die Query-Optimierung verwendet werden können. Die interne ODBC-Eigenschaftendatei für Oracle (für den Benutzer nicht sichtbar) enthält hierfür eine Option:

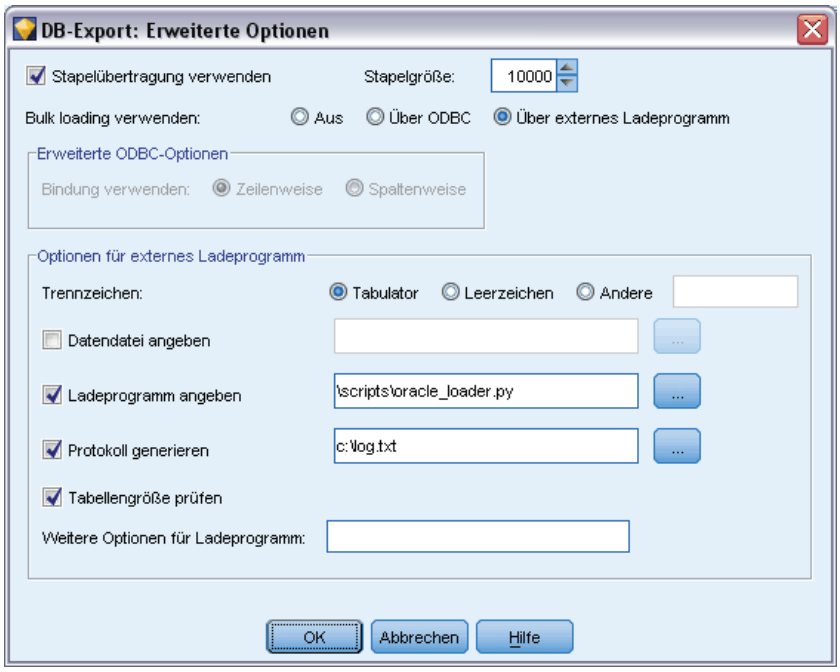
```
# Defines SQL to be executed after a table and any associated indexes
# have been created and populated
table_analysis_sql, 'ANALYZE TABLE <table-name> COMPUTE STATISTICS'
```

Dieser Schritt wird bei jeder Erstellung einer Tabelle in Oracle ausgeführt (unabhängig davon, ob Primärschlüssel oder Indizes definiert sind). Falls erforderlich, kann die ODBC-Eigenschaftsdatei für zusätzliche Datenbanken auf ähnliche Weise angepasst werden. Falls Sie Unterstützung benötigen, wenden Sie sich an den technischen Support von .

Erweiterte Optionen für den Datenbankexport

Wenn Sie im Dialogfeld für den Datenbankexportknoten auf die Schaltfläche “Erweitert” klicken, wird ein neues Dialogfeld geöffnet, in dem Sie die technischen Einzelheiten für das Exportieren der Ergebnisse in eine Datenbank festlegen können.

Abbildung 7-6
Festlegen erweiterter Optionen für den Datenbankexport



Stapelübertragung verwenden. Hiermit deaktivieren Sie die zeilenweise Übertragung an die Datenbank.

Stapelgröße. Hier können Sie die Anzahl der Datensätze angeben, die an die Datenbank gesendet werden sollen, bevor die Übertragung in den Speicher erfolgt. Wenn Sie hier einen niedrigeren Wert angeben, erzielen Sie eine größere Datenintegrität, jedoch auf Kosten der Übertragungsgeschwindigkeit. Nehmen Sie ggf. Feineinstellungen an diesem Wert vor, um so die optimale Leistung der Datenbank zu erreichen.

InfoSphere Warehouse-Optionen. Wird nur angezeigt, wenn Sie mit einer InfoSphere Warehouse-Datenbank verbunden sind (IBM DB2 9.7 oder höher). Do not log updates (Aktualisierungen nicht protokollieren) erlaubt Ihnen, das Protokollieren von Ereignissen zu deaktivieren, wenn Sie Tabellen erstellen und Daten einfügen.

Bulk loading verwenden. Gibt eine Methode an, mit der die Daten per Massenladen direkt aus IBM® SPSS® Modeler in die Datenbank übernommen werden. Möglicherweise müssen Sie ein wenig herumexperimentieren, um herauszufinden, welche Massenladeoperationen für ein bestimmtes Szenario angemessen sind.

- **Über ODBC.** Mit dieser Option lassen Sie Einfügungen mehrerer Zeilen durch die ODBC-API vornehmen; dies ist effizienter als der normale Export in die Datenbank. Wählen Sie die zeilenweise oder spaltenweise Bindung in den Optionen im unteren Bereich.
- **Über externes Ladeprogramm.** Es wird ein benutzerdefiniertes Massenladeprogramm verwendet, das speziell auf die Datenbank abgestimmt ist. Wenn Sie diese Option aktivieren, wird im unteren Bereich eine Reihe von Optionen eingeblendet.

Erweiterte ODBC-Optionen. Diese Optionen stehen nur dann zur Verfügung, wenn Sie die Option Über ODBC aktiviert haben. Beachten Sie, dass diese Funktionalität möglicherweise nicht von allen ODBC-Treibern unterstützt wird.

- **Zeilenweise.** Bei der zeilenweisen Bindung werden die Daten über den Aufruf SQLBulkOperations in die Datenbank geladen. Mit der zeilenweisen Bindung verbessern Sie in der Regel die Geschwindigkeit im Vergleich zu parametrisierten Einfügungen, bei denen die Daten jeweils Datensatz für Datensatz eingefügt werden.
- **Spaltenweise.** Die Daten werden mithilfe der spaltenweisen Bindung in die Datenbank geladen. Bei der spaltenweisen Bindung wird die Leistung gesteigert, indem die einzelnen Datenbankspalten (in einer parametrisierten INSERT-Anweisung) zu einem Array mit *n* Werten verbunden werden. Wenn Sie INSERT einmal ausführen, werden *n* Zeilen in die Datenbank eingefügt. Diese Methode kann die Leistung drastisch erhöhen.

Optionen für externes Ladeprogramm. Wenn Sie die Option Über externes Ladeprogramm aktiviert haben, wird eine Reihe von Optionen eingeblendet, mit denen Sie das Daten-Set in eine Datei exportieren und die Daten dann mithilfe eines benutzerdefinierten Ladeprogramms aus dieser Datei in die Datenbank laden können. SPSS Modeler kann mit externen Ladeprogrammen für viele beliebte Datenbanksysteme zusammenarbeiten. In der Software sind mehrere Skripts enthalten. Diese sind außerdem zusammen mit technischer Dokumentation im Unterverzeichnis *scripts* enthalten. Beachten Sie: Um diese Funktionen verwenden zu können, muss Python 2.7 auf demselben Computer installiert sein wie SPSS Modeler oder IBM® SPSS® Modeler Server und der Parameter python_exe_path muss in der Datei *options.cfg* festgelegt sein. Für weitere Informationen siehe Thema Programmierung des Massenladeprogramms auf S. 474.

- **Trennzeichen.** Hier können Sie angeben, welches Trennzeichen in der exportierten Datei verwendet werden soll. Bei der Option Tabulator erfolgt die Trennung mit Tabulatoren, bei der Option Leerzeichen entsprechend mit Leerzeichen. Mit der Option Andere können Sie ein anderes Zeichen angeben, beispielsweise ein Komma (,).
- **Datendatei angeben.** Geben Sie den Pfad für die Datendatei ein, die beim Massenladen geschrieben wird. Standardmäßig wird eine temporäre Datei im Verzeichnis “temp” auf dem Server angelegt.
- **Ladeprogramm angeben.** Hiermit können Sie ein Programm für das Massenladen auswählen. Standardmäßig wird das Unterverzeichnis *scripts* der SPSS Modeler-Installation nach einem Python-Skript durchsucht, das für eine bestimmte Datenbank ausgeführt werden soll. In der Software sind mehrere Skripts enthalten. Diese sind außerdem zusammen mit technischer Dokumentation im Unterverzeichnis *scripts* enthalten.
- **Protokoll generieren.** Im angegebenen Verzeichnis wird eine Protokolldatei erzeugt. Die Protokolldatei enthält Fehlerinformationen und ist von Nutzen, falls das Massenladen fehlschlägt.

- **Tabellengröße prüfen.** Hiermit lassen Sie eine Tabellenprüfung vornehmen, mit der Sie sicherstellen, dass der Anstieg der Tabellengröße mit der Anzahl der Zeilen übereinstimmt, die aus SPSS Modeler exportiert wurden.
- **Weitere Optionen für Ladeprogramm.** Hier können Sie weitere Argumente für das Ladeprogramm festlegen. Argumente, die ein Leerzeichen enthalten, müssen in (doppelte) Anführungszeichen eingeschlossen werden.

Sollen doppelte Anführungszeichen in optionale Argumente eingefügt werden, stellen Sie den Anführungszeichen jeweils einen umgekehrten Schrägstrich voran. Die als -comment "This is a \"comment\"" festgelegte Option beinhaltet beispielsweise sowohl das -comment-Flag als auch den Kommentar selbst, wiedergegeben als This is a "comment".

Um einen umgekehrten Schrägstrich einzufügen, stellen Sie diesem einen weiteren umgekehrten Schrägstrich voran. Die als -specialdir "C:\\Test Scripts\\" festgelegte Option beinhaltet beispielsweise sowohl das -specialdir-Flag als auch das Verzeichnis selbst, wiedergegeben als C:\Test Scripts\.

Programmierung des Massenladeprogramms

Der Datenbankexportknoten beinhaltet Optionen für das Massenladen im Dialogfeld "Erweiterte Optionen". Mit Massenladeprogrammen können Daten aus einer Textdatei in eine Datenbank geladen werden.

Mit der Option Bulk loading verwenden - Über externes Ladeprogramm wird IBM® SPSS® Modeler für drei Aktionen konfiguriert:

- Erstellen aller erforderlichen Datenbanktabellen.
- Exportieren der Daten in eine Textdatei.
- Aufrufen eines Massenladeprogramms, um die Daten aus dieser Datei in die Datenbanktabelle zu laden.

Normalerweise handelt es sich bei dem Massenladeprogramm nicht um das Ladedienstprogramm der Datenbank selbst (beispielsweise das Dienstprogramm sqlldr von Oracle), sondern um ein kleines Skript bzw. ein kleines Programm, das die richtigen Argumente bildet, alle erforderlichen datenbankspezifischen Hilfsdateien (beispielsweise eine Steuerungsdatei) erstellt und anschließend das Datenbank-Ladeprogramm aufruft. Anhand der Informationen in den folgenden Abschnitten können Sie ein bestehendes Massenladeprogramm bearbeiten.

Alternativ können Sie Ihr eigenes Programm für das Massenladen schreiben. Für weitere Informationen siehe Thema Entwickeln von Massenladeprogrammen auf S. 479.

Skripts für das Massenladen

SPSS Modeler wird mit einer Reihe von Massenladeprogrammen für verschiedene Datenbanken ausgeliefert, die mithilfe von Pythons-Skripts implementiert werden. Wenn Sie einen Stream, der einen Datenbankexportknoten enthält ausführen, während die Option Über externes Ladeprogramm ausgewählt ist, erstellt SPSS Modeler die Datenbanktabelle (sofern erforderlich) über ODBC, exportiert die Daten in eine temporäre Datei auf dem Host, auf dem IBM® SPSS® Modeler Server ausgeführt wird, und ruft anschließend das Massenlade-Skript auf. Dieses Skript führt

dann Dienstprogramme aus, die vom DBMS-Anbieter bereitgestellt wurden, um Daten aus den temporären Dateien in die Datenbank hochzuladen.

Hinweis: Die SPSS Modeler-Installation enthält keinen Python-Runtime-Interpreter, weshalb eine gesonderte Installation von Python erforderlich ist. Für weitere Informationen siehe Thema Erweiterte Optionen für den Datenbankexport auf S. 472.

Für die folgenden Datenbanken werden (im Ordner *\scripts* des SPSS Modeler-Installationsverzeichnisses) Skripts bereitgestellt.

Tabelle 7-1
Bereitgestellte Massenladeskripts

Datenbank	Skriptname	
IBM DB2	*db2_loader.py*	Für weitere Informationen siehe Thema Massenladen von Daten in IBM DB2-Datenbanken auf S. 475.
IBM Netezza	*netezza_loader.py*	Für weitere Informationen siehe Thema Massenladen von Daten in IBM Netezza-Datenbanken auf S. 476.
Oracle	*oracle_loader.py*	Für weitere Informationen siehe Thema Massenladen von Daten in Oracle-Datenbanken auf S. 477.
SQL Server	*mssql_loader.py*	Für weitere Informationen siehe Thema Massenladen von Daten in SQL Server-Datenbanken auf S. 478.
Teradata	*teradata_loader.py*	Für weitere Informationen siehe Thema Massenladen von Daten in Teradata-Datenbanken auf S. 479.

Massenladen von Daten in IBM DB2-Datenbanken

Die folgenden Punkte können Ihnen bei der Konfiguration für das Massenladen von IBM® SPSS® Modeler in eine IBM DB2-Datenbank mithilfe der Optionen für das externe Ladeprogramm im Dialogfeld “DB-Export: Erweiterte Optionen” behilflich sein.

Sicherstellen, dass das Dienstprogramm DB2 Command Line Processor (CLP)installiert ist

Das Skript *db2_loader.py* ruft den DB2 LOAD-Befehl auf. Vergewissern Sie sich, dass der Befehlszeilenprozessor (*db2* unter UNIX, *db2cmd* unter Windows) auf dem Server installiert ist, auf dem *db2_loader.py* ausgeführt werden soll (üblicherweise der Host, auf dem IBM® SPSS® Modeler Server ausgeführt wird).

Überprüfen, ob der Aliasname der lokalen Datenbank mit dem tatsächlichen Datenbanknamen übereinstimmt

Der Aliasname der lokalen DB2-Datenbank ist der Name, der von der DB2-Client-Software verwendet wird, um auf eine Datenbank in einer lokalen oder entfernten DB2-Instanz zu verweisen. Wenn der Aliasname der lokalen Datenbank vom Namen der Remote-Datenbank abweicht, geben Sie zusätzlich folgende Option für das Ladeprogramm an:

```
-alias <Alias_der_lokalen_Datenbank>
```

Hier ein Beispiel: Die Remote-Datenbank trägt den Namen STARS und befindet sich auf dem Host GALAXY, der Alias der lokalen DB2-Datenbank auf dem Host, auf dem SPSS Modeler Server ausgeführt wird, ist jedoch STARS_GALAXY. Verwenden Sie die zusätzliche Ladeprogramm-Option

-alias STARS_GALAXY

Datenkodierung mit Nicht-ASCII-Zeichen

Wenn Sie Massenladen von Daten durchführen, die nicht im ASCII-Format vorliegen, sollten Sie sicherstellen, dass die Codeseitenvariable im Konfigurationsabschnitt von *db2_loader.py* in Ihrem System richtig eingerichtet ist.

Leerstrings

Leerstrings werden als NULL-Werte in die Datenbank exportiert.

Massenladen von Daten in IBM Netezza-Datenbanken

Die folgenden Punkte können Ihnen bei der Konfiguration für das Massenladen von IBM® SPSS® Modeler in eine IBM Netezza-Datenbank mithilfe der Optionen für das externe Ladeprogramm im Dialogfeld “DB-Export: Erweiterte Optionen” behilflich sein.

Sicherstellen, dass das Netezza-Dienstprogramm “nzload” installiert ist

Das Skript *netezza_loader.py* ruft das Netezza-Dienstprogramm *nzload* auf. Vergewissern Sie sich, dass *nzload* installiert und ordnungsgemäß auf dem Server konfiguriert ist, auf dem *netezza_loader.py* ausgeführt werden soll.

Exportieren von Nicht-ASCII-Daten

Wenn Ihr Bericht Daten enthält, die nicht im ASCII-Format vorliegen, müssen Sie möglicherweise -encoding UTF8 zum Feld Weitere Optionen für Ladeprogramm im Dialogfeld “DB-Export: Erweiterte Optionen” hinzufügen. Dadurch sollte sichergestellt werden, dass Nicht-ASCII-Daten ordnungsgemäß hochgeladen werden.

Daten in den Formaten “Datum”, “Zeit” und “Zeitstempel”

Setzen Sie in den Stream-Eigenschaften das Datumsformat auf TT-MM-JJJJ und das Zeitformat auf HH:MM:SS.

Leerstrings

Leerstrings werden als NULL-Werte in die Datenbank exportiert.

Andere Spaltenreichenfolge in Stream- und Zieltabelle beim Einfügen von Daten in eine bestehende Tabelle

Wenn die Spaltenreihenfolge im Stream von der in der Zieltabelle abweicht, werden Datenwerte in die falschen Spalten eingefügt. Verwenden Sie einen Knoten vom Typ "Felder ordnen", um sicherzustellen, dass die Reihenfolge der Spalten im Stream mit der Reihenfolge in der Zieltabelle übereinstimmt. Für weitere Informationen siehe Thema Knoten "Felder ordnen" in Kapitel 4 auf S. 241.

Verfolgen des nzload-Fortschritts

Fügen Sie bei der Ausführung von SPSS Modeler im lokalen Modus -sts zum FeldWeitere Optionen für Ladeprogramm im Dialogfeld "DB-Export: Erweiterte Optionen" hinzu, um alle 10.000 Zeilen im Befehlsfenster, die vom Dialogfeld *nzload* geöffnet wurden, Statusmeldungen anzuzeigen.

Massenladen von Daten in Oracle-Datenbanken

Die folgenden Punkte können Ihnen bei der Konfiguration für das Massenladen von IBM® SPSS® Modeler in eine Oracle-Datenbank mithilfe der Optionen für das externe Ladeprogramm im Dialogfeld "DB-Export: Erweiterte Optionen" behilflich sein.

Sicherstellen, dass das Oracle-Dienstprogramm "sqlldr" installiert ist

Das Skript *oracle_loader.py* ruft das Oracle-Dienstprogramm *sqlldr* auf. Beachten Sie, dass *sqlldr* nicht automatisch in Oracle Client enthalten ist. Vergewissern Sie sich, dass *sqlldr* auf dem Server installiert ist, auf dem *oracle_loader.py* ausgeführt werden soll.

SID bzw. Service-Name der Datenbank angeben

Wenn Sie Daten an einen nichtlokalen Oracle-Server exportieren oder Ihr lokaler Oracle-Server mehrere Datenbanken enthält, müssen Sie im Feld Weitere Optionen für Ladeprogramm im Dialogfeld "DB-Export: Erweiterte Optionen" Folgendes angeben, um die SID bzw. den Service-Namen weiterzugeben:

```
-database <SID>
```

Bearbeiten des Konfigurationsabschnitts in oracle_loader.py

Bearbeiten Sie unter UNIX- (und optional: Windows-)Systemen den Konfigurationsabschnitt zu Beginn des Skripts *oracle_loader.py*. Hier können ggf. Werte für die Umgebungsvariablen ORACLE_SID, NLS_LANG, TNS_ADMIN und ORACLE_HOME angegeben werden, sowie der vollständige Pfad zum Dienstprogramm *sqlldr*.

Daten in den Formaten "Datum", "Zeit" und "Zeitstempel"

In den Stream-Eigenschaften sollten Sie normalerweise das Datumsformat auf JJJJ-MM-TT und das Zeitformat auf HH:MM:SS setzen.

Wenn Sie ein von den oben genannten Werten abweichendes Datums- und Zeitformat verwenden müssen, lesen Sie in Ihrer Oracle-Dokumentation nach und bearbeiten Sie die Skriptdatei *oracle_loader.py*.

Datenkodierung mit Nicht-ASCII-Zeichen

Wenn Sie Massenladen von Daten durchführen, die nicht im ASCII-Format vorliegen, sollten Sie sicherstellen, dass die Umgebungsvariable NLS_LANG in Ihrem System richtig eingerichtet ist. Dieses wird vom Oracle-Ladedienstprogramm *sqlldr* gelesen. Beispielsweise ist der richtige Wert für NLS_LANG für Shift-JIS unter Windows Japanese_Japan.JA16SJIS. Weitere Details zu NLS_LANG finden Sie in Ihrer Oracle-Dokumentation.

Leerstrings

Leerstrings werden als NULL-Werte in die Datenbank exportiert.

Massenladen von Daten in SQL Server-Datenbanken

Die folgenden Punkte können Ihnen bei der Konfiguration für das Massenladen von IBM® SPSS® Modeler in eine SQL Server-Datenbank mithilfe der Optionen für das externe Ladeprogramm im Dialogfeld “DB-Export: Erweiterte Optionen” behilflich sein.

Sicherstellen, dass das SQL Server-Dienstprogramm “bcp.exe” installiert ist

Das Skript *mssql_loader.py* ruft das SQL Server-Dienstprogramm *bcp.exe* auf. Vergewissern Sie sich, dass *bcp.exe* auf dem Server installiert ist, auf dem *mssql_loader.py* ausgeführt werden soll.

Die Verwendung von Leerzeichen als Trennzeichen funktioniert nicht

Vermeiden Sie, im Dialogfeld “DB-Export: Erweiterte Optionen” Leerzeichen als Trennzeichen auszuwählen.

Option “Tabellengröße prüfen” empfohlen

Es wird empfohlen, die Option Tabellengröße prüfen im Dialogfeld “DB-Export: Erweiterte Optionen” zu aktivieren. Fehler beim Massenladevorgang werden nicht immer erkannt und durch die Aktivierung dieser Option wird eine zusätzliche Prüfung durchgeführt, um sicherzustellen, dass die richtige Anzahl an Zeilen geladen wurde.

Leerstrings

Leerstrings werden als NULL-Werte in die Datenbank exportiert.

Massenladen von Daten in Teradata-Datenbanken

Die folgenden Punkte können Ihnen bei der Konfiguration für das Massenladen von IBM® SPSS® Modeler in eine Teradata-Datenbank mithilfe der Optionen für das externe Ladeprogramm im Dialogfeld "DB-Export: Erweiterte Optionen" behilflich sein.

Sicherstellen, dass das Teradata-Dienstprogramm "fastload" installiert ist

Das Skript *teradata_loader.py* ruft das Teradata-Dienstprogramm *fastload* auf. Vergewissern Sie sich, dass *fastload* installiert und ordnungsgemäß auf dem Server konfiguriert ist, auf dem *teradata_loader.py* ausgeführt werden soll.

Massenladen von Daten ist nur in leere Tabellen möglich

Als Ziele für das Massenladen sind nur leere Tabellen möglich. Wenn eine Zieltabelle bereits vor dem Massenladevorgang Daten enthält, kann der Vorgang nicht durchgeführt werden.

Daten in den Formaten "Datum", "Zeit" und "Zeitstempel"

Setzen Sie in den Stream-Eigenschaften das Datumsformat auf JJJJ-MM-TT und das Zeitformat auf HH:MM:SS.

Leerstrings

Leerstrings werden als NULL-Werte in die Datenbank exportiert.

Teradata-Prozess-ID (tdpid)

Standardmäßig exportiert *fastload* Daten mit `tdpid=dbc` in das Teradata-System. Normalerweise gibt es einen Eintrag in der HOSTS-Datei, der `dbccop1` mit der IP-Adresse des Teradate-Servers verknüpft. Wenn Sie einen anderen Server verwenden möchten, geben Sie Folgendes im Feld Weitere Optionen für Ladeprogramm im Dialogfeld "DB-Export: Erweiterte Optionen" ein, um die `tdpid` dieses Servers weiterzuleiten:

```
-tdpid <id>
```

Leerzeichen in Tabellen- und Spaltennamen

Wenn ein Tabellen- oder Spaltenname Leerzeichen enthält schlägt der Massenladevorgang fehl. Benennen Sie nach Möglichkeit die Tabellen bzw. Spalten um, um die Leerzeichen zu entfernen.

Entwickeln von Massenladeprogrammen

In diesem Thema wird erläutert, wie Sie ein Massenladeprogramm entwickeln können, das über IBM® SPSS® Modeler ausgeführt werden kann, um Daten aus einer Textdatei in eine Datenbank zu laden.

Verwenden von Python zum Erstellen von Massenladeprogrammen

Standardmäßig sucht SPSS Modeler anhand des Datenbanktyps nach einem Standard-Massenladeprogramm. Unter Tabelle 7-1 auf S. 475.

Sie können das Skript *test_loader.py* zur Unterstützung der Entwicklung von Massenladeprogrammen verwenden. Für weitere Informationen siehe Thema Testen von Massenladeprogrammen auf S. 482.

An das Massenladeprogramm weitergeleitete Objekte

SPSS Modeler schreibt zwei Dateien, die an das Massenladeprogramm weitergeleitet werden.

- **Datendatei.** Enthält die zu ladenden Daten im Textformat.
- **Schemadatei.** Dies ist eine XML-Datei, die die Namen und Typen der Spalten beschreibt und Informationen darüber bereitstellt, wie die Datendatei formatiert ist (beispielsweise, welches Zeichen als Trennzeichen zwischen Feldern verwendet wird).

Zusätzlich leitet SPSS Modeler weitere Informationen, wie Tabellenname, Benutzername und Passwort, als Argumente weiter, wenn das Massenladeprogramm aufgerufen wird.

Hinweis: Um SPSS Modeler einen erfolgreichen Abschluss des Vorgangs zu signalisieren, sollte das Massenladeprogramm die Schemadatei löschen.

An das Massenladeprogramm weitergeleitete Argumente

Folgende Argumente werden an das Programm weitergeleitet.

Tabelle 7-2
An das Massenladeprogramm weitergeleitete Argumente

Argument	Beschreibung
schemafile	Pfad zur Schemadatei
data file	Pfad zur Datendatei
servername	Name des DBMS-Servers; kann leer sein
databasename	Name der Datenbank innerhalb des DBMS-Servers; kann leer sein
username	Benutzername für die Anmeldung bei der Datenbank
password	Passwort für die Anmeldung bei der Datenbank
tablename	Name der zu ladenden Tabelle
ownername	Name des Tabellenbesitzers (auch als Schemaname bezeichnet)
logfilename	Name der Protokolldatei (wenn leer, wird keine Protokolldatei erstellt)
rowcount	Anzahl der Zeilen im Daten-Set

Alle im Feld Weitere Optionen für Ladeprogramm im Dialogfeld “DB-Export: Erweiterte Optionen” angegebenen Optionen werden nach diesen Standardargumenten an das Massenladeprogramm weitergeleitet.

Format der Datendatei

Daten werden im Textformat in die Datendatei geschrieben. Dabei werden die einzelnen Felder durch ein Trennzeichen getrennt, das im Dialogfeld "DB-Export: Erweiterte Optionen" angegebenen wurde. Im Folgenden sehen Sie ein Beispiel für das Erscheinungsbild einer tabulatorgetrennten Datendatei.

```
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
```

Die Datei wird in der von IBM® SPSS® Modeler Server verwendeten lokalen Codierung geschrieben (bzw. von SPSS Modeler, wenn keine Verbindung zu SPSS Modeler Server besteht. Ein Teil der Formatierung wird über die Stream-Einstellungen von SPSS Modeler festgelegt.

Format der Schemadatei

Die Schemadatei ist eine XML-Datei, die zur Beschreibung der Datendatei dient. Im Folgenden sehen Sie ein Beispiel, das zur oben stehenden Datendatei gehört.

```
<?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>
```

In den folgenden Tabellen werden die Attribute der Elemente <table> und <column> der Schemadatei aufgeführt.

Tabelle 7-3
Attribute des <table>-Elements

Attribut	Beschreibung
delimiter	Das Feldtrennzeichen (Tabulator wird als "\t" dargestellt)
commit_every	Das Intervall für die Stapelgröße (wie im Dialogfeld "DB-Export: Erweiterte Optionen")
date_format	Das zur Darstellung von Datumswerten verwendete Format
time_format	Das zur Darstellung von Zeitwerten verwendete Format
append_existing	true, wenn die zu ladende Tabelle bereits Daten enthält; ansonsten false
delete_datafile	true, wenn das Massenladeprogramm die Datendatei nach Abschluss des Ladevorgangs löschen soll

Tabelle 7-4
Attribute des <column>-Elements

Attribut	Beschreibung
name	Der Spaltenname
encoded_name	Der Spaltenname, in die selbe Codierung konvertiert wie die Datendatei und ausgegeben als Reihe von zweistelligen Hexadezimalzahlen
type	Der Datentyp der Spalte: integer, real, char, time, date oder datetime
size	Für den Datentyp char die maximale Breite der Spalte in Zeichen

Testen von Massenladeprogrammen

Sie können die Massenladefunktion mithilfe des Testskripts *test_loader.py* testen, das im Ordner *\scripts* des Installationsverzeichnisses von IBM® SPSS® Modeler enthalten ist. Dies ist nützlich für Entwicklung, Debugging und Fehlerbehandlung von Massenladeprogrammen oder Skripts zur Verwendung mit SPSS Modeler.

Gehen Sie zur Verwendung des Testskripts wie folgt vor.

- Führen Sie das Skript *test_loader.py* aus, um die Schemadatei und die Datendatei in die Dateien *schema.xml* und *data.txt* zu kopieren und eine Windows-Batchdatei (*test.bat*) zu erstellen.
- Bearbeiten Sie die Datei *test.bat*, um das zu testende Massenladeprogramm bzw. Skript auszuwählen.
- Führen Sie die Datei *test.bat* über eine Befehls-Shell aus, um das ausgewählte Massenladeprogramm bzw. Skript zu testen.

Hinweis: Bei der Ausführung von *test.bat* werden nicht tatsächlich Daten in die Datenbank geladen.

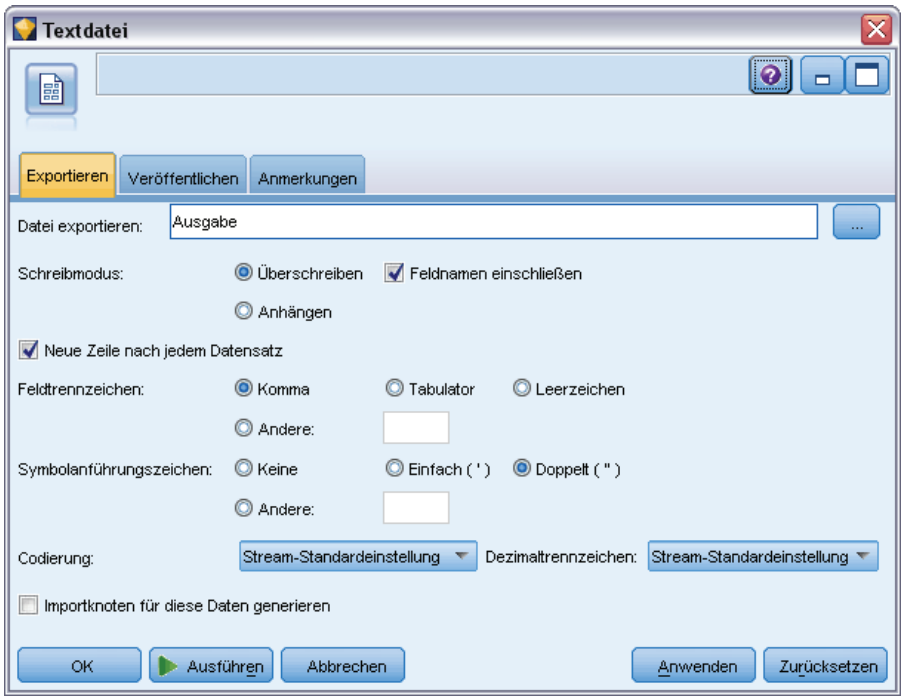
Textdatei-Exportknoten

Mit dem Knoten “Textdatei” schreiben Sie Daten in eine Textdatei, die mit Trennzeichen getrennt ist. Diese Vorgehensweise eignet sich für das Exportieren von Daten, die durch andere Analyse- oder Tabellenkalkulationsprogramme gelesen werden sollen.

Hinweis: Es ist nicht möglich, Dateien im alten Cacheformat zu schreiben, weil dieses Format nicht mehr für Cachedateien in IBM® SPSS® Modeler verwendet wird. Die SPSS Modeler-Cachedateien werden nunmehr im IBM® SPSS® Statistics-Format *.sav* gespeichert; dieses Format kann mit einem Statistics-Exportknoten geschrieben werden. Für weitere Informationen siehe Thema Statistikexportknoten in Kapitel 8 auf S. 514.

Registerkarte "Exportieren" beim Textdateiknoten

Abbildung 7-7
Knoten "Einfachdatei", Registerkarte "Exportieren"



Datei exportieren. Hier können Sie den Namen der Datei angeben. Geben Sie einen Dateinamen an oder klicken Sie auf die Feldauswahl-Schaltfläche und wechseln Sie zum Pfad der gewünschten Datei.

Schreibmodus. Wenn die Option Überschreiben aktiviert ist, werden alle vorhandenen Daten in der angegebenen Datei überschrieben. Ist die Option Anhängen aktiviert, wird die Ausgabe an die vorhandene Datei angehängt; die bereits vorhandenen Daten in dieser Datei werden also beibehalten.

- **Feldnamen einschließen.** Bei dieser Option werden die Dateinamen in die erste Zeile der Ausgabedatei geschrieben. Diese Option ist nur für den Schreibmodus Überschreiben verfügbar.

Neue Zeile nach jedem Datensatz. Bei dieser Option wird jeder Datensatz in eine eigene Zeile in der Ausgabedatei geschrieben.

Feldtrennzeichen. Dient zur Angabe des Zeichens, das als Trennzeichen zwischen den Feldwerten in der erzeugten Textdatei eingefügt werden soll. Die folgenden Optionen stehen zur Auswahl: Komma, Tabulator, Leerzeichen und Andere. Wenn Sie die Option Andere wählen, geben Sie das oder die gewünschten Trennzeichen in das Textfeld ein.

Symbolanführungszeichen. Hier können Sie die Art der Anführungszeichen angeben, die für Werte in symbolischen Feldern verwendet werden sollen. Die folgenden Optionen stehen zur Auswahl: Keine (die Werte werden nicht in Anführungszeichen eingeschlossen), Einfach (‘),

Doppelt (“) und Andere. Wenn Sie die Option Andere wählen, geben Sie das oder die gewünschten Anführungszeichen in das Textfeld ein.

Kodierung. Gibt die verwendete Textkodierungsmethode an. Sie haben die Wahl zwischen der System-Standardeinstellung, der Stream-Standardeinstellung und UTF-8.

- Die System-Standardeinstellung wird in der Windows-Systemsteuerung bzw. bei Ausführung im verteilten Modus auf dem Server-Computer angegeben.
- Der Stream-Standard wird im Dialogfeld “Stream-Eigenschaften” festgelegt.

Dezimaltrennzeichen. Hier können Sie das Trennzeichen für die Dezimalstellen in den Daten festlegen.

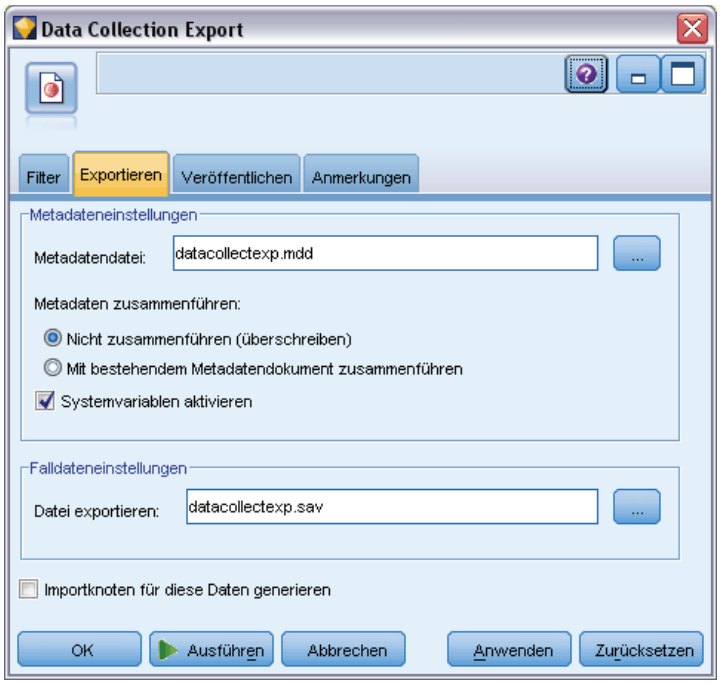
- **Stream-Standardeinstellung** Das Dezimaltrennzeichen, das durch die Standardeinstellung des aktuellen Streams definiert ist, wird verwendet. In der Regel ist dies das Dezimaltrennzeichen aus den Ländereinstellungen des Rechners.
- **Punkt (.).** Als Dezimaltrennzeichen wird ein Punkt verwendet.
- **Komma (,).** Als Dezimaltrennzeichen wird ein Komma verwendet.

Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen Quellenknotens für variable Dateien erzeugen, mit dem die exportierte Datendatei eingelesen wird. Für weitere Informationen siehe Thema Knoten “Datei (var.)” in Kapitel 2 auf S. 26.

IBM SPSS Data Collection-Exportknoten

Der IBM® SPSS® Data Collection-Exportknoten speichert Daten in dem von der Marktforschungssoftware Data Collection (beruht auf Data Collection Data Model) verwendeten Format. Bei diesem Format wird zwischen Falldaten – den tatsächlichen Antworten auf Fragen, die während einer Umfrage gesammelt werden – und Metadaten unterschieden, die beschreiben, wie der Fall gesammelt und organisiert wird. Metadaten bestehen aus Informationen wie Fragetexten, Variablennamen und -beschreibungen, Mehrfachantworten-Sets, Übersetzungen der verschiedenen Texte und der Definition der Struktur der Falldaten. Für weitere Informationen siehe Thema Data Collection Knoten in Kapitel 2 auf S. 35.

Abbildung 7-8
IBM SPSS Data Collection-Exportknoten, Registerkarte "Exportieren"



Hinweis: Für diesen Knoten ist Data Collection Data Model Version 4.0 oder höher erforderlich, das mit Software von Data Collection ausgeliefert wird. Weitere Informationen finden Sie auf der Data CollectionWebseite unter *http://www.ibm.com/software/analytics/spss/products/data-collection/*. Abgesehen von der Installation von Data Model, sind keine weiteren Konfigurationen erforderlich.

Metadaten-Datei. Gibt den Namen der Fragebogendefinitionsdatei (*.mdd*) an, in der die exportierten Metadaten gespeichert werden sollen. Auf der Grundlage der Informationen zum Feldtyp wird ein Standardfragebogen erstellt. So kann beispielsweise ein nominales (Set-) Feld als einzelne Frage dargestellt werden, wobei die Feldbeschreibung als Fragetext verwendet wird und für jeden definierten Wert ein gesondertes Kontrollkästchen vorhanden ist.

Metadaten zusammenführen. Hiermit können Sie angeben, ob die Metadaten die bestehenden Versionen überschreiben oder mit den bestehenden Metadaten zusammengeführt werden sollen. Wenn die Zusammenführungsoption ausgewählt wird, wird bei jeder Ausführung des Streams eine neue Version erstellt. Dadurch können die verschiedenen Versionen eines Fragebogens, der Änderungen unterzogen wird, dokumentiert werden. Jede Version lässt sich als Snapshot der für die Sammlung eines bestimmten Falldaten-Sets verwendeten Metadaten betrachten.

Systemvariablen aktivieren. Gibt an, ob Systemvariablen in die exportierte *.mdd*-Datei aufgenommen werden sollen. Dazu gehören Variablen wie *Respondent.Serial*, *Respondent.Origin* und *DataCollection.StartTime*.

Falldateneinstellungen. Gibt die IBM® SPSS® Statistics-Datendatei (*.sav*) an, in die die Falldaten exportiert werden sollen. Beachten Sie, dass hier alle Einschränkungen für Variablen- und Wertenamen gelten. So kann es beispielsweise erforderlich sein, auf die Registerkarte "Filter" zu wechseln und im Menü für Filteroptionen die Option "Umbenennen für SPSS Statistics" zu verwenden, um ungültige Zeichen in Feldnamen zu korrigieren.

Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen Data Collection-Quellenknoten erzeugen, mit dem die exportierte Datendatei eingelesen wird.

Mehrfachantworten-Sets. Etwaige im Stream definierte Mehrfachantworten-Sets bleiben beim Export der Datei automatisch erhalten. Mehrfachantworten-Sets können über jeden Knoten, der die Registerkarte "Filter" enthält, angezeigt und bearbeitet werden. Für weitere Informationen siehe Thema Bearbeiten von Mehrfachantworten-Sets in Kapitel 4 auf S. 160.

IBM Cognos BI-Exportknoten

Mit dem IBM Cognos BI-Exportknoten können Sie Daten aus einem IBM® SPSS® Modeler-Stream im UTF-8 Format in Cognos BI exportieren. Auf diese Weise kann Cognos BI transformierte oder gescorte Daten aus SPSS Modeler verwenden. Sie können mit Cognos BI Report Studio beispielsweise einen Bericht basierend auf den exportierten Daten erstellen, einschließlich Vorhersagen und Konfidenzwerten. Der Bericht könnte dann auf dem Cognos BI-Server gespeichert und an Cognos BI-Benutzer verteilt werden.

Hinweis: Sie können nur relationale Daten exportieren, keine OLAP-Daten.

Um Daten in Cognos BI zu exportieren, müssen Sie folgende Eingaben machen:

- Cognos-Verbindung – die Verbindung zum Cognos BI-Server
- ODBC-Verbindung – die Verbindung zum Cognos-Datenserver des Cognos BI-Servers

Innerhalb der Cognos-Verbindung geben Sie die zu verwendende Cognos-Datenquelle an. Diese Datenquelle muss dieselben Anmeldedaten verwenden wie die ODBC-Datenquelle.

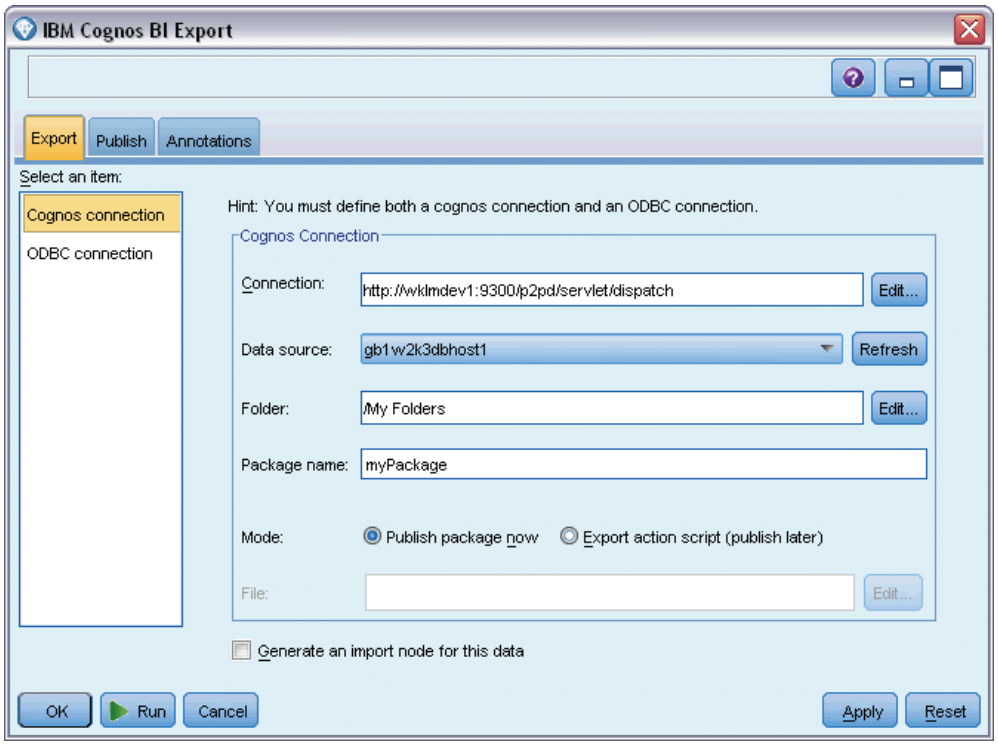
Sie exportieren die tatsächlichen Stream-Daten in den Daten-Server und die Paket-Metadaten in den Cognos BI-Server.

Wie bei allen Exportknoten können Sie auch die Registerkarte "Veröffentlichen" des Knoten-Dialogfelds verwenden, um den Stream mit IBM® SPSS® Modeler Solution Publisher zu veröffentlichen und bereitzustellen.

Cognos-Verbindung

Hier können Sie angeben, welche Verbindung zum Cognos BI-Server Sie für den Export verwenden möchten. Die Prozedur beinhaltet den Export der Metadaten in ein neues Paket auf dem Cognos BI-Server, während die Stream-Daten in den Cognos-Daten-Server exportiert werden.

Abbildung 7-9
Exportieren von Cognos-Daten



Verbindung. Klicken Sie auf die Schaltfläche Bearbeiten, um ein Dialogfeld anzuzeigen, in dem Sie die URL und andere Details für den Cognos BI-Server festlegen können, auf den die Daten exportiert werden sollen. Wenn Sie bereits bei einem Cognos BI-Server über IBM® SPSS® Modeler angemeldet sind, können Sie auch die Dateils der aktuellen Verbindung bearbeiten. Für weitere Informationen siehe Thema Cognos-Verbindungen in Kapitel 2 auf S. 48.

Datenquelle. Der Name der Cognos-Datenquelle (normalerweise eine Datenbank), in die Sie die Daten exportieren. Die Dropdown-Liste zeigt alle Cognos-Datenquellen an, auf die Sie über die aktuelle Verbindung zugreifen können. Klicken Sie zum Aktualisieren der Liste auf die Schaltfläche Aktualisieren.

Ordner. Der Pfad und Name des Ordners auf dem Cognos BI-Server, in dem das Export-Paket erstellt werden soll.

Paketname. Der Name des Pakets in einem angegebenen Ordner, der die exportierten Metadaten enthalten soll. Dabei muss es sich um neues Paket mit einem einzigen Abfragesubjekt handeln; Sie können nicht in vorhandene Pakete exportieren.

Modus. Legt fest, wie der Export durchgeführt werden soll:

- **Paket jetzt veröffentlichen.** (Standard) Führt den Exportvorgang aus, sobald Sie auf Ausführen klicken.
- **Aktionsskript exportieren.** Erstellt ein XML-Skript, das Sie später ausführen können (z. B. mit Framework Manager), um den Export durchzuführen. Geben Sie den Pfad und Datenamen für das Skript in das Feld Datei ein, oder klicken Sie auf die Schaltfläche Bearbeiten, um Namen und Speicherort der Skriptdatei festzulegen.

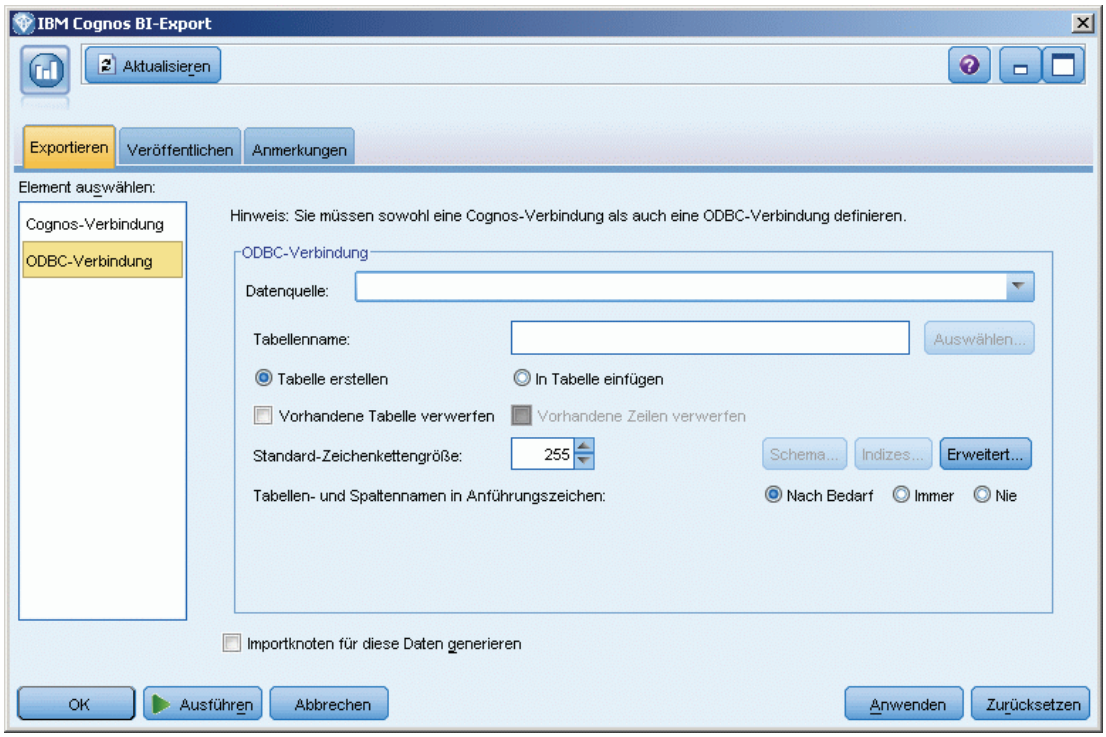
Importknoten für diese Daten generieren. Es wird ein Quellenknoten für die Daten erzeugt, die in die angegebene Datenquelle und Tabelle exportiert wurden. Wenn Sie auf Ausführen klicken, wird dieser Knoten in den Stream-Zeichenbereich aufgenommen.

ODBC-Verbindung

Hier geben Sie die Verbindung für den Cognos-Daten-Server (also die Datenbank) an, an den die Stream-Daten exportiert werden sollen.

Hinweis: Vergewissern Sie sich, dass die Datenquelle, die Sie hier angeben, auf dieselbe Datenquelle wie im Bereich Cognos-Verbindungen verweist. Außerdem müssen Sie sicherstellen, dass die Cognos-Datenquelle dieselben Anmeldedaten verwendet wie die ODBC-Datenquelle.

Abbildung 7-10
Exportieren von Cognos-Daten



Datenquelle. Zeigt die ausgewählte Datenquelle. Geben Sie den Namen ein oder wählen Sie einen Eintrag in der Dropdown-Liste aus. Wird die gewünschte Datenbank nicht in der Liste aufgeführt, wählen Sie Neue Datenbankverbindung hinzufügen und wechseln Sie im Dialogfeld "Datenbankverbindungen" zu dieser Datenbank. Für weitere Informationen siehe Thema Hinzufügen einer Datenbankverbindung in Kapitel 2 auf S. 18.

Tabellenname. Geben Sie den Namen der Tabelle ein, an die die Daten gesendet werden sollen. Bei der Option In Tabelle einfügen können Sie eine vorhandene Tabelle in der Datenbank auswählen, indem Sie auf die Schaltfläche Auswählen klicken.

Tabelle erstellen. Mit dieser Option können Sie eine neue Datenbanktabelle anlegen oder eine vorhandene Datenbanktabelle überschreiben.

In Tabelle einfügen. Mit dieser Option fügen Sie die Daten als neue Zeilen in eine vorhandene Datenbanktabelle ein.

Tabelle einlesen. (Wenn verfügbar) Aktivieren Sie diese Option, um ausgewählte Datenbankspalten mit Werten aus entsprechenden Quellendatenfeldern zu aktualisieren. Wenn Sie diese Option auswählen, wird die Schaltfläche Zusammenführen aktiviert, die ein Dialogfeld öffnet, in dem Sie Quellendatenfelder zu Datenbankspalten zuordnen können.

Vorhandene Tabelle verwerfen. Wenn Sie eine neue Tabelle erstellen, lassen Sie mit dieser Option alle vorhandenen Tabellen löschen, die denselben Namen besitzen wie die neu zu erstellende Tabelle.

Vorhandene Zeilen verwerfen. Wenn Sie Daten in eine Tabelle einfügen, lassen Sie mit dieser Option vorhandene Zeilen vor dem Exportieren aus der Tabelle löschen.

Hinweis: Bei den oben genannten beiden Optionen wird eine Überschreibungswarnung eingeblendet, sobald Sie den Knoten ausführen. Sollen diese Warnungen unterdrückt werden, deaktivieren Sie im Dialogfeld "Benutzeroptionen" auf der Registerkarte "Benachrichtigungen" die Option Warnen, wenn eine Datenbanktabelle durch einen Knoten überschrieben wird.

Standard-Zeichenkettengröße Felder, die Sie als "Ohne Typ" in einem aufwärts liegenden Typknoten gekennzeichnet haben, werden als Zeichenkettenfelder in die Datenbank geschrieben. Geben Sie die Größe der Zeichenketten an, die für Felder ohne Typ verwendet werden sollen.

Klicken Sie auf Schema, um ein Dialogfeld zu öffnen, in dem Sie verschiedene Exportoptionen festlegen können (für Datenbanken, die diese Funktion unterstützen), und geben Sie den Primärschlüssel für die Datenbankindizierung an. Für weitere Informationen siehe Thema Schemaoptionen für den Datenbankexport auf S. 465.

Klicken Sie auf Indizes, um Optionen für die Indizierung der exportierten Tabelle anzugeben und damit die Datenbankleistung zu verbessern. Für weitere Informationen siehe Thema Indexoptionen für den Datenbankexport auf S. 469.

Mit der Schaltfläche Erweitert können Sie Optionen für das Massenladen und die Datenbankübertragung festlegen. Für weitere Informationen siehe Thema Erweiterte Optionen für den Datenbankexport auf S. 472.

Tabellen- und Spaltennamen in Anführungszeichen. Wählen Sie die Optionen aus, die beim Senden der Anweisung CREATE TABLE an die Datenbank verwendet werden sollen. Enthält der Name von Tabellen und Spalten ein Leerzeichen oder ein Sonderzeichen, muss der Name in Anführungszeichen gesetzt werden.

- **Nach Bedarf.** Hiermit lassen Sie automatisch von Fall zu Fall durch IBM® SPSS® Modeler feststellen, ob Anführungszeichen erforderlich sind oder nicht.
- **Immer.** Die Tabellen- und Spaltennamen werden immer in Anführungszeichen eingeschlossen.
- **Nie.** Es werden keine Anführungszeichen verwendet.

Importknoten für diese Daten generieren. Es wird ein Quellenknoten für die Daten erzeugt, die in die angegebene Datenquelle und Tabelle exportiert wurden. Wenn Sie auf Ausführen klicken, wird dieser Knoten in den Stream-Zeichenbereich aufgenommen.

SAS-Exportknoten

Hinweis: Diese Funktion steht in SPSS Modeler Professional und SPSS Modeler Premium zur Verfügung.

Mit dem SAS-Exportknoten können Sie Daten im SAS-Format schreiben, die dann in SAS oder in SAS-kompatible Programme eingelesen werden können. Beim Export stehen drei SAS-Dateiformate zur Auswahl: SAS für Windows/OS2, SAS für UNIX sowie SAS Version 7/8.

Registerkarte "Exportieren" beim SAS-Exportknoten

Abbildung 7-11
SAS-Exportknoten, Registerkarte "Exportieren"

Datei exportieren. Geben Sie den Namen der Datei an. Geben Sie einen Dateinamen an oder klicken Sie auf die Feldauswahl-Schaltfläche und wechseln Sie zum Pfad der gewünschten Datei.

Exportieren. Legen Sie das Exportdateiformat fest. Die folgenden Optionen stehen zur Auswahl: SAS für Windows/OS2, SAS für UNIX sowie SAS Version 7/8.

Feldnamen als Variable exportieren. Wählen Sie die Optionen zum Exportieren der Feldnamen und Beschriftungen aus IBM® SPSS® Modeler, die in SAS genutzt werden sollen.

- **Namen.** Hiermit werden sowohl die Feldnamen als auch die Feldbeschriftungen aus SPSS Modeler exportiert. Die Namen werden als SAS-Variablennamen exportiert, die Beschriftungen entsprechend als SAS-Variablenlabels.
- **Labels.** Mit dieser Einstellung werden die SPSS Modeler-Feldnamen in SAS als Variablenlabels verwendet. Bei SPSS Modeler können verschiedene Zeichen in den Feldnamen verwendet werden, die bei SAS-Variablennamen nicht gültig sind. Um die mögliche Bildung ungültiger SAS-Namen zu vermeiden, wählen Sie stattdessen die Option Namen.

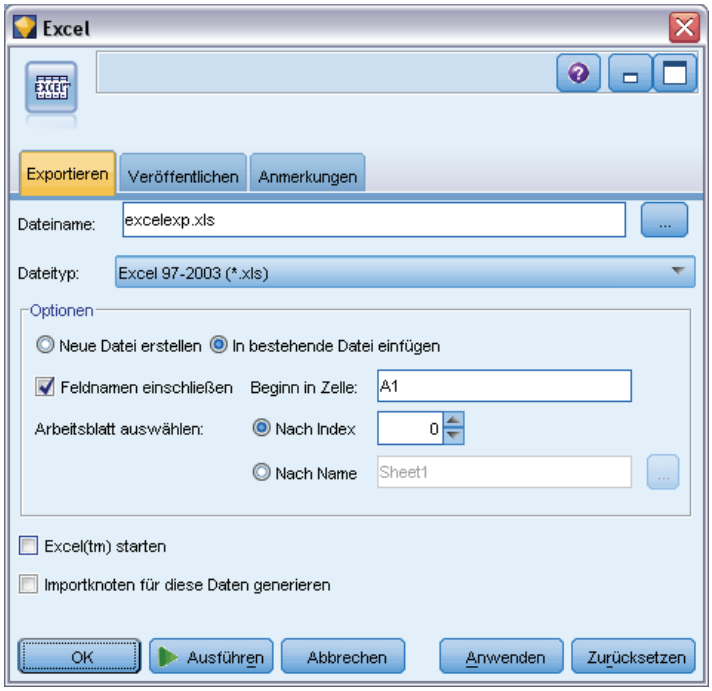
Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen SAS-Quellenknoten erzeugen, mit dem die exportierte Datendatei eingelesen wird. Für weitere Informationen siehe Thema SAS-Quellenknoten in Kapitel 2 auf S. 50.

Excel-Exportknoten

Der Excel-Exportknoten gibt Daten im Microsoft Excel-Format (*.xls*) aus. Optional können Sie auswählen, dass bei der Ausführung des Knotens Excel automatisch gestartet und die exportierte Datei geöffnet werden soll.

Registerkarte "Exportieren" beim Excel-Knoten

Abbildung 7-12
Excel-Exportknoten, Registerkarte "Exportieren"



Dateiname. Geben Sie einen Dateinamen an oder klicken Sie auf die Feldauswahl-Schaltfläche und wechseln Sie zum Pfad der gewünschten Datei. Der Standarddateiname lautet *excelexp.xls*.

Dateityp. Wählen Sie den Excel-Dateityp, den Sie exportieren möchten.

Neue Datei erstellen. Erstellt eine neue Excel-Datei.

In vorhandene Datei einfügen. Der Inhalt wird ersetzt, beginnend in der Zelle, die durch das Feld Start in Zelle angegeben ist. Andere Zellen im Arbeitsblatt behalten ihren ursprünglichen Inhalt.

Feldnamen einschließen. Gibt an, ob Feldnamen in die erste Zeile des Arbeitsblattes eingefügt werden sollen.

Start in Zelle. Die für den ersten Exportdatensatz (bzw. den ersten Feldnamen, falls Feldnamen einschließen aktiviert ist) verwendete Zellenposition. Daten werden ab dieser Anfangszelle nach rechts und unten gefüllt.

Arbeitsblatt wählen. Legt das Arbeitsblatt fest, an das Sie die Daten exportieren möchten. Sie können das Arbeitsblatt nach Index oder nach Name identifizieren:

- **Nach Index.** Wenn Sie eine neue Datei anlegen, bestimmen Sie eine Zahl zwischen 0 und 9, um das Arbeitsblatt zu identifizieren, das Sie exportieren möchten, beginnend mit 0 für das erste Arbeitsblat, 1 für das zweite usw. Werte von 10 und darüber können Sie nur verwenden, wenn an dieser Position bereits ein Arbeitsblatt vorhanden ist.
- **Nach Namen.** Wenn Sie eine neue Datei anlegen, bestimmen Sie den Namen, der für das Arbeitsblatt verwendet wird. Beim Einfügen in eine bestehende Datei werden die Daten in dieses Arbeitsblatt eingefügt, falls vorhanden, andernfalls wird ein neues Arbeitsblatt mit diesem Namen erstellt.

Excel starten. Gibt an, ob Excel bei der Ausführung des Knotens automatisch für die exportierte Datei gestartet werden soll. Beachten Sie, dass bei der Ausführung im verteilten Modus für IBM® SPSS® Modeler Server die Ausgabe im Dateisystem des Servers gespeichert und Excel auf dem Client mit einer Kopie der exportierten Datei gestartet wird.

Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen Excel-Quellenknoten erzeugen, mit dem die exportierte Datendatei eingelesen wird. Für weitere Informationen siehe Thema Excel-Quellenknoten in Kapitel 2 auf S. 52.

XML-Exportknoten

Mit dem XML-Exportknoten können Sie unter Verwendung der UTF-8-Kodierung Daten im XML-Format ausgeben. Optional können Sie einen XML-Quellenknoten erstellen, um die exportierten Daten wieder in der Stream einzulesen.

Abbildung 7-13
XML-Daten exportieren

XML-Exportdatei. Der vollständige Pfad und Dateiname der XML-Datei, an die Sie die Daten exportieren möchten.

XML-Schema verwenden. Aktivieren Sie dieses Kontrollkästchen, wenn Sie ein Schema oder DTD verwenden möchten, um die Struktur der exportierten Daten zu kontrollieren. Dadurch wird die unten beschriebene Schaltfläche Zuordnen aktiviert.

Wenn Sie kein Schema oder DTD verwenden, wird die folgende Standardstruktur für die exportierten Daten verwendet:

```
<Datensätze>
  <Datensatz>
    <Feldname1>value</Feldname1>
    <Feldname2>value</Feldname2>
  :
    <FeldnameN>value</FeldnameN>
  </Datensatz>
  <Datensatz>
  :
  :
  </Datensatz>
  :
  :
</Datensätze>
```

Leerfelder in einem Feldnamen werden durch Unterstriche ersetzt, so wird zum Beispiel “Mein Feld” zu <My_Field>.

Zuordnen. Falls Sie ein XML-Schema verwenden, öffnet diese Schaltfläche ein Dialogfeld, in dem Sie angeben können, welcher Teil der XML-Struktur für den Beginn jedes neuen Datensatzes verwendet werden soll. Für weitere Informationen siehe Thema XML-Zuordnungsdatensätze - Optionen auf S. 495.

Zugeordnete Felder. Zeigt die Anzahl der Felder an, die zugeordnet wurden

Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen XML-Quellenknoten erzeugen, mit dem die exportierte Datendatei wieder in den Stream eingelesen wird. Für weitere Informationen siehe Thema XML-Quellenknoten in Kapitel 2 auf S. 53.

XML-Daten schreiben

Wenn ein XML-Element angegeben wird, wird der Feldwert im Element-Tag platziert:

```
<Element>value</Element>
```

Wenn ein Attribut zugeordnet wird, wird der Feldwert als Wert für das Attribut platziert:

```
<Element Attribut="value">
```

Wenn ein Feld einem Element oberhalb des Elements <records> zugeordnet wird, wird das Feld nur einmal beschrieben und fungiert als Konstante für alle Datensätze. Der Wert für dieses Element kommt aus dem ersten Datensatz.

Wenn ein Nullwert geschrieben werden muss, geschieht dies durch Angeben von leerem Inhalt. Bei Elementen ist das:

<Element></Element>

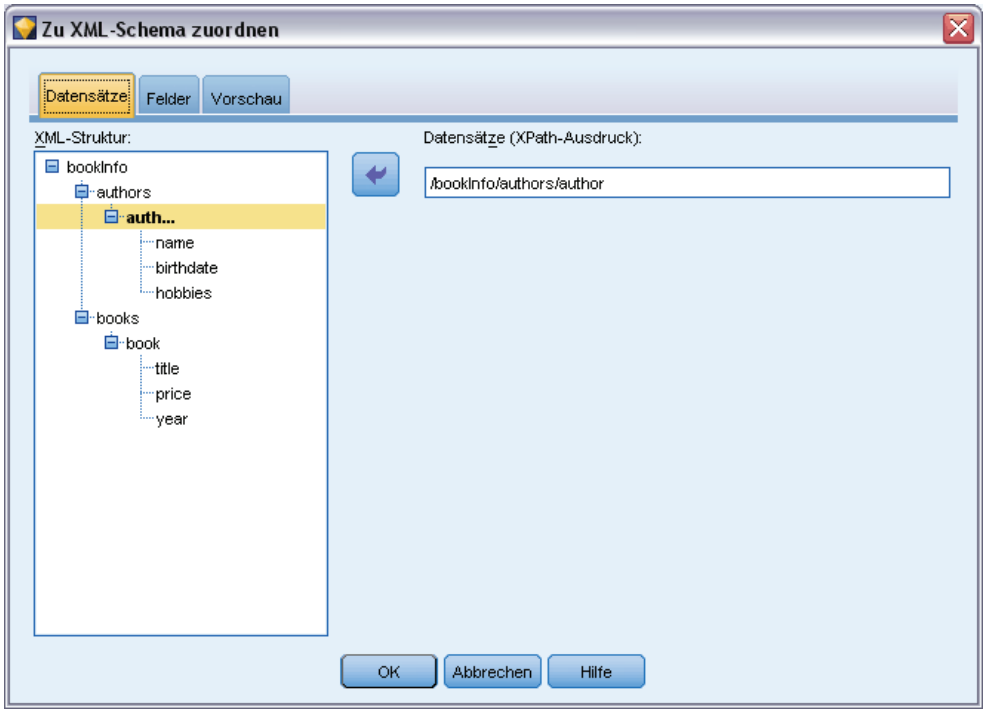
Bei Attributen ist es:

<Element Attribut="">

XML-Zuordnungsdatensätze - Optionen

Auf der Registerkarte “Datensätze” können Sie angeben, welcher Teil der XML-Struktur für den Beginn jedes neuen Datensatzes verwendet werden soll. Um korrekte Zuordnungen auf ein Schema durchführen zu können, müssen Sie das Datensatztrennzeichen festlegen.

Abbildung 7-14
XML-Zuordnungsdatensätze



XML-Struktur. Ein hierarchischer Baum zeigt die Sturktur des XML-Schemas, das auf dem vorangegangenen Bildschirm festgelegt wurde.

Datensätze (XPath-Ausdruck). Zum Festlegen des Datensatztrennzeichens wählen Sie ein Element in der XML-Struktur aus und klicken Sie auf die Rechtspfeil-Schaltfläche. Jedes Mal, wenn dieses Element in den Quelldaten gefunden wird, wird in der Ausgabedatei ein neuer Datensatz erstellt.

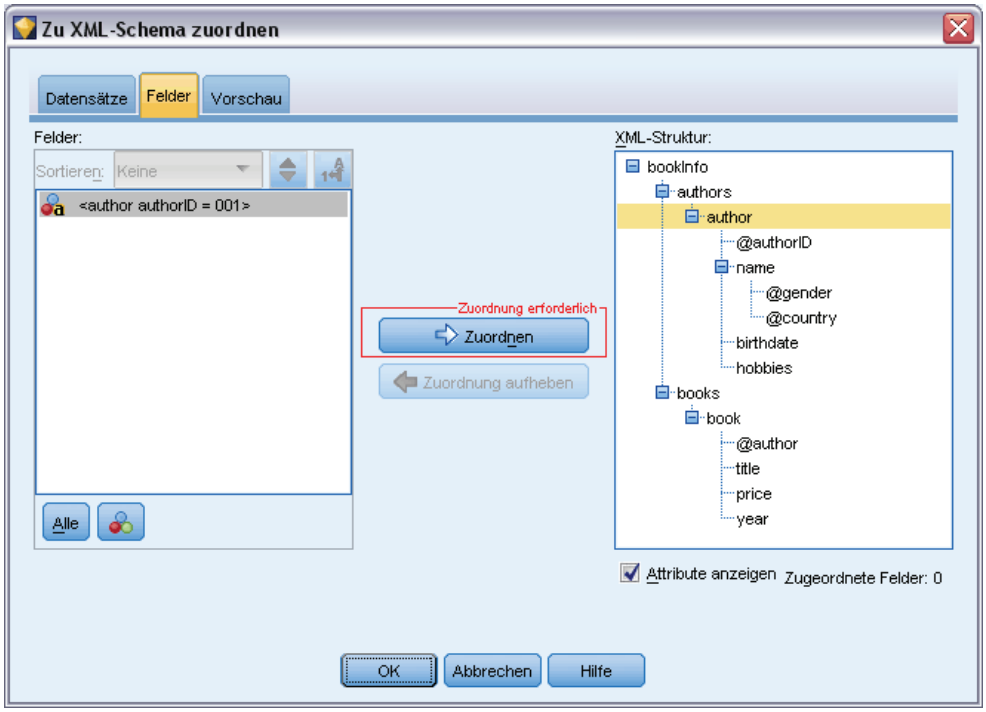
Hinweis: Wenn Sie das Wurzelelement in der XML-Struktur auswählen, kann nur ein einziger Datensatz geschrieben werden, während alle anderen Datensätze übergangen werden.

XML-Zuordnungsfelder - Optionen

Die Registerkarte “Felder” dient zum Zuordnen von Feldern im Daten-Set zu Elementen oder Attributen in der XML-Struktur, wenn eine Schemadatei verwendet wird.

Feldnamen, die einem Element- oder Attributnamen entsprechen, werden automatisch zugeordnet, so lange der Element- oder Attributname eindeutig ist. Wenn es also sowohl ein Element als auch ein Attribut mit dem Namen field1 gibt, kann keine automatische Zuordnung erfolgen. Wenn es nur ein Objekt in der Struktur mit der Bezeichnung field1 gibt, wird ein Feld mit diesem Namen in dem Stream automatisch zugeordnet.

Abbildung 7-15
XML-Zuordnungsfelder



Felder. Die Liste der Felder im Modell. Wählen Sie eines oder mehrere Felder als Quellenteil der Zuordnung aus. Sie können die Schaltflächen am Ende der Liste verwenden, um alle Felder oder alle Felder mit einem bestimmten Messniveau auszuwählen.

XML-Struktur. Wählen Sie ein Element in der XML-Struktur als Zuordnungsziel aus. Klicken Sie auf “Zuordnen”, um die Zuordnung zu erstellen. Anschließend wird die Zuordnung angezeigt. Die Anzahl der so zugeordneten Felder wird unterhalt dieser Liste angezeigt.

Um eine Zuordnung aufzuheben, wählen Sie das Objekt in der XML-Struktur aus und klicken Sie auf Zuordnung aufheben.

Attribute anzeigen. Zeigt die Attribute der XML-Elemente in der XML-Struktur an oder blendet diese aus, sofern vorhanden.

XML-Zuordnungsvorschau

Klicken Sie auf der Registerkarte “Vorschau” auf Aktualisieren, um eine Vorschau der XML-Datei zu sehen, die geschrieben wird.

Falls die Zuordnung nicht korrekt ist, kehren Sie zur Registerkarte “Datensätze” oder “Felder” zurück, um die Fehler zu korrigieren, und klicken erneut auf Aktualisieren, um sich das Ergebnis anzusehen.

Kapitel 8

IBM SPSS Statistics-Knoten

IBM SPSS Statistics-Knoten – Überblick

Zur Ergänzung von IBM® SPSS® Modeler und seinen Data-Mining-Funktionen bietet Ihnen IBM® SPSS® Statistics die Möglichkeit, weiterführende statistische Analysen durchzuführen und Daten zu verwalten.

Wenn Sie eine kompatible, lizenzierte Kopie von SPSS Statistics installiert haben, können Sie von SPSS Modeler eine Verbindung aufbauen und komplexe, aus mehreren Schritten bestehende Datenänderungen und Analysen ausführen, die andernfalls von SPSS Modeler nicht unterstützt werden. Für den erfahrenen Benutzer gibt es auch die Option, die Analysen mithilfe von Befehlssyntax weiter anzupassen. In den Versionshinweisen finden Sie Informationen zur Kompatibilität von Versionen.

Wenn verfügbar, werden die SPSS Statistics-Knoten auf einem eigenen Teil der Knotenpalette angezeigt.

Hinweis: Es empfiehlt sich, dass Sie Ihre Daten in einem Typenknoten instanziieren, bevor Sie die SPSS Statistics-Transformations-, Modell- oder Ausgabeknoten verwenden. Dies ist auch eine Voraussetzung für die Verwendung des Syntaxbefehls AUTORECODE.

Die SPSS Statistics-Palette enthält die folgenden Knoten:

Der Statistikdateiknoten liest Daten aus dem Dateiformat *.sav* ein, das von SPSS Statistics verwendet wird, sowie in SPSS Modeler gespeicherte Cache-Dateien, die ebenfalls dasselbe Format verwenden. Für weitere Informationen siehe Thema Statistikdateiknoten auf S. 499.

Der Statistiktransformationsknoten führt eine Auswahl von SPSS Statistics-Syntaxbefehlen an Datenquellen in SPSS Modeler aus. Für diesen Knoten ist eine lizenzierte Kopie von SPSS Statistics erforderlich. Für weitere Informationen siehe Thema Statistiktransformationsknoten auf S. 501.

Mithilfe des Knotens “Statistikmodell” können Sie Ihre Daten analysieren und bearbeiten, indem Sie SPSS Statistics-Prozeduren ausführen, die PMML erzeugen. Für diesen Knoten ist eine lizenzierte Kopie von SPSS Statistics erforderlich. Für weitere Informationen siehe Thema Statistikmodellknoten auf S. 505.

Mit dem Statistikausgabeknoten können Sie eineSPSS Statistics-Prozedur aufrufen, um Ihre SPSS Modeler-Daten zu analysieren. Es stehen zahlreiche SPSS Statistics-Analyseprozeduren zur Verfügung. Für diesen Knoten ist eine lizenzierte Kopie von SPSS Statistics erforderlich. Für weitere Informationen siehe Thema Statistikausgabeknoten auf S. 509.

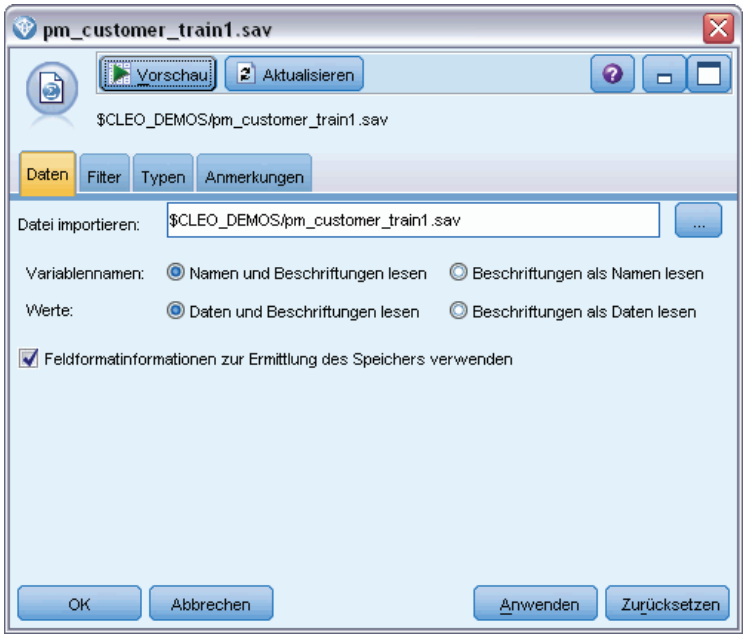
Der Statistikexportknoten gibt Daten im Format SPSS Statistics.*sav* aus. Die *.sav*-Dateien können von SPSS Statistics Base und anderen Produkten gelesen werden. Dieses Format wird auch für Cache-Dateien in SPSS Modeler verwendet. Für weitere Informationen siehe Thema Statistikexportknoten auf S. 514.

Hinweis: Wenn Ihre Kopie von SPSS Statistics nur für einen einzigen Benutzer lizenziert ist und Sie einen Stream mit mindestens zwei Verzweigungen ausführen, von denen jede einen SPSS Statistics-Knoten enthält, erhalten Sie möglicherweise einen Lizenzierungsfehler. Dieser Fall tritt dann ein, wenn die SPSS Statistics-Sitzung für eine Verzweigung noch nicht beendet wurde, bevor die Sitzung für eine andere Verzweigung zu starten versucht. Überarbeiten Sie den Stream nach Möglichkeit so, dass nicht mehrere Verzweigungen mit SPSS Statistics-Knoten parallel ausgeführt werden.

Statistikdateiknoten

Mit dem Statistikdateiknoten können Sie Daten direkt aus einer gespeicherten IBM® SPSS® Statistics-Datei (*.sav*) lesen. Dieses Format ersetzt nun die Cache-Datei aus früheren Versionen von IBM® SPSS® Modeler. Wenn Sie eine gespeicherte Cache-Datei importieren möchten, verwenden Sie am besten den SPSS Statistics-Dateiknoten.

Abbildung 8-1
Importieren einer .sav-Datei



Datei importieren. Geben Sie den Namen der Datei an. Zur Auswahl einer Datei können Sie einen Dateinamen eingeben oder auf die Schaltfläche mit den Auslassungspunkten (...) klicken. Der Dateipfad wird angezeigt, sobald Sie eine Datei ausgewählt haben.

Variablennamen. Wählen Sie eine Methode zur Behandlung von Variablennamen und -beschriftungen beim Importieren aus einer SPSS Statistics-Datei *.sav*. Metadaten, die Sie hier einschließen, bleiben während Ihrer Arbeit in SPSS Modeler erhalten und können zur Verwendung in SPSS Statistics wieder exportiert werden.

- **Namen und Beschriftungen lesen.** Wählen Sie diese Option, wenn sowohl Variablennamen als auch -beschriftungen in SPSS Modeler eingelesen werden sollen. Standardmäßig ist diese Option ausgewählt und Variablennamen werden im Typknoten angezeigt. Beschriftungen können je nach den im Dialogfeld "Stream-Eigenschaften" angegebenen Optionen in Diagrammen, Modellbrowsern und anderen Ausgabearten angezeigt werden. Standardmäßig ist die Anzeige von Beschriftungen in der Ausgabe deaktiviert.
- **Beschriftungen als Namen lesen.** Wählen Sie diese Option, um statt der kurzen Feldnamen die beschreibenden Variablenlabels aus der SPSS Statistics-Datei (*.sav*-Datei) zu lesen und diese Beschriftungen als Variablennamen in SPSS Modeler zu verwenden.

Werte. Wählen Sie eine Methode zur Behandlung von Werten und Beschriftungen beim Importieren aus einer SPSS Statistics-Datei *.sav*. Metadaten, die Sie hier einschließen, bleiben während Ihrer Arbeit in SPSS Modeler erhalten und können zur Verwendung in SPSS Statistics wieder exportiert werden.

- **Daten und Beschriftungen lesen.** Wählen Sie diese Option, um sowohl die tatsächlichen Werte als auch die Wertelabels in SPSS Modeler einzulesen. Standardmäßig ist diese Option ausgewählt und die Werte werden im Typknoten angezeigt. Wertelabels können je nach den im Dialogfeld "Stream-Eigenschaften" angegebenen Optionen in Expression Builder, Diagrammen, Modellbrowsern und anderen Ausgabearten angezeigt werden.
- **Beschriftungen als Daten lesen.** Wählen Sie diese Option, wenn Sie statt der numerischen oder symbolischen Codes, mit denen die Werte dargestellt werden, die Wertelabels aus der Datei *.sav* verwenden möchten. Bei Auswahl dieser Option z. B. für Daten mit dem Feld "Geschlecht", dessen Werte 1 und 2 für *männlich* und *weiblich* stehen, wird das Feld in eine Zeichenkette konvertiert und *männlich* und *weiblich* werden als tatsächliche Werte importiert.

  Vor Auswahl dieser Option müssen Sie Ihre SPSS Statistics-Daten auf fehlende Werte prüfen. Wenn ein numerisches Feld beispielsweise Beschriftungen nur für fehlende Werte verwendet (0 = *No Answer*, –99 = *Unknown*), werden bei Auswahl der obigen Option nur die Wertelabels *No Answer* und *Unknown* importiert und das Feld in eine Zeichenkette konvertiert. In diesem Fall sollten Sie die Werte selbst importieren und fehlende Werte in einem Typknoten festlegen.

Speichertyp anhand Feldformatinformationen bestimmen. Wenn dieses Kontrollkästchen markiert ist, werden Feldwerte, die in der *.sav*-Datei als Ganzzahlen formatiert sind (d. h. Felder, die in der Variablenansicht in SPSS Statistics als F*n*.0 angegeben sind), mit dem Speichertyp "Ganze Zahl" importiert. Alle übrigen Feldwerte mit Ausnahme von Zeichenketten werden als reelle Zahlen importiert.

Wenn das Kontrollkästchen unmarkiert ist (Standard), werden alle Feldwerte außer Zeichenketten als reelle Zahlen importiert, unabhängig davon, ob sie in der *.sav*-Datei als Ganzzahlen formatiert sind.

Mehrfachantworten-Sets. Etwaige in der SPSS Statistics-Datei definierte Mehrfachantworten-Sets bleiben beim Import der Datei automatisch erhalten. Mehrfachantworten-Sets können über jeden Knoten, der die Registerkarte "Filter" enthält, angezeigt und bearbeitet werden. Für weitere Informationen siehe Thema Bearbeiten von Mehrfachantworten-Sets in Kapitel 4 auf S. 160.

Statistiktransformationsknoten

Mit dem Statistiktransformationsknoten können Sie Datentransformationen mithilfe von IBM® SPSS® Statistics-Befehlssyntax durchführen. Dadurch kann eine Reihe von Transformationen durchgeführt werden, die von IBM® SPSS® Modeler nicht unterstützt werden, und die Automatisierung komplexer, aus mehreren Schritten bestehenden Transformationen ist möglich, einschließlich der Erstellung von Feldern aus einem einzelnen Knoten. Er ähnelt dem Statistikausgabeknoten, mit der Ausnahme, dass die Daten zur weiteren Analyse an SPSS Modeler ausgegeben werden, wohingegen die Daten beim Ausgabeknoten als angeforderte Ausgabeobjekte ausgegeben werden, beispielsweise als Diagramme oder Tabellen.

Um diesen Knoten verwenden zu können, muss eine kompatible Version von SPSS Statistics auf Ihrem Computer installiert und lizenziert sein. Für weitere Informationen siehe Thema IBM SPSS Statistics-Hilfsprogramme in Kapitel 6 auf S. 458. In den Versionshinweisen finden Sie Informationen zur Kompatibilität.

Falls erforderlich, können Sie mithilfe der Registerkarte "Filter" Felder filtern oder umbenennen, sodass sie den SPSS Statistics-Benennungsstandards entsprechen. Für weitere Informationen siehe Thema Umbenennen oder Filtern von Feldern für IBM SPSS Statistics auf S. 516.

Befehlssyntaxreferenz. Einzelheiten zu bestimmten SPSS Statistics-Prozeduren finden Sie im Handbuch zur *SPSS Statistics-Befehlssyntaxreferenz*, die in Ihrer Kopie der SPSS Statistics-Software inbegriffen ist. Um das Handbuch anzuzeigen, wählen Sie in der Registerkarte "Syntax" die Option Syntaxeditor und klicken Sie auf die SchaltflächeSPSS Statistics-Syntaxhilfe starten.

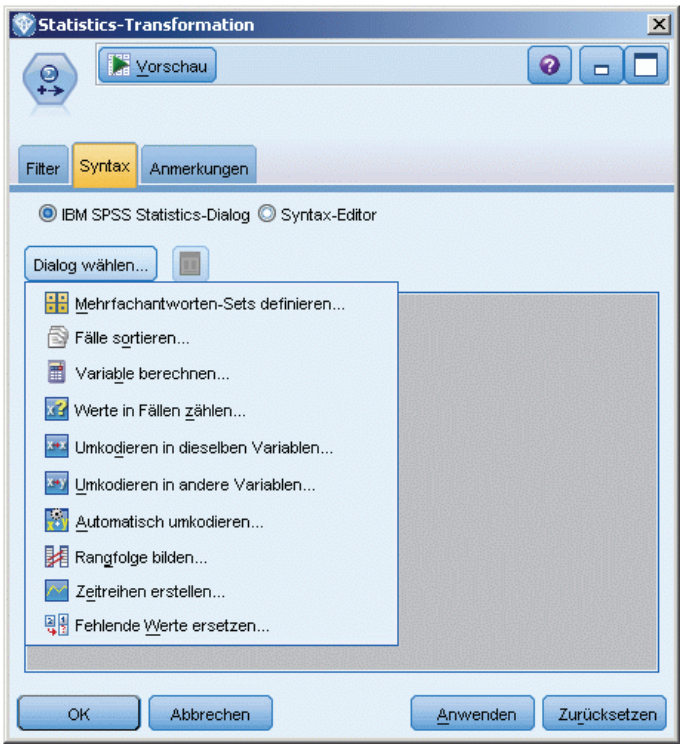
Hinweis: Dieser Knoten unterstützt nicht die gesamte SPSS Statistics-Syntax. Für weitere Informationen siehe Thema Zulässige Syntax auf S. 503.

Statistiktransformationsknoten - Registerkarte "Syntax"

IBM SPSS Statistics-Dialogoption

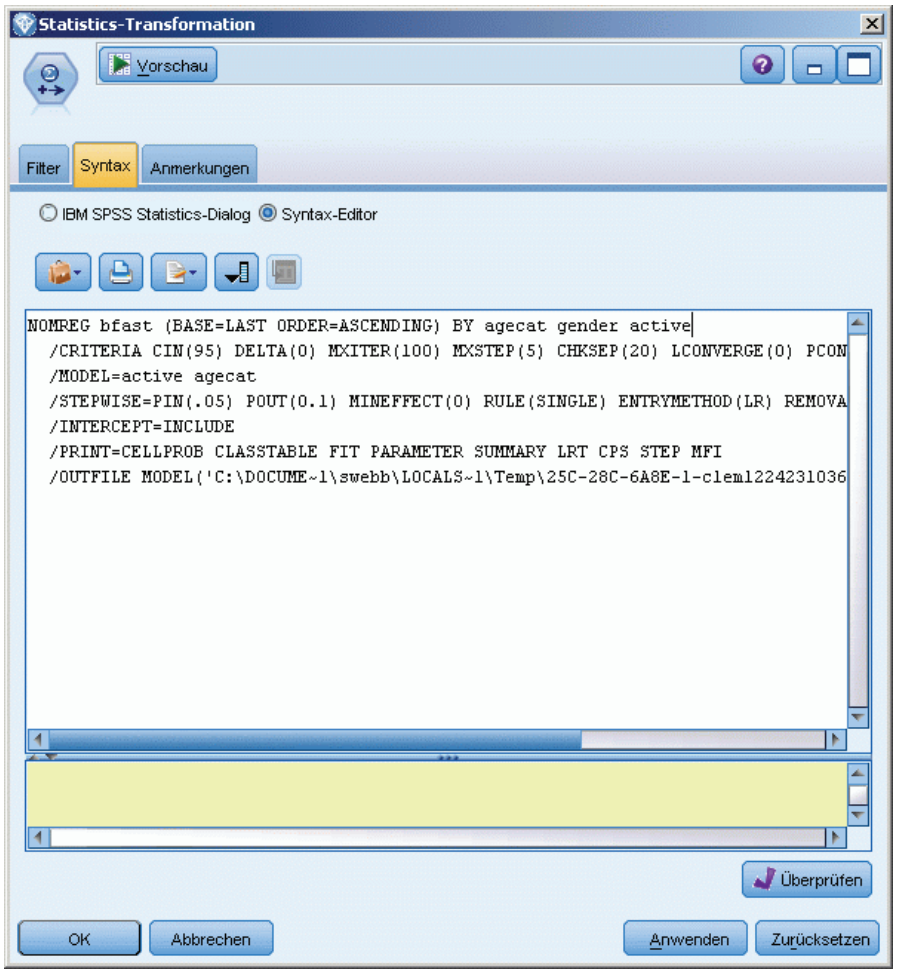
Wenn Sie nicht mit der IBM® SPSS® Statistics-Syntax für eine Prozedur vertraut sind, ist dies die einfachste Methode zur Erstellung von Syntax in IBM® SPSS® Modeler: Wählen Sie die Option IBM SPSS Statistics-Dialog, wählen Sie das Dialogfeld für die Prozedur aus, füllen Sie das Dialogfeld aus und klicken Sie auf "OK". Dadurch wird die Syntax auf der Registerkarte "Syntax" des SPSS Statistics-Knotens abgelegt, den Sie in SPSS Modeler verwenden. Anschließend können Sie den Stream ausführen, um die Ausgabe aus der Prozedur zu erhalten.

Abbildung 8-2
Statistiktransformationsknoten, Dialogfeldauswahl



IBM SPSS Statistics-Option Syntaxeditor

Abbildung 8-3
Statistiktransformationsknoten, Syntaxeditor



Überprüfen. Nachdem Sie Ihre Syntaxbefehle im oberen Bereich des Dialogfelds eingegeben haben, können Sie mit dieser Schaltfläche Ihre Einträge überprüfen. Etwaige falsche Syntax wird im unteren Teil des Dialogfelds angegeben.

Um sicherzustellen, dass die Überprüfung nicht zu lange dauert, wird bei der Syntaxvalidierung eine Überprüfung anhand einer repräsentativen Stichprobe Ihrer Daten durchgeführt, um sicherzustellen, dass die Einträge gültig sind. Auf eine Überprüfung anhand des gesamten Daten-Sets wird verzichtet.

Zulässige Syntax

Wenn Sie viel alte Syntax aus IBM® SPSS® Statistics verwenden oder mit den Datenvorbereitungsfunktionen von SPSS Statistics vertraut sind, können Sie viele Ihrer bestehenden Transformationen mithilfe des Statistiktransformationsknotens ausführen. Grob gesagt, ermöglicht der Knoten die Transformation von Daten auf vorhersehbare Weise,

beispielsweise durch die Ausführung von Schleifenbefehlen oder durch Ändern, Hinzufügen, Sortieren, Filtern oder Auswählen von Daten.

Hier einige Beispiele für Befehle, die ausgeführt werden können:

- Berechnung von Zufallszahlen gemäß einer Binomialverteilung:

```
COMPUTE newvar = RV.BINOM(10000,0.1)
```

- Umkodieren einer Variablen in eine neue Variable:

```
RECODE Age (Lowest thru 30=1)  (30 thru 50=2)  (50 thru Highest=3)  INTO  AgeRecoded
```

- Ersetzen fehlender Werte:

```
RMV Age_1=SMEAN(Age)
```

Die vom Statistiktransformationsknoten unterstützte SPSS Statistics-Syntax ist in folgender Tabelle aufgelistet:

Befehlsname

```
ADD VALUE LABELS
APPLY DICTIONARY
AUTORECODE
BREAK
CD
CLEAR MODEL PROGRAMS
CLEAR TIME PROGRAM
CLEAR TRANSFORMATIONS
COMPUTE
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
```

Befehlsname

INSERT
LEAVE
LOOP-END LOOP
MATRIX-END MATRIX
MISSING VALUES
N OF CASES
NUMERIC
PERMISSIONS
PRESERVE
RANK
RECODE
RENAME VARIABLES
RESTORE
RMV
SAMPLE
SELECT IF
SET
SORT CASES
SORT CASES
STRING
SUBTITLE
TEMPORARY
TITLE
UPDATE
V2C
VALIDATEDATA
VALUE LABELS
VARIABLE ATTRIBUTE
VARSTOCASES
VECTOR

Statistikmodellknoten

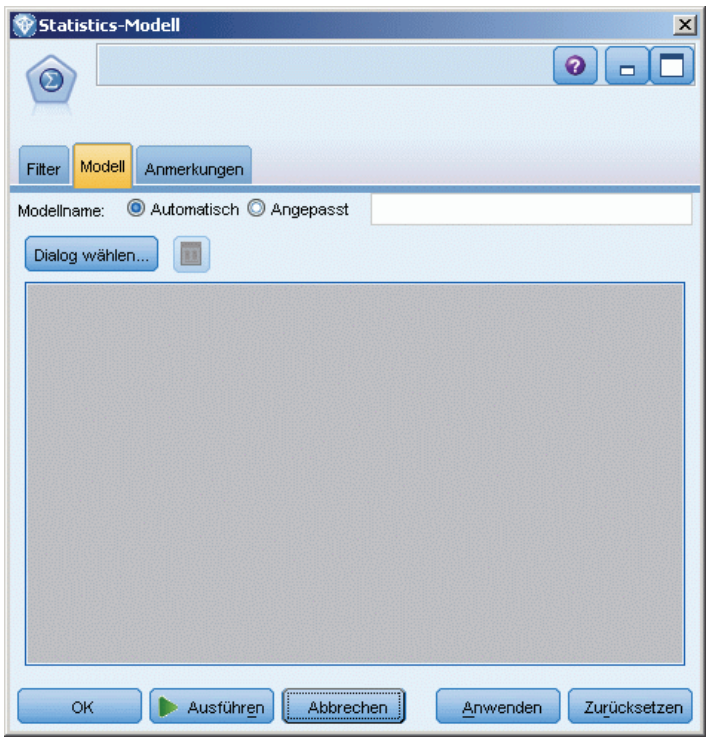
Mithilfe des Statistikmodellknotens können Sie Ihre Daten analysieren und bearbeiten, indem Sie IBM® SPSS® Statistics-Prozeduren ausführen, die PMML erzeugen. Die Modell-Nuggets, die Sie erzeugen, können dann wie üblich in IBM® SPSS® Modeler-Streams zum Scoring usw. verwendet werden.

Um diesen Knoten verwenden zu können, muss eine kompatible Version von SPSS Statistics auf Ihrem Computer installiert und lizenziert sein. Für weitere Informationen siehe Thema IBM SPSS Statistics-Hilfsprogramme in Kapitel 6 auf S. 458. In den Versionshinweisen finden Sie Informationen zur Kompatibilität.

Die verfügbaren SPSS Statistics-Analyseprozeduren hängen von Ihrer Lizenz ab.

Statistikmodellknoten - Registerkarte "Modell"

Abbildung 8-4
Statistikmodellknoten - Registerkarte "Modell"

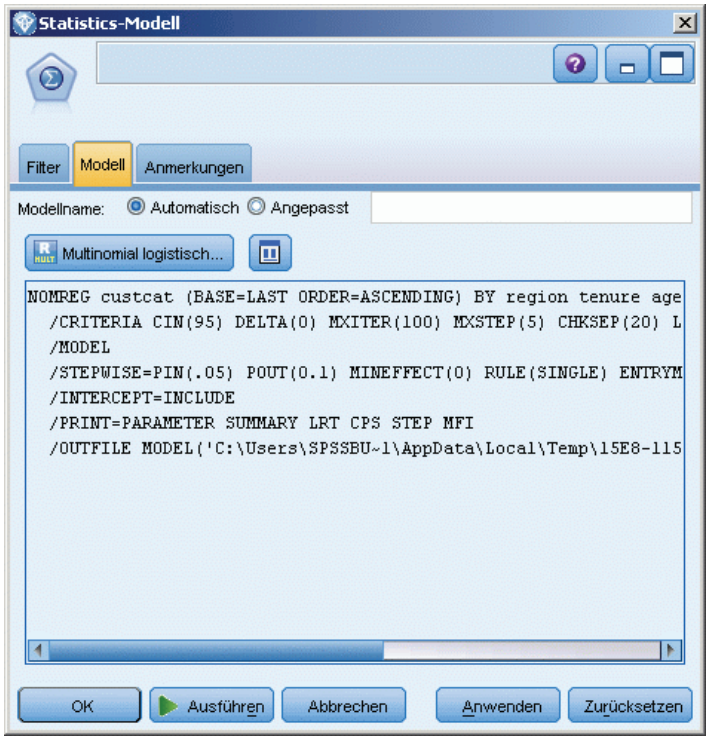


Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Wählen Sie ein Dialogfeld. Klicken Sie, um eine Liste verfügbarer IBM® SPSS® Statistics-Prozeduren anzuzeigen, die Sie auswählen und ausführen können. In der Liste werden nur die Prozeduren aufgeführt, die PMML produzieren und für die Sie eine Lizenz besitzen. Nicht enthalten sind benutzerdefinierte Prozeduren.

- Klicken Sie auf die gewünschte Prozedur. Das entsprechende SPSS Statistics-Dialogfeld wird geöffnet.
- Geben Sie im SPSS Statistics-Dialogfeld die Details für die Prozedur ein.
- Klicken Sie auf OK, um in den Statistikmodellknoten zurückzukehren; die SPSS Statistics-Syntax wird in der Registerkarte "Modell" angezeigt.

Abbildung 8-5
In der Registerkarte "Modell" angezeigte Syntax

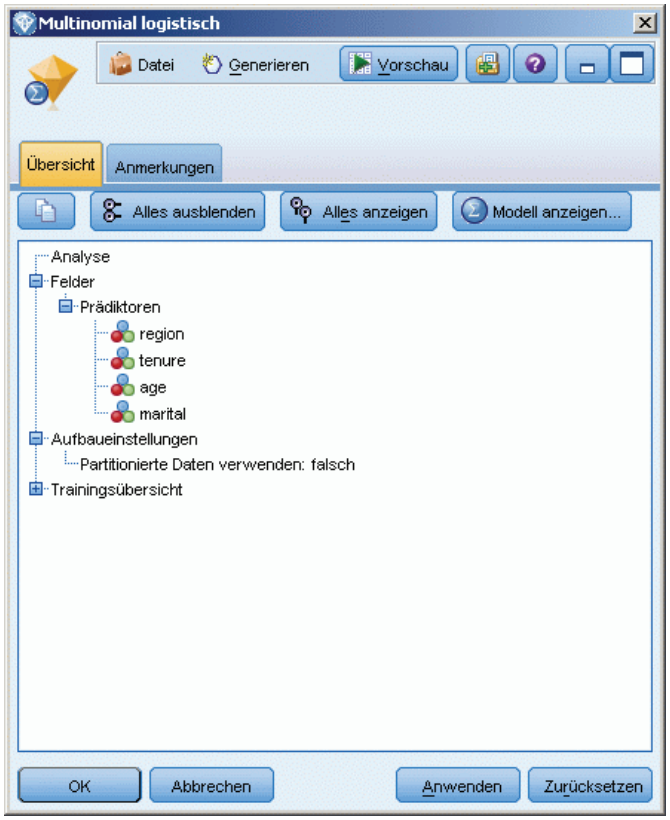


- Um zu einem beliebigen Zeitpunkt in das SPSS Statistics-Dialogfeld zurückzukehren, z. B. um Ihre Abfrage zu ändern, klicken Sie auf die Anzeigeschaltfläche für das SPSS Statistics-Dialogfeld rechts neben der Schaltfläche zur Prozedurauswahl.

Statistikmellknoten - Modell-Nugget-Übersicht

Wenn Sie den Statistikmodellknoten ausführen, führt dieser die zugehörige IBM® SPSS® Statistics-Prozedur aus und erstellt ein Modell-Nugget, das Sie zum Scoring in IBM® SPSS® Modeler-Streams verwenden können.

Abbildung 8-6
Statistikmodell-Nugget – Registerkarte "Übersicht"

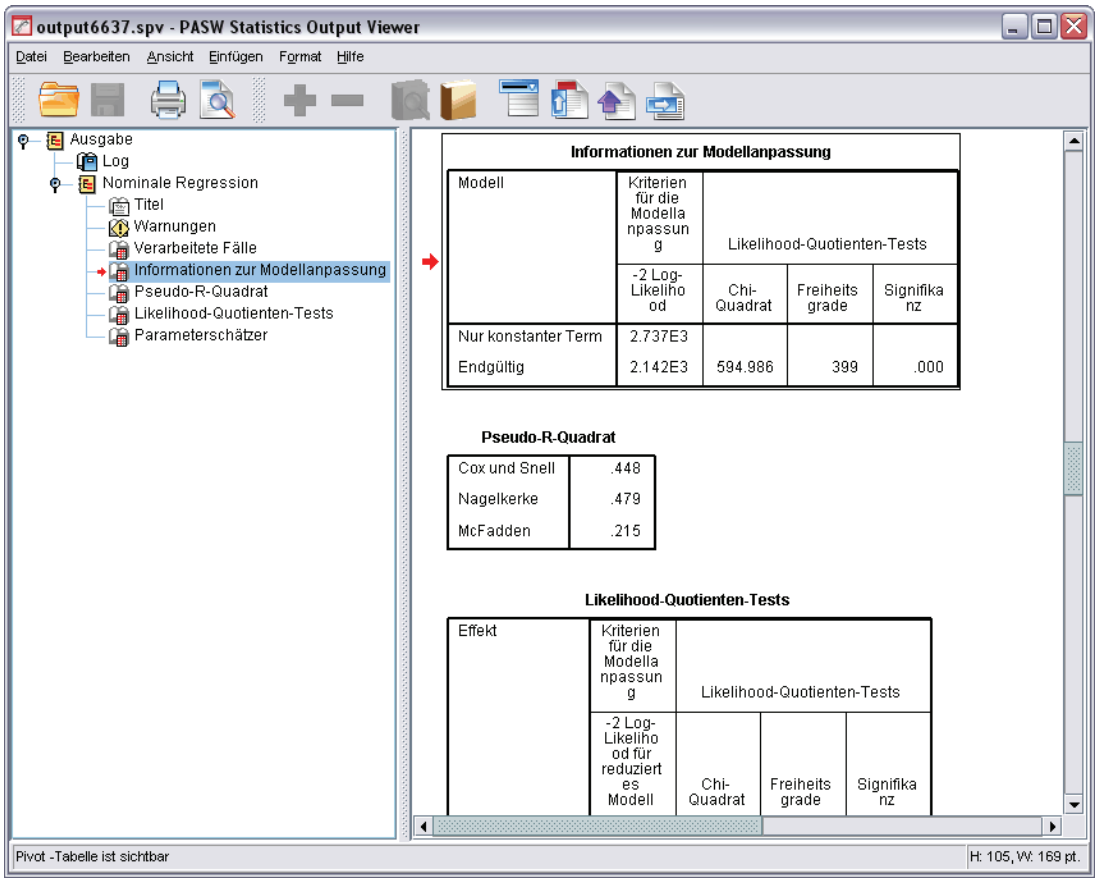


Auf der Registerkarte "Übersicht" für das Modell-Nugget werden Informationen über die Felder, die Aufbaueinstellungen und die Modellschätzung angezeigt. Die Ergebnisse werden in einer Baumansicht dargestellt, die durch Klicken auf bestimmte Elemente erweitert bzw. reduziert werden kann.

Die Schaltfläche Modell anzeigen zeigt die Ergebnisse in einer modifizierten Variante des SPSS Statistics-Ausgabe-Viewer. Weitere Informationen zu diesem Viewer finden Sie in der SPSS Statistics-Dokumentation.

Das Menü "Datei" enthält die üblichen Befehle zum Exportieren und Drucken. Für weitere Informationen siehe Thema Anzeigen der Ausgabe in Kapitel 6 auf S. 399.

Abbildung 8-7
Statistikmodell-Nugget – Registerkarte "Erweitert"



Statistikausgabeknoten

Mit dem Statistikausgabeknoten können Sie eineIBM® SPSS® Statistics-Prozedur aufrufen, um Ihre IBM® SPSS® Modeler-Daten zu analysieren. Lassen Sie die Ergebnisse in einem Browser-Fenster anzeigen oder speichern Sie sie im SPSS Statistics-Ausgabedateiformat. In SPSS Statistics stehen zahlreiche SPSS Modeler-Analyseprozeduren zur Verfügung.

Um diesen Knoten verwenden zu können, muss eine kompatible Version von SPSS Statistics auf Ihrem Computer installiert und lizenziert sein. Für weitere Informationen siehe Thema IBM SPSS Statistics-Hilfsprogramme in Kapitel 6 auf S. 458. In den Versionshinweisen finden Sie Informationen zur Kompatibilität.

Falls erforderlich, können Sie mithilfe der Registerkarte "Filter" Felder filtern oder umbenennen, sodass sie den SPSS Statistics-Benennungsstandards entsprechen. Für weitere Informationen siehe Thema Umbenennen oder Filtern von Feldern für IBM SPSS Statistics auf S. 516.

Befehlssyntaxreferenz. Einzelheiten zu bestimmten SPSS Statistics-Prozeduren finden Sie im Handbuch zur *SPSS Statistics-Befehlssyntaxreferenz*, die in Ihrer Kopie der SPSS Statistics-Software inbegriffen ist. Um das Handbuch anzuzeigen, wählen Sie in der Registerkarte

"Syntax" die Option Syntaxeditor und klicken Sie auf die SchaltflächeSPSS Statistics-Syntaxhilfe starten.

Statistikausgabeknoten - Registerkarte "Syntax"

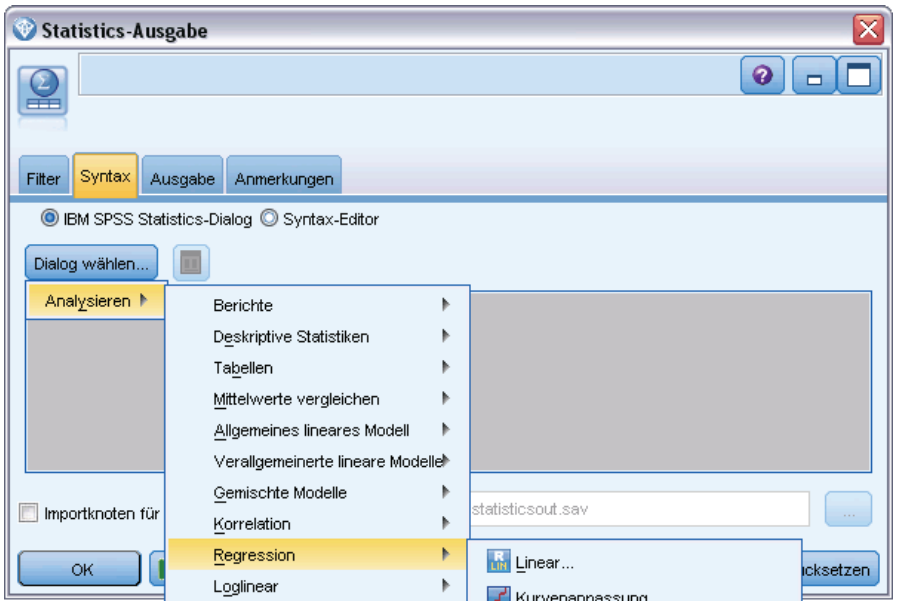
Mit dieser Registerkarte können Sie die Syntax für die IBM® SPSS® Statistics-Prozedur erstellen, mit der Sie Ihre Daten analysieren möchten. Die Syntax besteht aus zwei Teilen: die **Anweisung** und die zugehörigen **Optionen**. Die Anweisung bezeichnet die auszuführende Analyse oder Option sowie die zu verwendenden Felder. In den Optionen sind alle anderen Angaben festgelegt, z. B. die anzuzeigende Statistik oder die zu speichernden abgeleiteten Felder.

IBM SPSS Statistics-Dialogoption

Wenn Sie nicht mit der SPSS Statistics-Syntax für eine Prozedur vertraut sind, ist dies die einfachste Methode zur Erstellung von Syntax in IBM® SPSS® Modeler: Wählen Sie die Option IBM SPSS Statistics-Dialog, wählen Sie das Dialogfeld für die Prozedur aus, füllen Sie das Dialogfeld aus und klicken Sie auf "OK". Dadurch wird die Syntax auf der Registerkarte "Syntax" des SPSS Statistics-Knotens abgelegt, den Sie in SPSS Modeler verwenden. Anschließend können Sie den Stream ausführen, um die Ausgabe aus der Prozedur zu erhalten.

Sie können optional einen Statistikdatei-Quellenknoten zum Importieren der resultierenden Daten generieren. Dies ist beispielsweise dann nützlich, wenn eine Prozedur zusätzlich zur Anzeige der Ausgabe Felder, beispielsweise für Scores, in das aktive Daten-Set schreibt.

Abbildung 8-8
Statistikausgabeknoten, Dialogfeldauswahl



So erstellen Sie die Syntax:

► Klicken Sie auf die Schaltfläche Dialogfeld auswählen.

▶ Wählen Sie eine der Optionen aus:

- **Analysieren.** Listet den Inhalt des SPSS Statistics-Analysemenüs auf; wählen Sie die Prozedur aus, die Sie verwenden möchten.
- **Sonstige.** Wenn diese Option angezeigt wird, werden dort Dialogfelder aufgelistet, die Sie mit dem Custom Dialog Builder in SPSS Statistics erstellt haben, sowie andere SPSS Statistics-Dialogfelder, die nicht im Analysemenü erscheinen und für die Sie eine Lizenz besitzen. Wenn keine Dialogfelder zur Verfügung stehen, wird diese Option nicht angezeigt.

Hinweis: Die Dialogfelder zur automatischen Datenaufbereitung werden nicht angezeigt.

Bei einem benutzerdefinierten SPSS Statistics-Dialogfeld, das neue Felder erstellt, können diese Felder nicht in SPSS Modeler verwendet werden, da es sich bei dem Statistikausgabeknoten um einen Endknoten handelt.

▶ Optional können Sie das Kontrollkästchen Importknoten für resultierende Daten generieren aktivieren, um einen Statistikdatei-Quellenknoten zu erstellen, mit dem die resultierenden Daten in einen anderen Stream importiert werden können. Der Knoten wird auf dem Bildschirm-Zeichenbereich abgelegt, wobei die in der *.sav*-Datei enthaltenen Daten im Feld Datei angegeben werden (Standardspeicherort ist das Installationsverzeichnis von SPSS Modeler).

Option Syntaxeditor

Gehen Sie wie folgt vor, um Syntax zu speichern, die für eine häufig verwendete Prozedur erstellt wurde:

▶ Klicken Sie auf die Schaltfläche “Dateioptionen” (die erste in der Symbolleiste).

▶ Wählen Sie im Menü Speichern oder Speichern unter.

▶ Speichern Sie die Datei im Format *sps*.

Gehen Sie wie folgt vor, um früher erstellte Syntaxdateien zu verwenden und den aktuellen Inhalt des Syntaxeditors zu ersetzen, falls vorhanden:

▶ Klicken Sie auf die Schaltfläche “Dateioptionen” (die erste in der Symbolleiste).

▶ Wählen Sie Öffnen im Menü aus.

▶ Wählen Sie eine *.sps*-Datei aus, um den Inhalt dieser Datei in die Registerkarte “Syntax” für den Ausgabeknoten einzufügen.

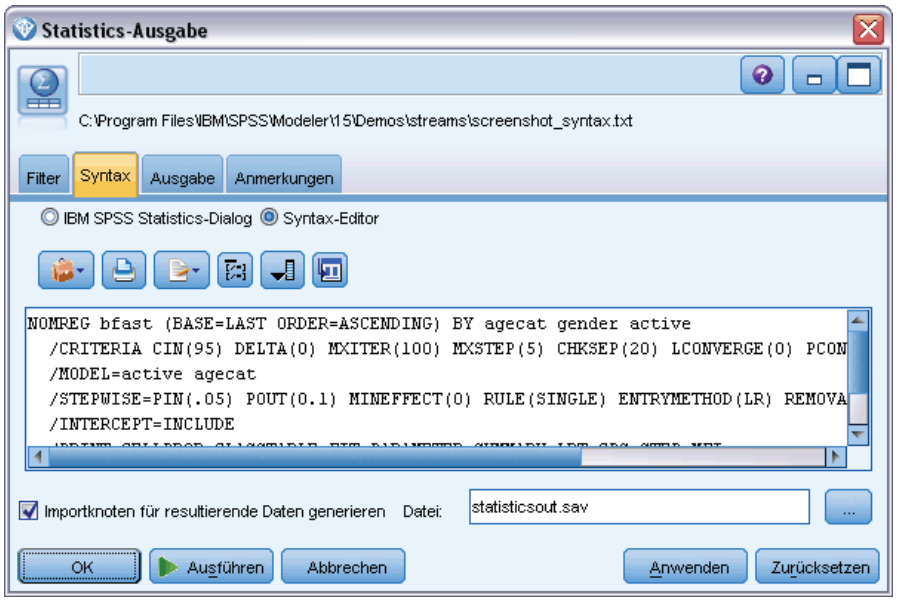
Gehen Sie wie folgt vor, um früher gespeicherte Syntax ohne Ersetzen des aktuellen Inhalts einzufügen:

▶ Klicken Sie auf die Schaltfläche “Dateioptionen” (die erste in der Symbolleiste).

▶ Wählen Sie Einfügen im Menü aus.

▶ Wählen Sie eine *.sps*-Datei aus, um den Inhalt dieser Datei für den Ausgabeknoten an der Cursorposition einzufügen.

Abbildung 8-9
Statistikausgabeknoten, Syntaxeditor



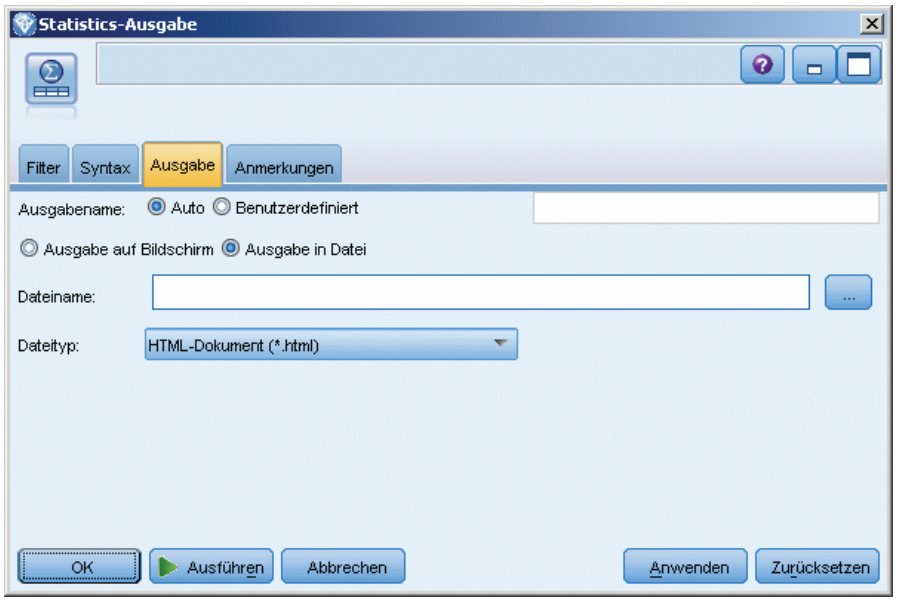
- Optional können Sie das Kontrollkästchen Importknoten für resultierende Daten generieren aktivieren, um einen Statistikdatei-Quellenknoten zu erstellen, mit dem die resultierenden Daten in einen anderen Stream importiert werden können. Der Knoten wird auf dem Bildschirm-Zeichenbereich abgelegt, wobei die in der *.sav*-Datei enthaltenen Daten im Feld Datei angegeben werden (Standardspeicherort ist das Installationsverzeichnis von SPSS Modeler).

  Beim Klicken auf Ausführen werden die Ergebnisse im SPSS Statistics-Ausgabe-Viewer angezeigt. Weitere Informationen zum Viewer finden Sie in der SPSS Statistics-Dokumentation.

Statistikausgabeknoten - Registerkarte "Ausgabe"

Auf der Registerkarte "Ausgabe" legen Sie Format und Position der Ausgabe fest. Sie können auswählen, dass die Ergebnisse auf dem Bildschirm angezeigt werden sollen, oder sie an einen der verfügbaren Dateitypen senden.

Abbildung 8-10
Statistikausgabeknoten - Registerkarte "Ausgabe"



Ausgabename. Bestimmt den Namen der Ausgabe, die beim Ausführen des Knotens erstellt wird. Mit Auto wird ein Name auf der Grundlage des Knotens bestimmt, mit dem die Ausgabe erzeugt wird. Optional können Sie auch Angepasst auswählen und einen anderen Namen angeben.

Ausgabe auf Bildschirm (Standardeinstellung)**.** Erstellt ein Ausgabeobjekt für die Online-Anzeige. Das Ausgabeobjekt wird auf der Registerkarte "Ausgaben" im Manager-Fenster dargestellt, wenn der Ausgabeknoten ausgeführt wird.

Ausgabe in Datei. Speichert die Ausgabe in einer Datei, wenn der Knoten ausgeführt wird. Wenn Sie diese Option wählen, geben Sie einen Dateinamen im Feld Dateiname an (oder wechseln Sie zu einem Verzeichnis und geben Sie einen Dateinamen mithilfe der Feldauswahl-Schaltfläche an) und wählen Sie einen Dateityp aus.

Dateityp. Wählen Sie den Dateiyp, an den Sie die Ausgabe senden möchten.

- **HTML-Dokument (*.html).** Schreibt die Ausgabe im HTML-Format.
- **SPSS Statistics Viewer-Datei (*.spv).** Schreibt die Ausgabe in einem Format, das vom IBM® SPSS® Statistics-Ausgabe-Viewer gelesen werden kann.
- **SPSS Statistics Web Reports-Datei (*.spw).** Schreibt die Ausgabe in einem SPSS Statistics Web Reports-Format, das in einem IBM SPSS Collaboration and Deployment Services-Repository veröffentlicht und anschließend in einem Webbrowser angezeigt werden kann. Für weitere Informationen siehe Thema Im Web veröffentlichen in Kapitel 6 auf S. 400.

Statistikexportknoten

Mit dem Statistikexportknoten können Sie die Daten im IBM® SPSS® Statistics-Format *.sav* exportieren. SPSS Statistics.*sav*-Dateien können von SPSS Statistics Base und anderen Modulen gelesen werden. Dieses Format wird auch für die IBM® SPSS® Modeler-Cache-Dateien verwendet.

Beim Zuordnen von SPSS Modeler-Feldnamen zu SPSS Statistics-Variablennamen entsteht hin und wieder ein Fehler, weil die SPSS Statistics-Variablennamen maximal 64 Zeichen umfassen können und bestimmte Zeichen wie beispielsweise Leerzeichen, Dollarzeichen ($) und Gedankenstriche (–) nicht zulässig sind. Diese Einschränkungen können auf zweierlei Weise umgangen werden:

- Benennen Sie die Felder so um, dass die Namen den Anforderungen für SPSS Statistics-Variablennamen genügen. Klicken Sie hierzu auf die Registerkarte “Filter”. Für weitere Informationen siehe Thema Umbenennen oder Filtern von Feldern für IBM SPSS Statistics auf S. 516.
- Legen Sie fest, dass sowohl die Feldnamen als auch die Beschriftungen aus SPSS Modeler exportiert werden sollen.

Hinweis: SPSS Modeler schreibt *.sav*-Dateien in Unicode UTF-8-Format. SPSS Statistics unterstützt nur Dateien in Unicode UTF-8-Format aus Version 16.0 und höher. Um die Möglichkeit beschädigter Daten zu vermeiden, sollten *.sav*-Dateien nicht in SPSS Statistics-Versionen vor 16.0 verwendet werden. Weitere Informationen finden Sie in der Hilfe zu SPSS Statistics.

Mehrfachantworten-Sets. Etwaige im Stream definierte Mehrfachantworten-Sets bleiben beim Export der Datei automatisch erhalten. Mehrfachantworten-Sets können über jeden Knoten, der die Registerkarte “Filter” enthält, angezeigt und bearbeitet werden. Für weitere Informationen siehe Thema Bearbeiten von Mehrfachantworten-Sets in Kapitel 4 auf S. 160.

Statistikexportknoten - Registerkarte "Exportieren"

Abbildung 8-11
Statistikexportknoten, Registerkarte "Exportieren"

Datei exportieren. Hier können Sie den Namen der Datei angeben. Geben Sie einen Dateinamen an oder klicken Sie auf die Feldauswahl-Schaltfläche und wechseln Sie zum Pfad der gewünschten Datei.

Feldnamen als Variable exportieren. Dient zur Angabe einer Methode, wie Variablennamen und Beschriftungen beim Exportieren aus IBM® SPSS® Modeler in eine Datei im IBM® SPSS® Statistics-Format *.sav* behandelt werden sollen.

- **Namen.** Hiermit werden sowohl die Feldnamen als auch die Feldbeschriftungen aus SPSS Modeler exportiert. Die Namen werden als SPSS Statistics-Variablennamen exportiert, die Beschriftungen entsprechend als SPSS Statistics-Variablenlabels.
- **Labels.** Mit dieser Einstellung werden die SPSS Modeler-Feldnamen in SPSS Statistics als Variablenlabels verwendet. Bei SPSS Modeler können verschiedene Zeichen in den Feldnamen verwendet werden, die bei SPSS Statistics-Variablennamen nicht gültig sind. Um die mögliche Bildung ungültiger SPSS Statistics-Namen zu vermeiden, wählen Sie stattdessen die Option Labels oder passen die Feldnamen auf der Registerkarte "Filter" an.

Anwendung starten. Wenn SPSS Statistics oder AnswerTree auf dem Computer installiert ist, können Sie die Anwendung mit dieser Option direkt für die gespeicherte Datendatei aufrufen. Die Optionen zum Starten der Anwendung müssen im Dialogfeld "Hilfsprogramme" angegeben werden. Für weitere Informationen siehe Thema IBM SPSS Statistics-Hilfsprogramme in Kapitel 6 auf S. 458.Soll lediglich eine Datei im SPSS Statistics-Format *.sav* erstellt werden, ohne ein externes Programm zu öffnen, deaktivieren Sie diese Option.

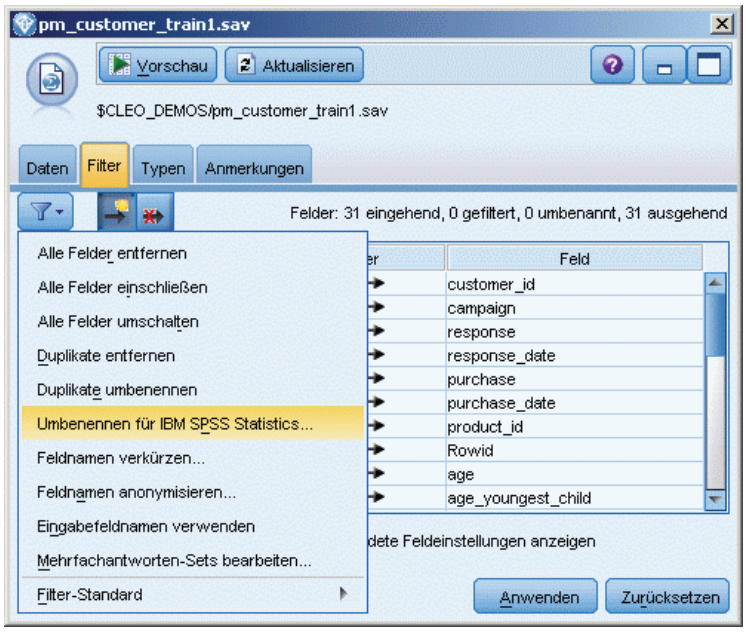
Importknoten für diese Daten generieren. Mit dieser Option lassen Sie automatisch einen Quellenknoten für eine Statistikdatei erzeugen, mit dem die exportierte Datendatei eingelesen wird. Für weitere Informationen siehe Thema Statistikdateiknoten auf S. 499.

Umbenennen oder Filtern von Feldern für IBM SPSS Statistics

Vor dem Exportieren oder Bereitstellen von Daten aus IBM® SPSS® Modeler in externe Anwendungen wie IBM® SPSS® Statistics müssen die Feldnamen ggf. umbenannt oder angepasst werden. Die Dialogfelder "Statistiktransformation", "Statistikausgabe" und "Statistikexport" beinhalten jeweils die Registerkarte "Filter", mit der dieser Vorgang erleichtert wird.

Eine ausführliche Beschreibung der Funktionen auf der Registerkarte "Filter" finden Sie an anderer Stelle in diesem Handbuch. Für weitere Informationen siehe Thema Festlegen der Filteroptionen in Kapitel 4 auf S. 157. In diesem Thema finden Sie Tipps zum Einlesen von Daten in SPSS Statistics.

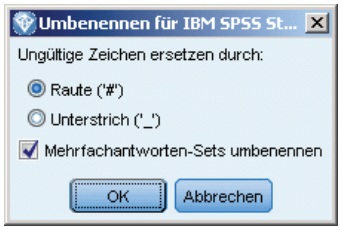
Abbildung 8-12
Umbenennen von Feldern für IBM SPSS Statistics auf der Registerkarte "Filter" im Statistikdateiknoten



Führen Sie folgende Schritte aus, um die Feldnamen an das SPSS Statistics-Namensschema anzupassen:

- Klicken Sie auf der Registerkarate "Filter" auf die Symbolleistenschaltfläche "Optionen im Filtermenü" (die erste in der Symbolleiste).
- Wählen Sie den Befehl Umbenennen für SPSS Statistics.

Abbildung 8-13
Umbenennen von Feldern



- Im Dialogfeld Umbenennen für SPSS Statistics können Sie auswählen, ob ungültige Zeichen in Dateinamen durch ein Rautenzeichen (#) oder einen Unterstrich (_) ersetzt werden.

  Umbenennen von Mehrfachantworten-Sets. Wählen Sie diese Option aus, wenn Sie die Namen von Mehrfachantworten-Sets bearbeiten wollen, die mithilfe eines Statistikdatei-Quellenknoten in SPSS Modeler importiert werden können. Sie werden zum Aufzeichnen von Daten verwendet, die mehr als einen Wert für jeden Fall haben, wie beispielsweise bei Umfrageantworten.

Kapitel
9

Superknoten

Überblick über Superknoten

Einer der Gründe, warum der Umgang mit der visuellen Programmierschnittstelle von IBM® SPSS® Modeler so leicht zu erlernen ist, liegt darin, dass jeder Knoten eine klar definierte Funktion erfüllt. Für eine komplexe Verarbeitung ist jedoch eventuell eine lange Sequenz von Knoten erforderlich. Dadurch können die Elemente im Stream-Zeichenbereich unübersichtlich werden und es kann schwierig werden, den Stream-Diagrammen zu folgen. Es gibt zwei Methoden zur Vermeidung eines langen und komplexen Streams:

- Sie können eine Verarbeitungssequenz in mehrere Streams aufteilen, die einander als Datengrundlage dienen. So könnte der erste Stream beispielsweise eine Datendatei erstellen, die der zweite Stream als Eingabe verwendet. Der zweite erstellt eine Datei, die der dritte Stream als Eingabe verwendet, usw. Diese Streams können Sie verwalten, indem Sie sie in einem **Projekt** speichern. Ein Projekt kann mehrere Streams und deren Ausgaben organisieren. Projektdateien enthalten jedoch nur einen Verweis auf die Objekte, die sie enthalten, und Sie müssen noch immer mehrere Stream-Dateien verwalten.
- Sie können einen **Superknoten** als als effizientere Alternative bei der Arbeit mit komplexen Stream-Prozessen erstellen.

Superknoten fassen mehrere Knoten zu einem einzigen zusammen, indem sie Bereiche eines Daten-Streams verkapseln. Dies bietet zahlreiche Vorteile für das Data Mining:

- Streams sind überschaubarer und können besser verwaltet werden.
- Knoten können zu einem geschäftsspezifischen Superknoten zusammengefasst werden.
- Superknoten können in Bibliotheken exportiert und in mehreren Data Mining-Projekten wieder verwendet werden.

Typen von Superknoten

Superknoten werden im Daten-Stream durch ein sternförmiges Symbol angezeigt. Das Symbol ist schattiert, um den Superknotentyp und die Richtung anzugeben, in der der Stream zum Superknoten hin bzw. von ihm weg fließen muss.

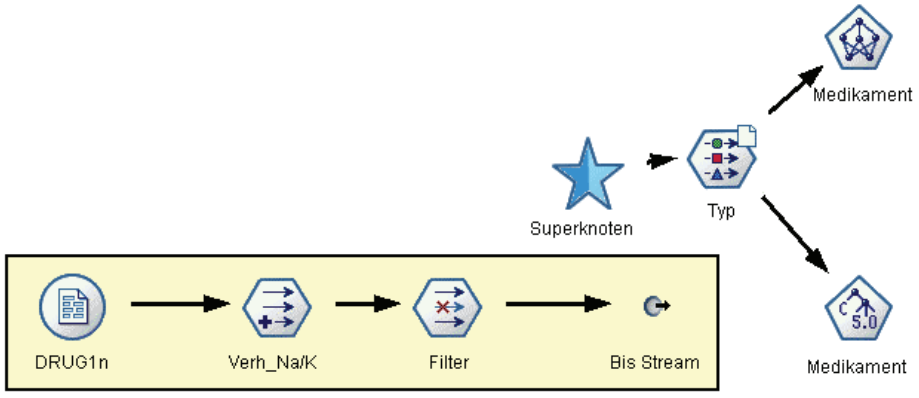
Es gibt drei Typen von Superknoten:

- Quellen-Superknoten
- Prozess-Superknoten
- End-Superknoten

Quellen-Superknoten

Quellen-Superknoten enthalten eine Datenquelle, genau wie ein normaler Quellenknoten, und können an jeder Stelle verwendet werden, an der auch ein normaler Quellenknoten eingesetzt werden kann. Die linke Seite eines Quellen-Superknotens ist schattiert, um anzuzeigen, dass er auf der linken Seite "geschlossen" ist und dass die Daten *vom* Superknoten nach unten im Stream fließen müssen.

Abbildung 9-1
Quellen-Superknoten mit vergrößerter Version über dem Stream

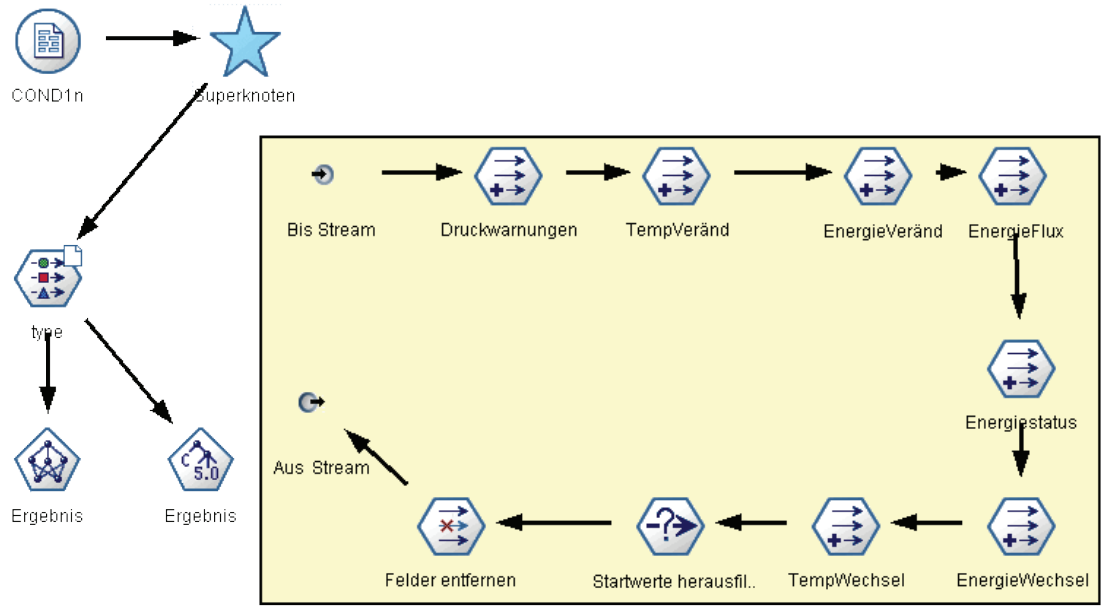


Quellen-Superknoten weisen nur einen einzigen Verbindungspunkt auf der rechten Seite auf, der anzeigt, dass die Daten den Superknoten verlassen und nach unten im Stream fließen.

Prozess-Superknoten

Prozess-Superknoten enthalten nur Prozessknoten und sind nicht schattiert, um anzuzeigen, dass die Daten bei diesem Superknotentyp sowohl in den Knoten *hinein*- als auch aus ihm *heraus*fließen können.

Abbildung 9-2
Process-Superknoten mit vergrößerter Version über dem Stream



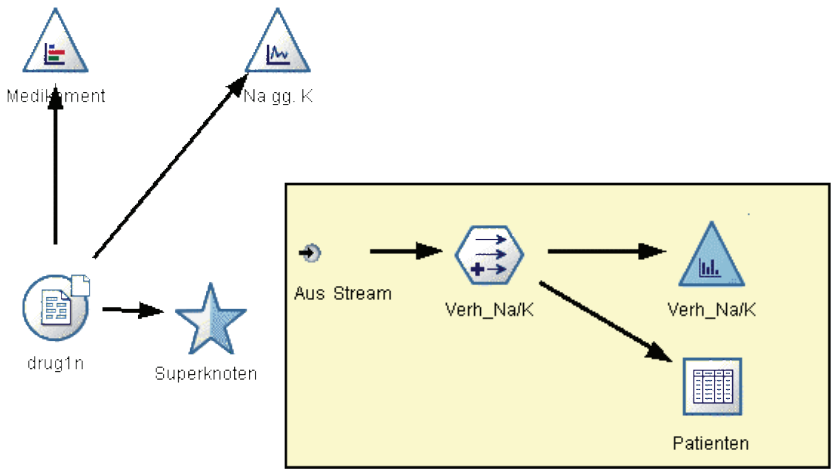
Prozess-Superknoten weisen sowohl links als auch rechts Verbindungspunkte auf, was anzeigt, dass die Daten in den Superknoten eintreten und ihn dann wieder verlassen und wieder in den Stream eintreten. Superknoten können zwar zusätzliche Stream-Fragmente und sogar zusätzliche Streams enthalten, beide Verbindungspunkte müssen aber durch einen einzelnen Pfad fließen, der die Punkte *Aus Stream* und *Bis Stream* verbindet.

Hinweis: Prozess-Superknoten werden manchmal als *Manipulations-Superknoten* bezeichnet.

End-Superknoten

End-Superknoten enthalten mindestens einen Endknoten (Diagramm, Tabelle usw.) und können auf dieselbe Weise verwendet werden wie Endknoten. Die rechte Seite eines Quellen-Superknotens ist schattiert, um anzuzeigen, dass er auf der rechten Seite “geschlossen” ist und dass die Daten nur *in* den End-Superknoten fließen können.

Abbildung 9-3
End-Superknoten mit vergrößerter Version über dem Stream



Quellen-Superknoten weisen nur einen einzigen Verbindungspunkt auf der rechten Seite auf, der anzeigt, dass die Daten aus dem Stream in den Superknoten eintreten und dort enden.

End-Superknoten können auch Skripts enthalten, die für alle Knoten innerhalb des Superknotens die Reihenfolge der Ausführung festlegen. Für weitere Informationen siehe Thema Superknoten und Skripts auf S. 536.

Erstellen von Superknoten

Beim Erstellen von Superknoten wird der Daten-Stream reduziert, indem mehrere Knoten zu einem Knoten verkapselt werden. Nach dem Erstellen bzw. Laden eines Streams im Zeichenbereich gibt es mehrere Möglichkeiten zum Erstellen eines Superknotens.

Mehrfachauswahl

Die einfachste Methode zum Erstellen eines Superknotens besteht in der Auswahl aller Knoten, die verkapselt werden sollen:

- Mithilfe der Maus können Sie mehrere Knoten im Stream-Zeichenbereich auswählen. Außerdem können Sie mit Umschalt-Klicken einen Stream oder einen Abschnitt eines Streams auswählen. *Hinweis*: Die ausgewählten Knoten müssen aus einem kontinuierlichen oder gegabelten Stream stammen. Knoten, die nicht aneinander angrenzen oder in irgendeiner Weise verbunden sind, können nicht ausgewählt werden.

- Anschließend verkapseln Sie die ausgewählten Knoten unter Verwendung einer der folgenden drei Methoden:
  - Klicken Sie auf das Superknoten-Symbol (sternförmig) in der Symbolleiste.

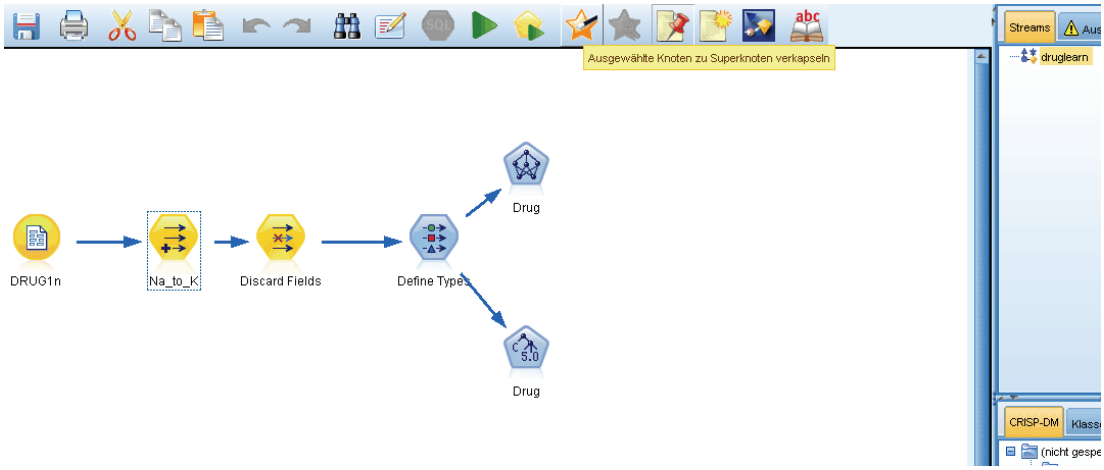
- Klicken Sie mit der rechten Maustaste auf den Superknoten und wählen Sie aus dem Kontextmenü folgende Optionen:

Superknoten erstellen > Aus Auswahl

- Wählen Sie im Superknoten-Menü folgende Befehlsfolge aus:

Superknoten erstellen > Aus Auswahl

Abbildung 9-4
Erstellen eines Superknotens mithilfe der Mehrfachauswahl



Bei allen drei Optionen werden die Knoten in einem Superknoten verkapselt, dessen Typ (Quellen-, Prozess- oder End-Superknoten) durch die Schattierung angezeigt wird. Die Grundlage dafür bildet der jeweilige Inhalt.

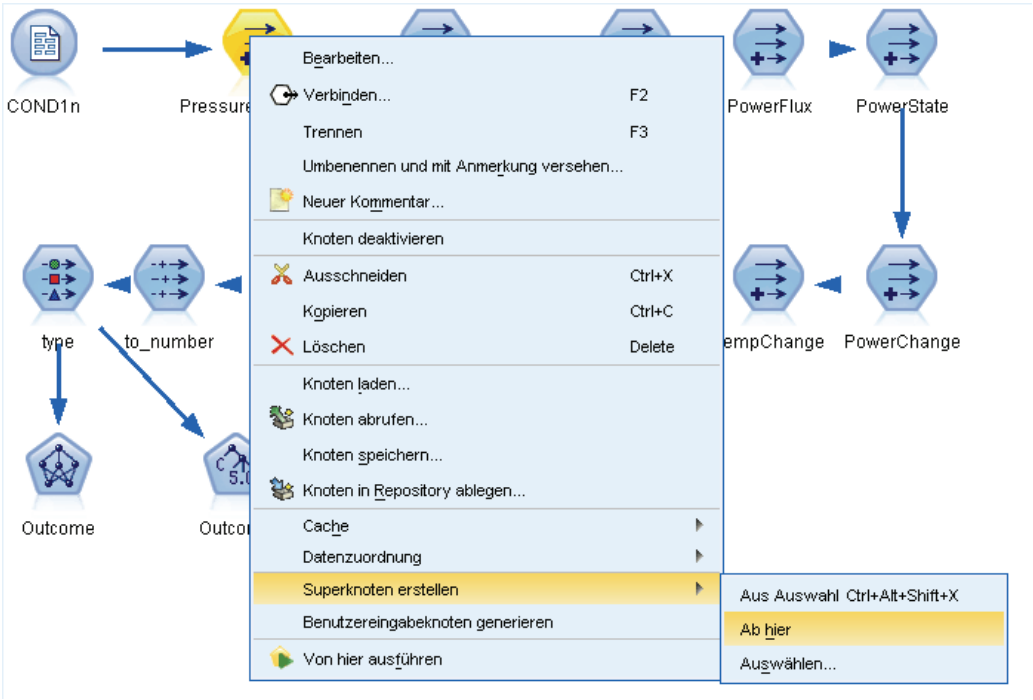
Einzelauswahl

Außerdem können Sie einen Superknoten erstellen, indem Sie einen einzelnen Knoten auswählen und mithilfe von Menüoptionen den Start und das Ende des Superknotens festlegen oder alle Elemente verkapseln, die im Stream hinter dem ausgewählten Knoten liegen.

- Klicken Sie auf den Knoten, der den Start der Verkapselung bestimmt.
- Wählen Sie im Superknoten-Menü folgende Befehlsfolge aus:

Superknoten erstellen > Ab hier

Abbildung 9-5
Erstellen eines Superknotens mithilfe des Kontextmenüs für die Einzelauswahl



Superknoten können außerdem auf mehr interaktive Weise erstellt werden. Dazu wählen Sie den Start und das Ende des Stream-Abschnitts aus, um die Knoten zu verkapseln:

- Klicken Sie auf den ersten oder letzten Knoten, der in den Superknoten aufgenommen werden soll.
- Wählen Sie im Superknoten-Menü folgende Befehlsfolge aus:
  Superknoten erstellen > Auswählen...
- Alternativ können Sie die Optionen des Kontextmenüs verwenden. Klicken Sie dazu mit der rechten Maustaste auf den gewünschten Knoten.
- Der Cursor wird zu einem Superknoten-Symbol, wodurch angezeigt wird, dass ein weiterer Punkt im Stream ausgewählt werden muss. Gehen Sie entweder nach unten oder nach oben im Stream zum “anderen Ende” des Superknoten-Fragements und klicken Sie auf einen Knoten. Dadurch werden alle dazwischenliegenden Knoten durch das Sternsymbol des Superknotens ersetzt.

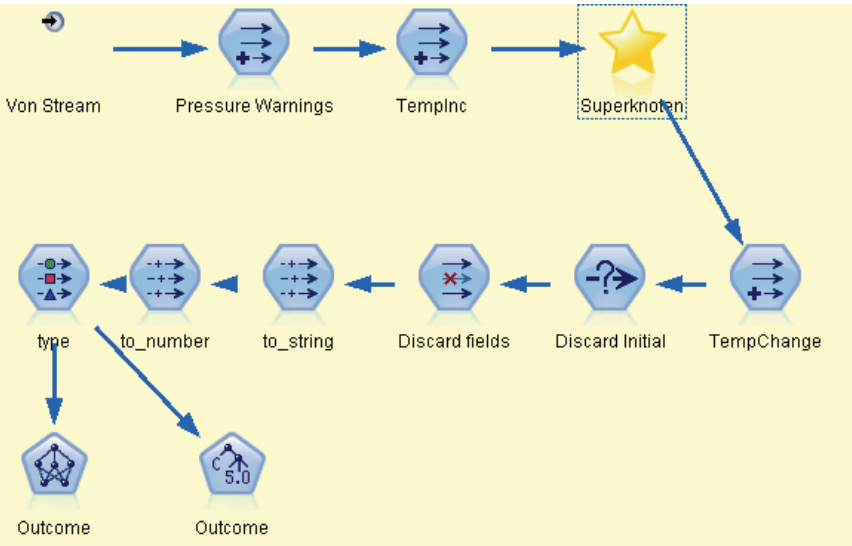
  Hinweis: Die ausgewählten Knoten müssen aus einem kontinuierlichen oder gegabelten Stream stammen. Knoten, die nicht aneinander angrenzen oder in irgendeiner Weise verbunden sind, können nicht ausgewählt werden.

Verschachteln von Superknoten

Superknoten können innerhalb von anderen Superknoten verschachtelt werden. Die Regeln für die einzelnen Superknotentypen (Quellen-, Prozess- und End-Superknoten) gelten uneingeschränkt auch für verschachtelte Superknoten. Bei einem Prozess-Superknoten mit Verschachtelung muss

ein kontinuierlicher Datenfluss durch alle verschachtelten Superknoten vorliegen, damit er ein Prozess-Superknoten bleiben kann. Wenn es sich bei einem der verschachtelten Superknoten um einen End-Superknoten handelt, fließen die Daten nicht mehr durch die Hierarchie.

Abbildung 9-6
In einem anderen Prozess-Superknoten verschachtelter Prozess-Superknoten



End- und Quellen-Superknoten können andere Typen verschachtelter Superknoten enthalten, doch die grundlegenden Regeln für das Erstellen von Superknoten gelten weiterhin.

Beispiele für gültige Superknoten

Fast alle in IBM® SPSS® Modeler erstellten Elemente können in einem Superknoten verkapselt werden. Im Folgenden finden Sie Beispiele für gültige Superknoten:

Abbildung 9-7
Gültiger Prozess-Superknoten mit zwei Verbindungen in einem gültigen Stream-Fluss

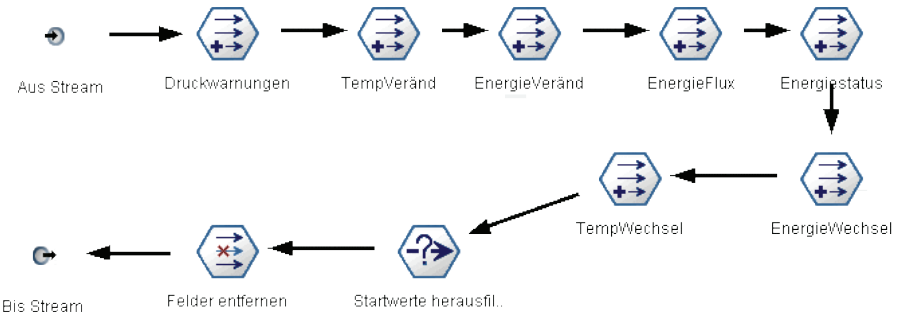


Abbildung 9-8
Gültiger End-Superknoten mit einem separatem Stream, der zum Testen der generierten Modelle verwendet wird

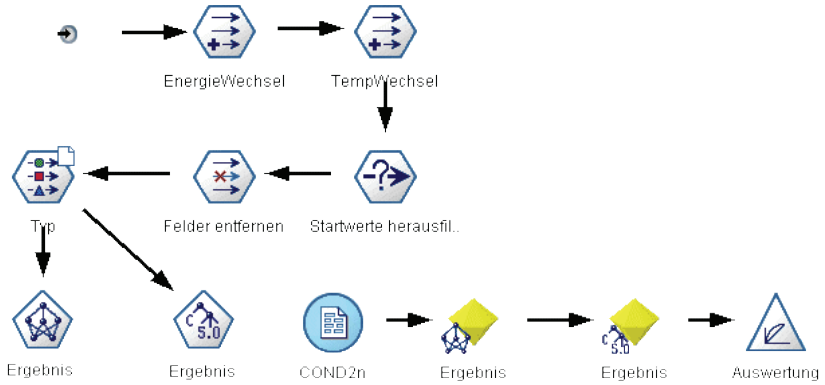
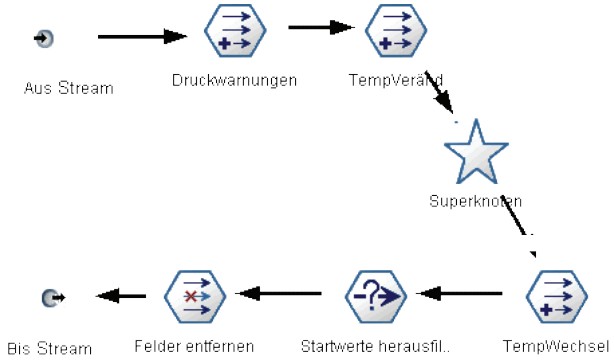


Abbildung 9-9
Gültiger Prozess-Superknoten mit einem verschachtelten Superknoten



Beispiele für ungültige Superknoten

Der wichtigste Aspekt beim Erstellen gültiger Superknoten besteht darin, sicherzustellen, dass die Daten linear durch die Superknoten-Verbindungen fließen. Wenn zwei Verbindungen bestehen (Prozess-Superknoten), müssen die Daten in einem Stream von der Anfangsverbindung zur Endverbindung fließen. In ähnlicher Weise muss ein Quellen-Superknoten zulassen, dass Daten vom Quellenknoten zu der einzelnen Verbindung fließen, die die Daten an den nicht vergrößerten Stream übergibt.

Abbildung 9-10
Ungültiger Quellen-Superknoten: Quellenknoten nicht mit Datenflusspfad verbunden

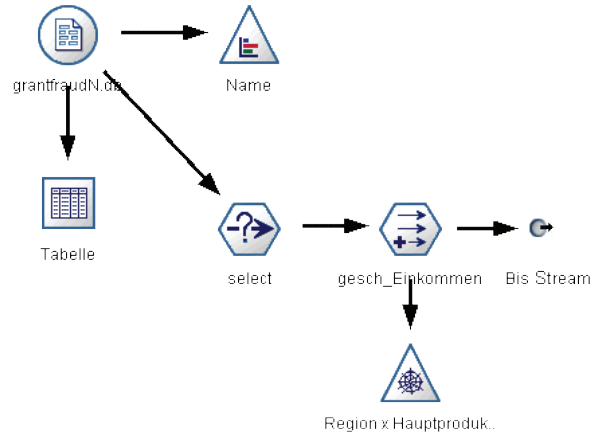
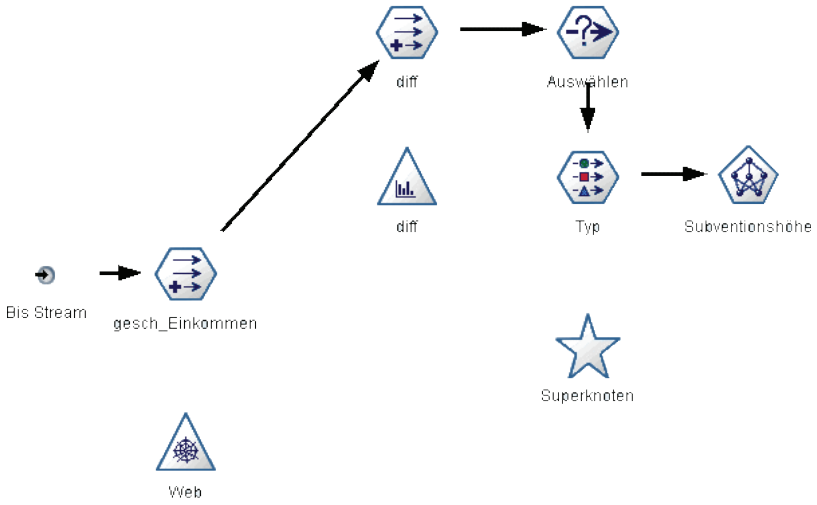


Abbildung 9-11
Ungültiger End-Superknoten: Verschachtelter Superknoten nicht mit Datenflusspfad verbunden



Sperren von Superknoten

Nach dem Erstellen eines Superknotens können Sie ihn mit einem Passwort schützen, um eine Änderung zu verhindern. Dies ist z. B. möglich, wenn Sie Streams oder Teile von Streams als Vorlagen mit festem Wert für andere Benutzer in Ihrer Organisation erstellen, die weniger erfahren mit dem Einrichten von IBM® SPSS® Modeler-Abfragen sind.

Für einen gesperrten Superknoten können Benutzer in der Registerkarte “Parameter” immer noch Werte für Parameter eingeben, die definiert wurden. Außerdem kann ein gesperrter Superknoten ohne Eingabe des Passworts ausgeführt werden.

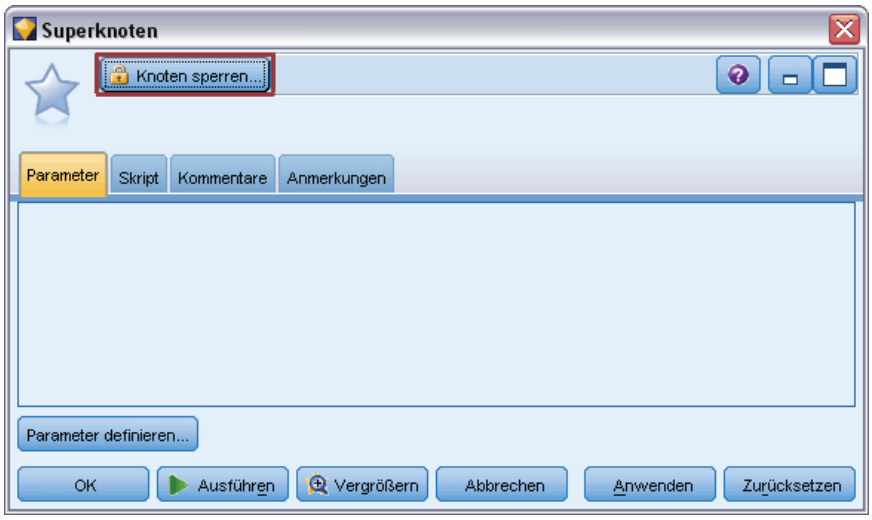
Hinweis: Das Sperren und Entsperren mithilfe von Skripts ist nicht möglich.

Sperren und Entsperren eines Superknotens

Achtung: Vergessene Passwörter können nicht wiederhergestellt werden.

Sie können einen Superknoten von einer der drei Registerkarten aus sperren oder entsperren.

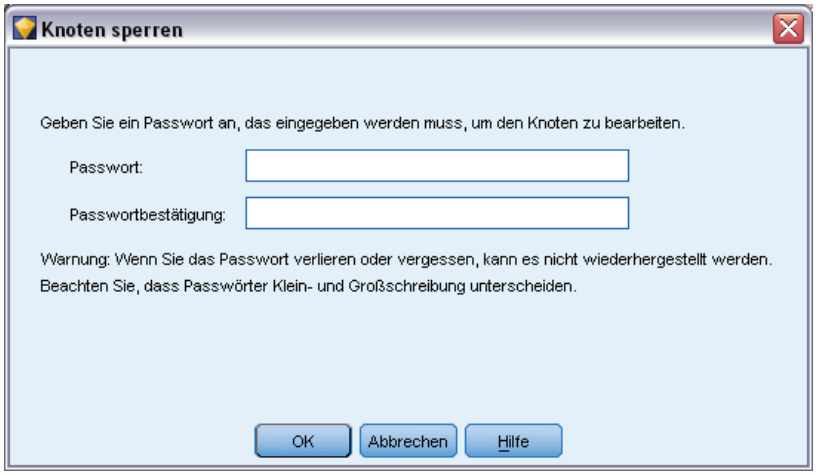
Abbildung 9-12
Sperren eines Superknotens



Klicken Sie auf Knoten sperren.

Geben Sie das Passwort ein und bestätigen Sie es.

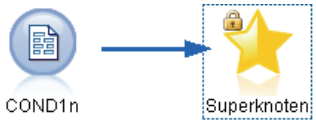
Abbildung 9-13
Superknoten-Passwort eingeben und bestätigen



▶ Klicken Sie auf OK.

Ein passwortgeschützter Superknoten wird im Stream-Zeichenbereich durch ein kleines Vorhängeschlosssymbol in der oberen linken Ecke des Superknotensymbols markiert.

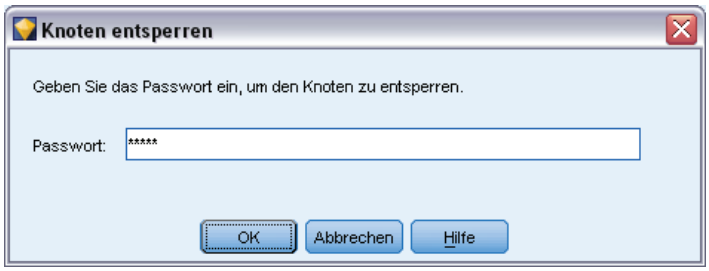
Abbildung 9-14
Gesperrter Quell-Superknoten als Teil eines Streams



Entsperren eines Superknotens

▶ Sie entfernen den Passwortschutz permanent, indem Sie auf Knoten entsperren klicken. Sie werden dann zur Eingabe des Passworts aufgefordert.

Abbildung 9-15
Passwort eingeben, um einen Superknoten zu entsperren

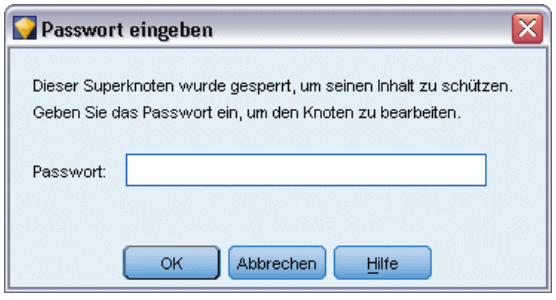


▶ Geben Sie das Passwort ein und klicken Sie auf OK. Der Superknoten ist nun nicht mehr passwortgeschützt und an seinem Symbol im Stream wird kein Vorhängeschloss-Symbol mehr angezeigt.

Bearbeiten eines gesperrten Superknotens

Wenn Sie versuchen, Parameter für einen gesperrten Superknoten zu definieren oder zu zoomen, um einen gesperrten Superknoten anzuzeigen, werden Sie aufgefordert, das Passwort einzugeben.

Abbildung 9-16
Passwort eingeben für Zoom oder zum Bearbeiten eines Superknotens



▶ Geben Sie das Passwort ein und klicken Sie auf OK.

Sie können nun die Parameterdefinitionen bearbeiten und wie gewünscht zoomen, bis Sie den Stream schließen, in dem sich der Superknoten befindet.

Beachten Sie, dass damit nicht der Passwortschutz entfernt wird, sondern Ihnen nur ermöglicht wird, mit dem Superknoten zu arbeiten. Für weitere Informationen siehe Thema Sperren und Entsperren eines Superknotens auf S. 527.

Bearbeiten von Superknoten

Nach dem Erstellen eines Superknotens können Sie ihn genauer untersuchen, indem Sie ihn vergrößern. Wenn der Superknoten gesperrt ist, werden Sie zur Eingabe des Passworts aufgefordert. Für weitere Informationen siehe Thema Bearbeiten eines gesperrten Superknotens auf S. 528.

Wenn Sie den Inhalt eines Superknotens anzeigen möchten, können Sie dazu entweder das Vergrößerungssymbol der IBM® SPSS® Modeler-Symbolleiste oder die folgende Methode verwenden:

- Klicken Sie mit der rechten Maustaste auf einen Superknoten.
- Wählen Sie im Kontextmenü die Option Vergrößern aus.

Der Inhalt des ausgewählten Superknotens wird in einer leicht abweichenden SPSS Modeler-Umgebung angezeigt, in der die Verbindungen den Fluss der Daten durch den Stream bzw. das Stream-Fragment anzeigen. Auf dieser Ebene im Stream-Zeichenbereich können Sie mehrere Aufgaben durchführen:

- Ändern des Superknotentyps – Quellen-, Prozess- oder End-Superknoten.
- Erstellen von Parametern bzw. Bearbeiten der Werte eines Parameters. Parameter werden zur Skripterstellung und für CLEM-Ausdrücke verwendet.
- Festlegen von Caching-Optionen für den Superknoten und seine Unterknoten.
- Erstellen oder Bearbeiten eines Superknoten-Skripts (nur bei End-Superknoten).

Ändern der Superknotentypen

Unter gewissen Umständen kann es sinnvoll sein, den Typ eines Superknotens zu ändern. Diese Funktion ist nur verfügbar, wenn Sie die Ansicht des Superknotens vergrößert haben, und sie bezieht sich nur auf den Superknoten auf dieser Stufe. Es gibt folgende drei Typen von Superknoten:

Quellen-Superknoten	Eine ausgehende Verbindung
Prozess-Superknoten	Zwei Verbindungen: eine eingehende und eine ausgehende
End-Superknoten	Eine eingehende Verbindung

So ändern Sie den Typ eines Superknotens:

- Stellen Sie sicher, dass sie die Ansicht des Superknotens vergrößert haben.

- Wählen Sie im Superknotenmenü die Option Superknotentyp und wählen Sie anschließend den Typ aus.

Anmerkungen für Superknoten und Umbenennen von Superknoten

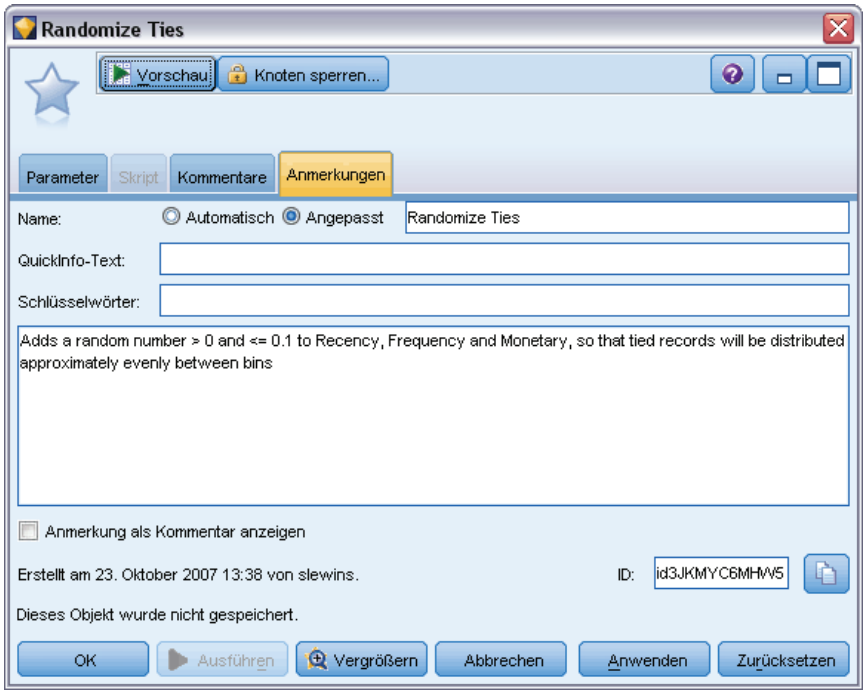
Sie können einen Superknoten umbenennen, wenn er im Stream angezeigt wird, sowie Anmerkungen schreiben, die in einem Projekt oder Bericht verwendet werden. So können Sie auf diese Eigenschaften zugreifen:

- Klicken Sie mit der rechten Maustaste auf einen Superknoten (verkleinert) und wählen Sie die Option Umbenennen und mit Anmerkung versehen.

- Alternativ wählen Sie im Superknoten-Menü die Option Umbenennen und mit Anmerkung versehen. Diese Option ist sowohl im vergrößerten als auch im verkleinerten Modus verfügbar.

In beiden Fällen wird ein Dialogfeld geöffnet, bei dem die Registerkarte "Anmerkungen" ausgewählt ist. Mit den hier verfügbaren Optionen können Sie den im Stream-Zeichenbereich angezeigten Namen anpassen und eine Dokumentation zu den Superknoten-Operationen bereitstellen.

Abbildung 9-17
Erstellen von Anmerkungen für einen Superknoten



Verwenden von Kommentaren mit Superknoten

Wenn Sie aus einem kommentierten Knoten oder Nugget einen Superknoten erstellen, müssen Sie den Kommentar in die Auswahl aufnehmen, falls dieser im Superknoten erscheinen soll. Wenn Sie den Kommentar aus der Auswahl weglassen, bleibt der Kommentar ohne Verbindung im Stream, nachdem der Superknoten erstellt wurde.

Wenn Sie einen Superknoten erweitern, der Kommentare enthielt, werden die Kommentare wieder an ihrer ursprünglichen Position (vor der Erstellung des Superknotens) eingesetzt.

Wenn Sie einen Superknoten erweitern, der kommentierte Objekte enthielt, aber die Kommentare nicht in den Superknoten aufgenommen wurden, werden die Objekte wieder an ihrer ursprünglichen Position eingesetzt, aber die Kommentare werden nicht erneut verknüpft.

Superknoten-Parameter

In IBM® SPSS® Modeler haben Sie die Möglichkeit, benutzerdefinierte Variablen festzulegen, beispielsweise *Minvalue*, deren Werte bei der Verwendung in der Skripterstellung oder in CLEM-Ausdrücken angegeben werden können. Diese Variablen heißen **Parameter**. Sie können Parameter für Streams, Sitzungen und Superknoten festlegen. Alle für einen Superknoten festgelegten Parameter sind bei der Erstellung von CLEM-Ausdrücken in diesem Superknoten oder etwaigen verschachtelten Knoten verfügbar. Die für verschachtelte Superknoten festgelegten Parameter stehen nicht für den übergeordneten Superknoten zur Verfügung.

Es gibt zwei Schritte zum Erstellen und Festlegen von Parametern für Superknoten:

- Definieren Sie die Parameter für den Superknoten.
- Geben Sie anschließend den Wert für die einzelnen Parameter des Superknotens an.

Diese Parameter können dann in CLEM-Ausdrücken für alle verschachtelten Knoten verwendet werden.

Festlegen von Superknoten-Parametern

Parameter für einen Superknoten können sowohl im vergrößerten als auch im verkleinerten Modus definiert werden. Die definierten Parameter gelten für alle verkapselten Knoten. Zur Definition der Parameter eines Superknotens müssen Sie zunächst im Dialogfeld des Superknotens die Registerkarte “Parameter” aufrufen. Das Dialogfeld lässt sich auf folgende Weisen öffnen:

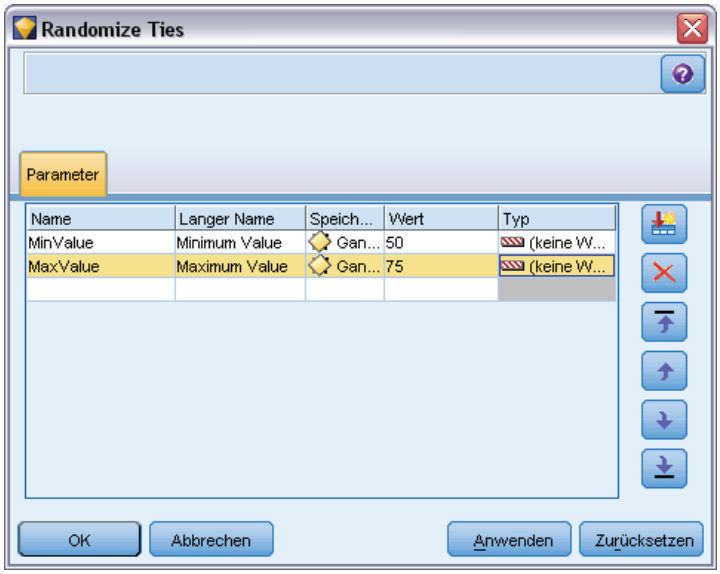
- Doppelklicken Sie auf einen Superknoten im Stream.
- Wählen Sie im Superknoten-Menü den Befehl Parameter festlegen aus.
- Alternativ wählen Sie bei vergrößerter Superknoten-Ansicht die Option Parameter festlegen aus dem Kontextmenü.

Nach dem Öffnen des Dialogfelds wird die Registerkarte “Parameter” mit allen zuvor definierten Parametern angezeigt.

So definieren Sie einen neuen Parameter:

▶ Klicken Sie auf die Schaltfläche Parameter definieren, um das Dialogfeld zu öffnen.

Abbildung 9-18
Definieren der Parameter für einen Superknoten.



Name. Hier werden die Parameternamen aufgelistet. Sie können einen neuen Parameter erstellen, indem Sie in diesem Feld einen Namen eingeben. Um beispielsweise einen Parameter für die Mindesttemperatur zu erstellen, könnten Sie minvalue eingeben. Verwenden Sie nicht das Präfix $P-, das Parameter in CLEM-Ausdrücken kennzeichnet. Dieser Name wird auch zur Anzeige im CLEM Expression Builder verwendet.

Langer Name. Listet den beschreibenden Namen für die einzelnen erstellten Parameter auf.

Speichertyp. Wählen Sie einen Speichertyp aus der Liste aus. Der Speichertyp gibt an, wie die Datenwerte im Parameter gespeichert werden. Wenn Sie z. B. mit Werten arbeiten, die führende Nullen enthalten und die Sie beibehalten möchten (wie 008), sollten Sie Zeichenkette als Speichertyp wählen. Andernfalls werden die Nullen vom Wert abgezogen. Verfügbare Speichertypen sind "Zeichenkette", "Ganze Zahl", "Reelle Zahl", "Uhrzeit", "Datum" und "Zeitstempel". Beachten Sie, dass bei Datumsparametern die Werte gemäß der im nächsten Absatz erläuterten ISO-Standardnotierung eingegeben werden müssen.

Wert. Listet den aktuellen Wert für die einzelnen Parameter auf. Ändern Sie den Parameter wie gewünscht. Datumsparameter müssen in ISO-Standardnotierung angegeben werden (d. h. in der Form YYYY-MM-DD). Datumsangaben in anderen Formaten sind nicht zulässig.

Typ (optional). Wenn Sie den Stream für eine externe Anwendung bereitstellen möchten, wählen Sie aus der Liste ein Messniveau aus. Andernfalls sollten Sie die Spalte *Typ* so belassen, wie sie ist. Wenn Sie Wertebeschränkungen für den Parameter festlegen möchten, z. B. die Ober- und Untergrenze für einen numerischen Bereich, wählen Sie Angeben aus der Liste aus.

Die Optionen "Langer Name", "Speichertyp" und "Typ" können nur über die Benutzeroberfläche für Parameter festgelegt werden. Die Festlegung dieser Optionen mithilfe von Skripts ist nicht möglich.

Klicken Sie auf die Pfeile rechts, um den ausgewählten Parameter in der Liste verfügbarer Parameter weiter nach oben oder weiter nach unten zu verschieben. Verwenden Sie die Schaltfläche zum Löschen (mit einem *X* markiert), um den ausgewählten Parameter zu entfernen.

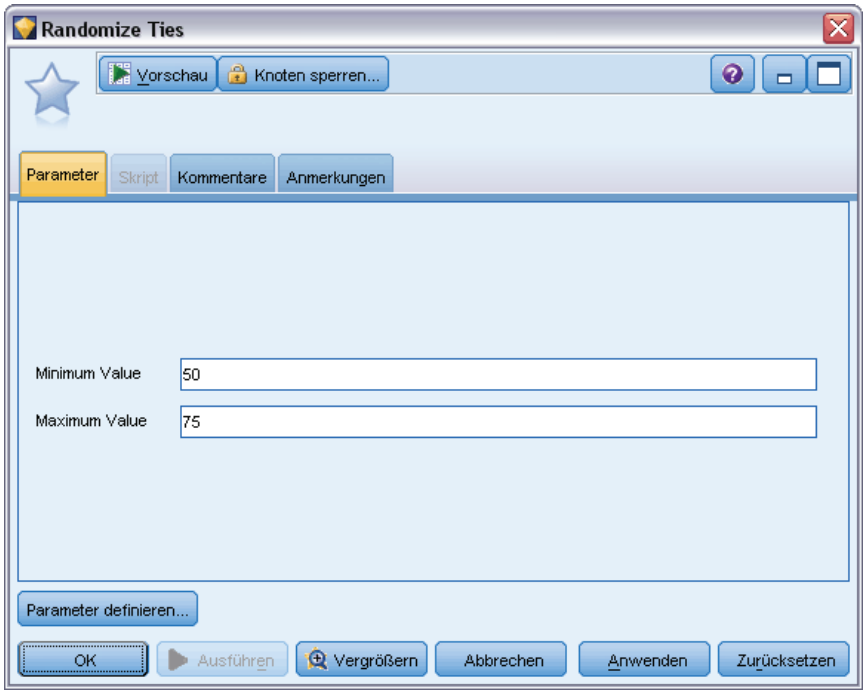
Festlegen von Werten für Superknoten-Parameter

Nach der Definition von Parametern für einen Superknoten können Sie mithilfe der Parameter in einem CLEM-Ausdruck bzw. einem -Skript Werte angeben.

So geben Sie die Parameter eines Superknotens an:

- Doppelklicken Sie auf das Superknoten-Symbol, um das Dialogfeld für den Superknoten zu öffnen.
- Alternativ können Sie im Superknoten-Menü den Befehl Parameter festlegen auswählen.
- Klicken Sie auf die Registerkarte Parameter. *Hinweis*: Bei den Feldern in diesem Dialogfeld handelt es sich um die Felder, die durch Klicken auf die Schaltfläche Parameter definieren auf dieser Registerkarte definiert wurden.
- Geben Sie für jeden erstellten Parameter einen Wert in das Textfeld ein. Beispielsweise können Sie den Wert *minvalue* auf einen bestimmten Schwellenwert festsetzen. Dieser Parameter kann anschließend in verschiedenen Operationen verwendet werden, beispielsweise bei der Auswahl der Datensätze oberhalb oder unterhalb dieses Schwellenwerts zur weiteren Untersuchung.

Abbildung 9-19
Angeben von Parametern für einen Superknoten.



Verwenden von Superknotenparametern zum Zugriff auf Knoteneigenschaften

Superknotenparameter können außerdem zur Definition von Knoteneigenschaften (auch als **Slot-Parameter** bezeichnet) für verkapselte Knoten verwendet werden. Beispiel: Angenommen, Sie möchten festlegen, dass ein Superknoten einen verkapselten Netzwerkknoten eine bestimmte Zeit lang mithilfe einer Zufallsstichprobe der verfügbaren Daten trainiert. Mithilfe von Parametern können Sie Werte für die Zeitdauer und und den Prozentsatz der Stichprobe angeben.

Abbildung 9-20
In einem Superknoten verkapseltes Stream-Fragment

Dieser Beispiel-Superknoten enthält einen Stichprobenknoten mit der Bezeichnung *Sample* (Stichprobe) und einen Netzwerkknoten mit der Bezeichnung *Train* (Trainieren). Mit den Knotendialogfeldern können Sie für die Einstellung **Stichprobe** des Stichprobenknotens Zufällig % und für die Einstellung **Stopp bei** des Netzwerkknotens Zeit festlegen. Nachdem diese Optionen angegeben wurden, können Sie auf die Knoteneigenschaften mit Parametern zugreifen und bestimmte Werte für den Superknoten angeben. Klicken Sie im Dialogfeld "Superknoten" auf Parameter definieren und erstellen Sie folgende Parameter:

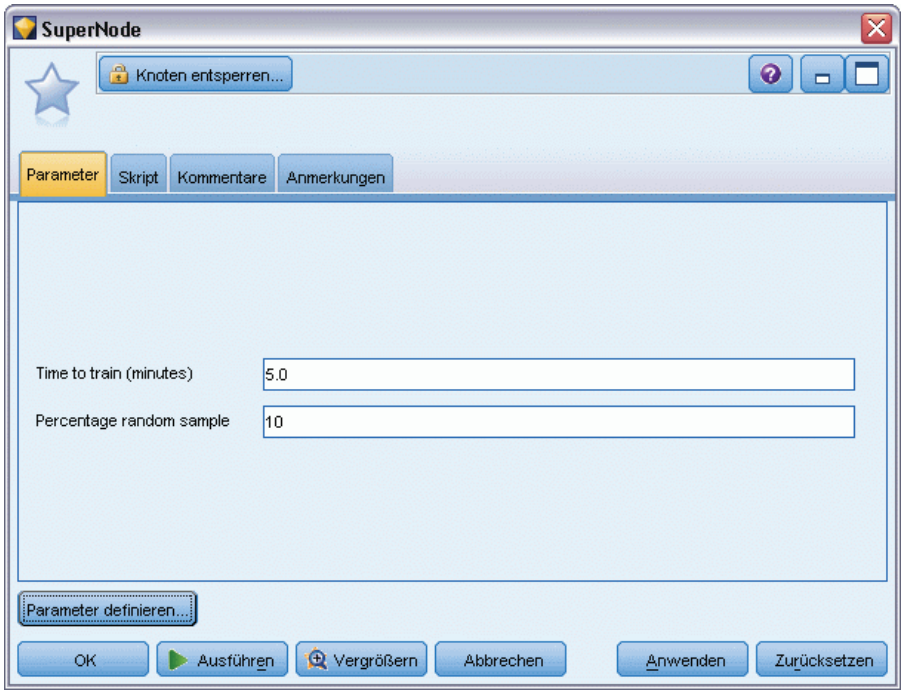
Abbildung 9-21
Definieren von Parametern für den Zugriff auf Knoteneigenschaften

Hinweis: Bei den Parameternamen, beispielsweise *Sample.random*, wird korrekte Syntax für Referenzen auf Knoteneigenschaften verwendet. Dabei steht *Sample* für den Namen des Knotens und *random* ist eine Knoteneigenschaft.

Nach der Definition dieser Parameter können Sie problemlos diese Werte für die beiden Stichproben- und Netzwerkknoten-Eigenschaften angeben, ohne dass die einzelnen Dialogfelder erneut geöffnet werden müssen. Wählen Sie stattdessen einfach im Superknoten-Menü die Option

Parameter festlegen aus, um über das Dialogfeld des Superknotens die Registerkarte “Parameter” aufzurufen, auf der Sie neue Werte für Zufällig % und Zeit angeben können. Dies ist besonders nützlich bei der Untersuchung der Daten während mehrerer Iterationen der Modellerstellung.

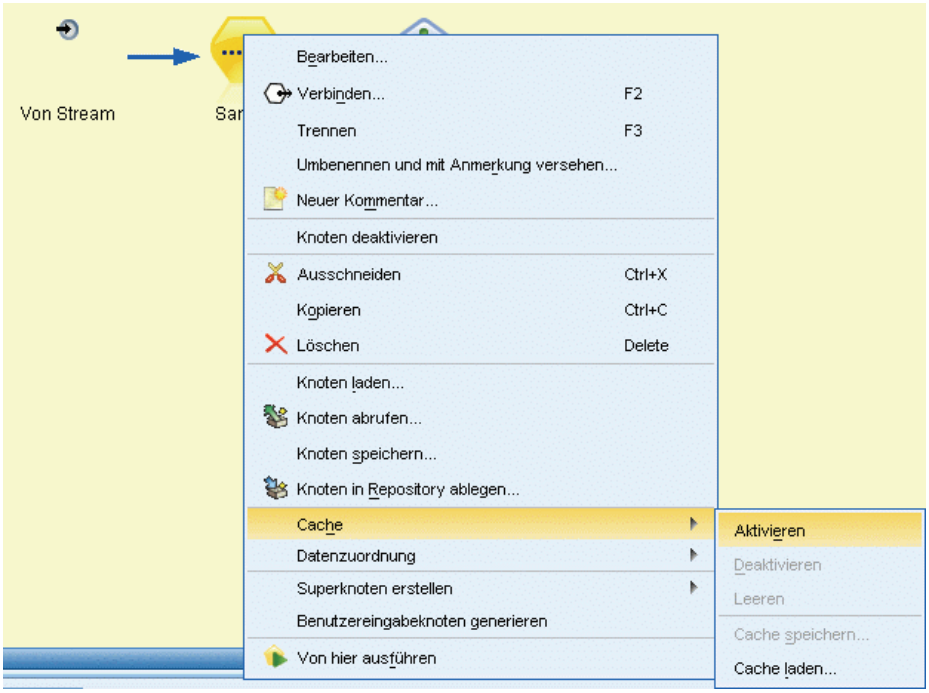
Abbildung 9-22
Angabe von Werten für Knoteneigenschaften auf der Registerkarte “Parameter” im Superknoten-Dialogfeld



Superknoten und Caching

Aus einem Superknoten können alle Knoten außer Endknoten im Cache gespeichert werden. Das Caching wird durch Rechtsklicken auf einen Knoten und Auswahl einer von mehreren Optionen aus dem Cache-Kontextmenü gesteuert. Diese Menüoption ist sowohl von außerhalb eines Superknotens als auch für die in einem Superknoten verkapselten Knoten verfügbar.

Abbildung 9-23
Auswählen der Caching-Optionen für einen Superknoten



Es gibt verschiedene Richtlinien für Superknoten-Caches:

- Wenn bei mindestens einem der in einem Superknoten verkapselten Knoten Caching aktiviert ist, ist es auch beim Superknoten aktiviert.
- Wenn der Cache für einen Superknoten deaktiviert wird, wird der Cache auch für *alle* verkapselten Knoten deaktiviert.
- Beim Aktivieren des Caching bei einem Superknoten wird der Cache tatsächlich für den letzten Caching-fähigen Superknoten aktiviert. Anders ausgedrückt: Wenn der letzte Superknoten ein Auswahlknoten ist, wird der Cache für diesen Auswahlknoten aktiviert. Wenn es sich beim letzten Unterknoten um einen Endknoten handelt (bei dem Caching nicht möglich ist), wird der nächste Knoten weiter oben im Stream, der Caching unterstützt, aktiviert.
- Nach der Festlegung von Caches für die Unterknoten eines Superknotens werden bei jeder Aktivität oberhalb des Knotens, für den die Cache-Speicherung erfolgte, wie beispielsweise Hinzufügen und Bearbeiten von Knoten, die Caches geleert.

Superknoten und Skripts

Sie können mithilfe der IBM® SPSS® Modeler-Skriptsprache einfache Programme schreiben, mit denen der Inhalt eines End-Superknotens bearbeitet und ausgeführt werden kann. Beispielsweise können Sie für komplexe Streams die Reihenfolge der Ausführung festlegen. Wenn ein Superknoten beispielsweise einen Globalwerteknoten enthält, der vor einem Plot-Knoten ausgeführt werden muss, können Sie ein Skript erstellen, mit dem zuerst der Globalwerteknoten ausgeführt wird. Die durch diesen Knoten berechneten Werte, wie Durchschnitt oder

Standardabweichung, können anschließend bei der Ausführung des Plotknotens verwendet werden.

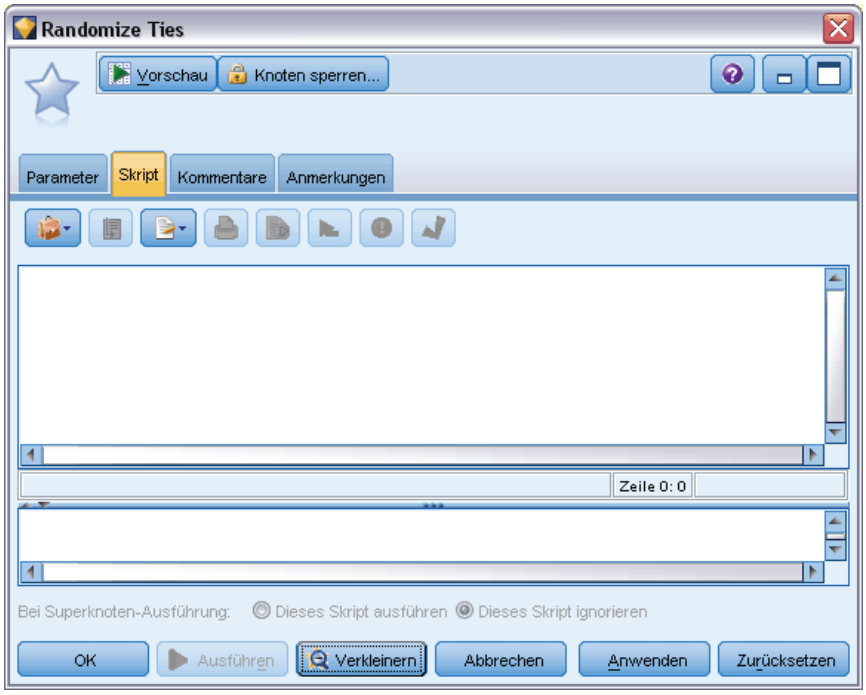
Die Registerkarte “Skript” des Dialogfelds “Superknoten” ist nur für End-Superknoten verfügbar.

So öffnen Sie das Skript-Dialogfeld für einen End-Superknoten:

- Klicken Sie mit der rechten Maustaste auf den Zeichenbereich des Superknotens und wählen Sie die Option Superknoten-Skript aus:
- Alternativ können Sie sowohl im vergrößerten als auch im verkleinerten Modus im Superknoten-Menü die Option Superknoten-Skript auswählen.

Hinweis: Superknoten-Skripts werden nur mit dem Stream und dem Superknoten ausgeführt, wenn im Dialogfeld Dieses Skript ausführen ausgewählt wurde.

Abbildung 9-24
Erstellen eines Skripts für einen Superknoten



Spezifische Optionen für die Skripts und ihre Verwendung in SPSS Modeler finden Sie im *Handbuch für Skripterstellung und Automatisierung* auf der SPSS ModelerDVD.

Speichern und Laden von Superknoten

Einer der Vorteile von Superknoten besteht darin, dass sie gespeichert und in anderen Streams wieder verwendet werden können. Beim Speichern und Laden von Superknoten werden *.slb*-Erweiterungen verwendet.

So speichern Sie einen Superknoten:

- Vergrößern Sie den Superknoten.
- Wählen Sie im Superknoten-Menü den Befehl Superknoten speichern aus.
- Geben Sie im Dialogfeld einen Dateinamen und ein Verzeichnis an.
- Wählen Sie aus, ob der gespeicherte Superknoten zum aktuellen Projekt hinzugefügt werden soll.
- Klicken Sie auf Speichern.

So laden Sie einen Superknoten:

- Wählen Sie im Menü "Einfügen" im IBM® SPSS® Modeler-Fenster die Option Superknoten aus.
- Wählen Sie eine Superknoten-Datei (*.slb*) aus dem aktuellen Verzeichnis aus oder wechseln Sie zu einem anderen Verzeichnis.
- Klicken Sie auf Laden.

Hinweis: Bei importierten Superknoten werden für alle Parameter Standardwerte verwendet. Zum Ändern der Parameter doppelklicken Sie auf einen Superknoten im Stream-Zeichenbereich.

Anhang

A

Hinweise

Diese Informationen wurden für weltweit angebotene Produkte und Dienstleistungen erarbeitet.

IBM bietet die in diesem Dokument behandelten Produkte, Dienstleistungen oder Merkmale möglicherweise nicht in anderen Ländern an. Informationen zu den derzeit in Ihrem Land erhältlichen Produkten und Dienstleistungen erhalten Sie bei Ihrem zuständigen IBM-Mitarbeiter vor Ort. Mit etwaigen Verweisen auf Produkte, Programme oder Dienste von IBM soll nicht behauptet oder impliziert werden, dass nur das betreffende Produkt oder Programm bzw. der betreffende Dienst von IBM verwendet werden kann. Stattdessen können alle funktional gleichwertigen Produkte, Programme oder Dienste verwendet werden, die keine geistigen Eigentumsrechte von IBM verletzen. Es obliegt jedoch der Verantwortung des Benutzers, die Funktionsweise von Produkten, Programmen oder Diensten von Drittanbietern zu bewerten und zu überprüfen.

IBM verfügt möglicherweise über Patente oder hat Patentanträge gestellt, die sich auf in diesem Dokument beschriebene Inhalte beziehen. Durch die Bereitstellung dieses Dokuments werden Ihnen keinerlei Lizenzen an diesen Patenten gewährt. Lizenzanfragen können schriftlich an folgende Adresse gesendet werden:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

Bei Lizenzanfragen in Bezug auf DBCS-Daten (Double-Byte Character Set) wenden Sie sich an die für geistiges Eigentum zuständige Abteilung von IBM in Ihrem Land. Schriftliche Anfragen können Sie auch an folgende Adresse senden:

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

Der folgende Abschnitt findet in Großbritannien und anderen Ländern keine Anwendung, in denen solche Bestimmungen nicht mit der örtlichen Gesetzgebung vereinbar sind: INTERNATIONAL BUSINESS MACHINES STELLT DIESE VERÖFFENTLICHUNG IN DER VERFÜGBAREN FORM OHNE GARANTIEN BEREIT, SEIEN ES AUSDRÜCKLICHE ODER STILLSCHWEIGENDE, EINSCHLIESSLICH JEDOCH NICHT NUR DER GARANTIEN BEZÜGLICH DER NICHT-RECHTSVERLETZUNG, DER GÜTE UND DER EIGNUNG FÜR EINEN BESTIMMTEN ZWECK. Manche Rechtsprechungen lassen den Ausschluss ausdrücklicher oder implizierter Garantien bei bestimmten Transaktionen nicht zu, sodass die oben genannte Ausschlussklausel möglicherweise nicht für Sie relevant ist.

Diese Informationen können technische Ungenauigkeiten oder typografische Fehler aufweisen. An den hierin enthaltenen Informationen werden regelmäßig Änderungen vorgenommen. Diese Änderungen werden in neuen Ausgaben der Veröffentlichung aufgenommen. IBM kann jederzeit und ohne vorherige Ankündigung Optimierungen und/oder Änderungen an den Produkten und/oder Programmen vornehmen, die in dieser Veröffentlichung beschrieben werden.

Jegliche Verweise auf Drittanbieter-Websites in dieser Information werden nur der Vollständigkeit halber bereitgestellt und dienen nicht als Befürwortung dieser. Das Material auf diesen Websites ist kein Bestandteil des Materials zu diesem IBM-Produkt und die Verwendung erfolgt auf eigene Gefahr.

IBM kann die von Ihnen angegebenen Informationen verwenden oder weitergeben, wie dies angemessen erscheint, ohne Ihnen gegenüber eine Verpflichtung einzugehen.

Lizenznehmer dieses Programms, die Informationen dazu benötigen, wie (i) der Austausch von Informationen zwischen unabhängig erstellten Programmen und anderen Programmen und (ii) die gegenseitige Verwendung dieser ausgetauschten Informationen ermöglicht wird, wenden sich an:

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

Derartige Informationen stehen ggf. in Abhängigkeit von den jeweiligen Geschäftsbedingungen sowie in einigen Fällen der Zahlung einer Gebühr zur Verfügung.

Das in diesem Dokument beschriebene lizenzierte Programm und sämtliche dafür verfügbaren lizenzierten Materialien werden von IBM gemäß dem IBM-Kundenvertrag, den Internationalen Nutzungsbedingungen für Programmpakete der IBM oder einer anderen zwischen uns getroffenen Vereinbarung bereitgestellt.

Jegliche hier enthaltene Daten zur Leistung wurden in einer überwachten Umgebung ermittelt. Aus diesem Grund können in anderen Betriebsumgebungen gewonnene Ergebnisse stark davon abweichen. Einige Messungen wurden unter Umständen auf Systemen im Entwicklungsstadium durchgeführt und es kann nicht garantiert werden, dass diese Messungen auf allgemein verfügbaren Systemen zum gleichen Ergebnis führen. Darüber hinaus wurden einige Messungen unter Umständen durch Extrapolation bestimmt. Die tatsächlichen Ergebnisse können hiervon abweichen. Die Benutzer dieses Dokuments sollten die entsprechenden Daten für die jeweils vorliegende Umgebung prüfen.

Informationen zu Produkten von Drittanbietern wurden von den Anbietern des jeweiligen Produkts, aus deren veröffentlichten Ankündigungen oder anderen, öffentlich verfügbaren Quellen bezogen. IBM hat diese Produkte nicht getestet und kann die Genauigkeit bezüglich Leistung, Kompatibilität oder anderen Behauptungen nicht bestätigen, die sich auf Drittanbieter-Produkte beziehen. Fragen bezüglich der Funktionen von Drittanbieter-Produkten sollten an die Anbieter der jeweiligen Produkte gerichtet werden.

Alle Aussagen bezüglich der zukünftigen Ausrichtung von IBM oder der Absichten des Unternehmens können ohne vorherige Ankündigung geändert oder zurückgenommen werden und stellen lediglich Ziele und Vorgaben dar.

Diese Informationen enthalten Beispiele zu Daten und Berichten, die im täglichen Geschäftsbetrieb Verwendung finden. Um diese so vollständig wie möglich zu illustrieren, umfassen die Beispiele Namen von Personen, Unternehmen, Marken und Produkten. Alle diese Namen sind fiktiv und jegliche Ähnlichkeit mit Namen und Adressen realer Unternehmen ist rein zufällig.

Unter Umständen werden Fotografien und farbige Abbildungen nicht angezeigt, wenn Sie diese Informationen nicht in gedruckter Form verwenden.

Marken

IBM, das IBM-Logo, ibm.com und SPSS sind Marken der IBM Corporation und in vielen Ländern weltweit registriert. Eine aktuelle Liste der IBM-Marken finden Sie im Internet unter *http://www.ibm.com/legal/copytrade.shtml*.

Intel, das Intel-Logo, Intel Inside, das Intel Inside-Logo, Intel Centrino, das Intel Centrino-Logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium und Pentium sind Markten oder eingetragene Marken der Intel Corporation oder der Tochtergesellschaften des Unternehmens in den USA und anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA, anderen Ländern oder beidem.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken der Microsoft Corporation in den USA, anderen Ländern oder beidem.

UNIX ist eine eingetragene Marke der The Open Group in den USA und anderen Ländern.

Java und alle Java-basierten Marken sowie Logos sind Marken von Sun Microsystems, Inc. in den USA, anderen Ländern oder beidem.

Andere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein.

Index