IBM SPSS Modeler 18.5 Source，Process 和 Output 节点

注

在使用本资料及其支持的产品之前，请阅读第 305 页的『注意事项』中的信息。

产品信息

本版本适用于的版本 18、发行版 4、IBM® SPSS Modeler 的修订 0 以及所有后续版本和修改，除非在新版本中另有说明

内容

前言

IBM SPSS Modeler 是 IBM 企业强度的数据挖掘工作台。 SPSS Modeler 通过深度的数据分析帮助组织改进与客户和市民的关系。 组织通过借助源自 SPSS Modeler 的洞察力可以留住优质客户，识别交叉销售机遇，吸引新客户，检测欺诈，降低风险，促进政府服务交付。

SPSS Modeler 的可视化界面让用户可以应用他们自己的业务专长，这将生成更加强有力的预测模型，缩减实现解决方案所需时间。 SPSS Modeler 提供了多种建模技术，例如预测、分类、分割和关联检测算法。 模型创建成功后，通过 IBM SPSS Modeler Solution Publisher，在广泛的企业内交付给决策者，或通过数据库交付。

关于 IBM Business Analytics

IBM Business Analytics 软件提供完整、一致和正确的信息，决策人依据此信息来提高业务性能。 企业智能、预测分析、财务业绩和战略管理的完整产品组合，和分析应用程序一起提供对当前业绩的清晰、直接和实用的洞察力，以及预测未来结果的能力。 结合丰富的行业解决方案，久经证明的实践和专业服务以及各种规模的组织都能够实现最高生产力、确信地自动作出决策以及获取更好的结果。

作为此产品服务组合的组成部分，IBM SPSS Predictive Analytics 软件可帮助组织预测未来事件，并在该洞察的基础上提前行动以实现更好的业务结果。 减少欺诈和降低风险时，世界范围的商业、政府和学术客户都依赖 IBM SPSS 技术作为吸引、保留和增加客户的竞争优势。 通过在日常活动中融入 IBM SPSS 软件，成为预测企业的组织可指引并实现决策的自动化，以满足企业目标并实现可衡量的竞争优势。 有关详细信息或要联系一位代表，请访问 http://www.ibm.com/spss。

技术支持

技术支持可供维护客户使用。 客户可就 IBM 产品使用问题或某一受支持硬件环境的安装帮助寻求技术支持。 要寻求技术支持，请访问 IBM Web 站点：http://www.ibm.com/support。 请求帮助时，请准备好标识您自身、组织和支持协议。

第 1 章 关于 IBM SPSS Modeler

IBM SPSS Modeler 是一组数据挖掘工具，通过这些工具可以采用商业技术快速建立预测性模型，并将其应用于商业活动，从而改进决策过程。 IBM SPSS Modeler 参照行业标准 CRISP-DM 模型设计而成，可支持从数据到更优商业成果的整个数据挖掘过程。

IBM SPSS Modeler 提供了各种借助机器学习、人工智能和统计学的建模方法。 通过建模选项板中的方法，您可以根据数据生成新的信息以及开发预测模型。 每种方法各有所长，同时适用于解决特定类型的问题。

SPSS Modeler 可以作为独立产品购买，也可以作为客户端与 SPSS Modeler Server 一起使用。 同时提供了大量其他选项，以下各节将对这些选项进行概述。 有关更多信息，请参阅 https://www.ibm.com/analytics/us/en/technology/spss/。

IBM SPSS Modeler 产品

IBM SPSS Modeler 系列产品及关联的软件包括以下各项。

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console（包含在 IBM SPSS Deployment Manager 中）
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS 协作和部署服务 的 IBM SPSS Modeler Server 适配器

IBM SPSS Modeler

SPSS Modeler 是具有完整功能的产品，它安装并运行于个人计算机上。 您可以在本地方式作为独立产品运行 SPSS Modeler，也可以在分布方式下将其与 IBM SPSS Modeler Server 一起使用来提高大型数据集的性能。

借助 SPSS Modeler，您可以快速直接地构建准确的预测模型，而不进行编程。 通过使用唯一可视界面，您可以轻松地查看数据挖掘过程。 借助该产品随附的高级分析支持，您可以发现数据中先前隐藏的模式和趋势。 您可以构建结果模型并了解影响结果的因素，从而利用业务机会并降低风险。

SPSS Modeler 推出了两个版本：SPSS Modeler Professional 和 SPSS Modeler Premium。 有关更多信息，请参阅主题 第 2 页的『IBM SPSS Modeler 版本』。

IBM SPSS Modeler Server

SPSS Modeler 使用客户端/服务器体系结构将资源集约型操作的请求分发给功能强大的服务器软件，因而使大数据集的传输速度大大加快。

SPSS Modeler Server 是一个单独授权的产品，在分布分析方式下，该产品在安装了一个或多个 IBM SPSS Modeler 的服务器主机上持续运行。 这种运行方式大大提高了 SPSS Modeler Server 对大型数据集的处理速度，因为在服务器上可以运行耗用内存的操作，并且无需将数据下载到客户端计算机上。 IBM SPSS Modeler Server 还提供对 SQL 优化和数据库内建模功能的支持，从而在性能和自动化方面带来更多优势。

IBM SPSS Modeler Administration Console

Modeler Administration Console 是一个图形用户界面，用于管理多个 SPSS Modeler Server 配置选项，这些选项还可以通过选项文件进行配置。 控制台包含在 IBM SPSS Deployment Manager，可以用于监视和配置 SPSS Modeler Server 安装，并且可供当前 SPSS Modeler Server 客户免费使用。 应用程序只能安装在 Windows 计算机上；但是它可以管理安装在任何受支持平台上的服务器。

IBM SPSS Modeler Batch

数据挖掘通常是交互过程，因此，还可以从命令行运行 SPSS Modeler 而不需要图形用户界面。 例如，您可能具有长时间运行或重复任务，并且希望在用户不进行干预的情况下执行这些任务。 SPSS Modeler Batch 是该产品的一个特殊版本，可提供对 SPSS Modeler 完整分析性能的支持，而无需访问常规的用户界面。 要使用 SPSS Modeler Batch，需要 SPSS Modeler Server。

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher 是一个工具，它使您能够创建 SPSS Modeler 流的打包版本，该版本的流可以由外部运行时引擎运行或者可以嵌入在外部应用程序中。 通过这种方式，您可以发布和部署完整的 SPSS Modeler 流以用于未安装 SPSS Modeler 的环境。 SPSS Modeler Solution Publisher 作为 IBM SPSS 协作和部署服务-评分 服务的组成部分分发，需要单独的许可证。 通过此许可证，您可以接收 SPSS Modeler Solution Publisher Runtime，它使您能够执行已发布的流。

有关 SPSS Modeler Solution Publisher 的更多信息，请参阅 IBM SPSS 协作和部署服务 文档。 IBM SPSS 协作和部署服务 IBM 文档包含名为“IBM SPSS Modeler Solution Publisher”和“IBM SPSS Analytics Toolkit”的部分。

IBM SPSS 协作和部署服务 的 IBM SPSS Modeler Server 适配器

IBM SPSS 协作和部署服务 的一些适配器使 SPSS Modeler 和 SPSS Modeler Server 能够与 IBM SPSS 协作和部署服务 存储库进行交互。 通过这种方式，部署到存储库的 SPSS Modeler 流可以由多个用户共享，或者从瘦客户端应用程序 IBM SPSS Modeler Advantage 进行访问。 请将适配器安装在托管存储库的系统上。

IBM SPSS Modeler 版本

SPSS Modeler 推出了下列版本。

SPSS Modeler Professional

SPSS Modeler Professional 提供处理大多数类型的结构化数据所需要的所有工具，例如 CRM 系统中跟踪的行为和交互、人口统计信息、采购行为和销售数据。

SPSS Modeler Premium

SPSS Modeler Premium 是一个单独授权的产品，它对 SPSS Modeler Professional 进行了扩展，以便后者能够处理专门的数据和非结构化文本数据。 SPSS Modeler Premium 包含 IBM SPSS Modeler 文本分析：

IBM SPSS Modeler 文本分析 采用先进语言技术和自然语言处理 (NLP)，以快速处理大量非结构化文本数据，提取和组织关键概念，以及将这些概念分为各种类别。 提取的概念和类别可以与现有的结构化数据（例如人口统计信息）相结合，并且可借助 IBM SPSS Modeler 的全套数据挖掘工具进行建模，以此实现更好更集中的决策。

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription 提供与传统 IBM SPSS Modeler 客户端完全相同的预测性分析功能。 通过 Subscription 版本，您可以定期下载产品更新。

文档

可从 SPSS Modeler 中的**帮助**菜单获取文档。 这样会打开始终可在产品外部访问的在线 IBM 文档。

每个产品的完整文档（包括安装指示信息）也在以下位置以 PDF 格式提供：https://www.ibm.com/support/pages/spss-modeler-185-documentation。

SPSS Modeler Professional 文档

SPSS Modeler Professional 文档套件（安装指示信息除外）如下。

- **IBM SPSS Modeler 用户指南**。 对于使用 SPSS Modeler 的一般简介，包括如何构建数据流、处理缺失值、构建 CLEM 表达式处理项目和报告，以及将用于部署的流打包到 IBM SPSS 协作和部署服务 或 IBM SPSS Modeler Advantage。
- **IBM SPSS Modeler Source、Process 和 Output 节点**。 描述用于以不同格式读取、处理和输出数据的所有节点。 实际上这表示所有节点而非建模节点。
- **IBM SPSS Modeler Modeling 节点**。 描述所有用于创建数据挖掘模型的节点。 IBM SPSS Modeler 提供了各种借助机器学习、人工智能和统计学的建模方法。
- **IBM SPSS Modeler 应用程序指南**。 本指南中的示例旨在为具体的建模方法和技术提供具有针对性的简介。 还可以从“帮助”菜单获取本指南的联机版本。 有关更多信息，请参阅主题 第 3 页的『应用程序示例』。
- **IBM SPSS Modeler Python 脚本编制和自动化**。 通过编写 Python 脚本实现系统自动化的相关信息，其中包括可以用于处理节点和流的属性的信息。
- **IBM SPSS Modeler 部署指南**。 有关在 IBM SPSS Deployment Manager 下以处理作业的步骤形式运行 IBM SPSS Modeler 流的信息。
- **IBM SPSS Modeler 数据库内挖掘指南**。 有关如何利用数据库的功能通过第三方算法来改进性能并增强分析功能的信息。
- **IBM SPSS Modeler Server 管理和性能指南**。 提供有关如何配置和管理 IBM SPSS Modeler Server 的信息。
- **IBM SPSS Deployment Manager 用户指南**。 有关使用 Deployment Manager 应用程序中包含的管理控制台用户界面来监视和配置 IBM SPSS Modeler Server 的信息。
- **IBM SPSS Modeler CRISP-DM 指南**。 借助 CRISP-DM 方法进行 SPSS Modeler 数据挖掘的分步指南。
- **IBM SPSS Modeler Batch 用户指南**。 提供在批处理方式下使用 IBM SPSS Modeler 的完整指导，包括批处理方式执行和命令行自变量的详细信息。 本指南仅以 PDF 格式提供。

SPSS Modeler Premium 文档

SPSS Modeler Premium 文档套件（安装指示信息除外）如下。

- **SPSS Modeler 文本分析 用户指南**。 提供有关将文本分析与 SPSS Modeler 配合使用的信息，包括文本挖掘节点、交互式工作台、模板和其他资源。

应用程序示例

SPSS Modeler 中的数据挖掘工具可以帮助解决很多业务和组织问题，应用程序示例将提供有关特定建模方法和技术的简要的针对性说明。 此处使用的数据集比某些数据挖掘器管理的大量数据存储小得多，但涉及的概念和方法可扩展到实际应用程序。

要访问示例，请在 SPSS Modeler 中单击“帮助”菜单中的**应用程序示例**。

数据文件和样本流安装在产品安装目录下的 Demos 文件夹中。 有关更多信息，请参阅 第 3 页的『Demos 文件夹』。

数据库建模示例。 请参阅 *IBM SPSS Modeler* 数据库内挖掘指南 中的示例。

脚本编制示例。 请参阅 *IBM SPSS Modeler* 脚本编写与自动化指南 中的示例。

Demos 文件夹

与应用程序示例配合使用的数据文件和样本流安装在产品安装目录下的 Demos 文件夹中（例如：`C:\Program Files\IBM\SPSS\Modeler\<version>\Demos`）。也可以从 Windows"开始"菜单上的 IBM SPSS Modeler 程序组访问此文件夹，或者通过单击**文件** > **打开流**对话框中最近的目录列表中的 Demos 来进行访问。

许可证跟踪

当您使用 SPSS Modeler 时，系统会定期跟踪并记录许可证使用情况。 所记录的许可证度量为 *AUTHORIZED_USER* 和 *CONCURRENT_USER*，并且记录的度量类型取决于您针对 SPSS Modeler 具有的许可证类型。

产生的日志文件可由 IBM License Metric Tool 处理，通过该工具可生成许可证使用情况报告。

许可证日志文件是在记录 SPSS Modeler Client 日志文件的同一目录中创建的（缺省情况下，为 `%ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log`）。

第 2 章 源节点

概述

通过源节点，可以导入以多种格式存储的数据，例如平面文件、IBM SPSS Statistics (.sav)、SAS、Microsoft Excel 和 ODBC 兼容关系数据。 也可以使用用户输入节点生成综合数据。

“源”选项板包含下列节点：

Analytic Server 源使您可以在 Hadoop 分布式文件系统 (HDFS) 上运行流。 Analytic Server 数据源中的信息可以来源于各种位置，例如文本文件和数据库。 有关更多信息，请参阅主题 第 9 页的『Analytic Server 源节点』。

“数据库”节点可用于使用 ODBC（开放数据库连接）从多种其他数据包中导入数据，这些数据包包括 Microsoft SQL Server、Db2 和 Oracle 等。 有关更多信息，请参阅主题 第 13 页的『“数据库源”节点』。

自由格式文件节点读取自由格式字段文本文件中的数据，即，其记录包含固定数量的字段，但包含不定数量字符的文件。 此节点对于具有固定长度标题文本和某些特定类型注解的文件也非常有用。 有关更多信息，请参阅主题 第 22 页的『“变量文件”节点』。

固定文件节点会从固定字段文本文件（即文件字段不定界，而是从相同的位置开始且长度固定）中导入数据。 机器生成的数据或遗存数据通常以固定字段格式存储。 有关更多信息，请参阅主题 第 25 页的『固定文件节点』。

Statistics 文件节点从 IBM SPSS Statistics 使用的 *.sav* 或 *.zsav* 文件格式以及保存在 IBM SPSS Modeler 中的缓存文件（也使用同一格式）读取数据。

数据收集 节点从符合 数据收集 数据模型的市场调查软件所用的各种格式中导入调查数据。 必须安装 数据收集 Developer Library 才可使用此节点。 有关更多信息，请参阅主题 第 26 页的『数据收集 节点』。

The IBM Cognos 源节点从 Cognos Analytics 数据库导入数据。

IBM Cognos TM1 源节点从 Cognos TM1 数据库导入数据。

SAS 文件节点可将 SAS 数据导入到 IBM SPSS Modeler 中。 有关更多信息，请参阅主题 第 34 页的『SAS 源节点』。

Excel 节点可从 Microsoft Excel 以 .xlsx 文件格式导入数据。 不要求指定 ODBC 数据源。 有关更多信息，请参阅主题 第 35 页的『Excel 源节点』。

“XML 源”节点将 XML 格式的数据导入到流中。 可以导入单个文件，也可以导入某个目录中的所有文件。 您可以选择性地指定模式文件，以便从中读取 XML 结构。

用户输入节点提供了一种用于创建综合数据的简单方式 - 可以从头开始创建也可以通过更改现有数据进行创建。 此节点非常有用，例如，在希望为建模创建测试数据集时，即可使用此节点。 有关更多信息，请参阅主题 第 37 页的『用户输入节点』。

“模拟生成”节点提供了一种生成模拟数据的简单方法 - 使用用户指定的统计分布从头开始生成数据，或者使用对现有历史数据运行“模拟拟合”节点而获取的分布自动生成数据。 当您想要在模型输入存在不确定性的情况下评估预测模型的结果时，这十分有用。

您可以使用“地理空间”源节点将地图或空间数据引入到数据挖掘会话中。 有关更多信息，请参阅主题 第 52 页的『“地理空间”源节点』。

JSON 源节点从 JSON 文件导入数据。

要开始执行流，可将源节点添加到流工作区中。 然后，双击该节点以打开其对话框。 通过此对话框中的各种选项卡能够读取数据；查看字段和值；设置各种选项，包括过滤器、数据类型、字段角色和缺失值检查。

设置字段存储类型和格式

使用固定“文件节点”、“变量文件”、“XML 源”和“用户输入”节点的“数据”选项卡中的选项，可以在 IBM SPSS Modeler 中导入或创建字段时为字段指定存储类型。 对于“固定文件”节点、“变量文件”节点和“用户输入”节点，您还可以指定字段格式化以及其他元数据。

对于从其他源读取的数据，存储自动确定，但是可以在“填充”节点或“派生”节点中使用转换函数（例如，`to_integer`）进行更改。

字段 使用**字段**列可以查看并选择当前数据集中的字段。

覆盖 选中**覆盖**列中的复选框可以激活**存储**和**输入格式**列中的选项。

数据存储

存储格式描述数据在某个字段中的存储方式。 例如，值为 1 和 0 的字段存储整型数据。 这点与测量级别明显不同，测量级别描述的是数据的使用方法，而且不影响存储。 例如，您可能希望将值为 1 和 0 的某个整数字段的测量级别设置为标志。 这通常表明 1 = 真 ， 0 = 假 。 存储格式必须在数据源中确定，而测量级别可以使用“类型”节点在流中的任意点上进行更改。 有关更多信息，请参阅主题 第 104 页的『测量级别』。

可用存储类型有：

- **字符串** 用于包含非数字数据（也称为字母数字数据）的字段。 字符串可以包含任何字符序列，比如 *fred* 、 *Class 2* 或 *1234* 。 注意：字符串中的数字不能用于计算。
- **整数** 值为整数的字段。
- **实数** 值为可能包含小数（不限于整数）的数字。 显示格式在“流属性”对话框中指定，并且可以被“类型”节点（“格式”选项卡）中的各个字段覆盖。

- **日期** 以标准格式指定的日期值，例如年月日（例如 2007-09-26）。 具体格式在“流属性”对话框中指定。
- **时间** 以持续时间形式测量的时间。 例如，某个服务电话持续 1 小时 26 分 38 秒，该时间可以根据“流属性”对话框中指定的当前时间格式表示为：01:26:38。
- **时间戳记** 同时包含日期和时间组成部分的值，例如 2007-09-26 09:04:00，具体同样取决于“流属性”对话框中的当前日期和时间格式。 请注意，需要用双引号将时间戳值括起来，以确保将此值解释为单一值而非单独的日期和时间值。 （同样适用于在用户输入节点中输入值时的情况。）
- **列表** 在 SPSS Modeler V17 中，随新测量级别“地理空间”和“集合”一起引入了“列表”存储字段，对于单个记录，此字段包含多个值。 存在所有其他存储类型的列表版本。

表 *1:* 列表存储类型图标

图标	存储类型
[A]	字符串列表
[◇]	整数列表
[#]	实数列表
[🕒]	时间列表
[▦]	日期列表
[▦]	时间戳记列表
[[]]	深度大于零的列表

另外，为了与“集合”测量级别配合使用，提供了下列测量级别的列表版本。

表 *2:* 列表测量级别图标

图标	测量级别
[✎]	连续值列表
[▮]	分类值列表
[🔑]	标志列表
[●]	名义值列表
[▮]	有序值列表

可以通过三个源节点（“Analytic Server”、“地理空间”或“变量文件”）中的某一个将列表导入到 SPSS Modeler 中，也可以在流中使用“派生”或“填充”字段操作节点创建列表。

有关“列表”及其与“集合”和“地理空间”测量级别的交互的更多信息，请参阅第 8 页的『列表存储以及相关联的测量级别』。

存储转换。 用户可使用各种转换函数来转换某个字段的存储格式，比如“填充”节点中的 to_string 和 to_integer。 有关更多信息，请参阅主题 第 120 页的『使用填充节点进行存储类型转换』。 请注意，转换函数（以及需要特定类型输入（如日期或时间值）的任何其他函数）取决于“流属性”对话框中指定的当前格式。 例如，如果要将值为 *Jan 2018*、*Feb 2018* 等等的字符串字段转换为日期存储格式，请选择 **MON YYYY** 作为流的缺省日期格式。 “派生”节点中也有可用的转换函数，用于派生计算过程中的临时转换。 另外，还可以使用“派生”节点来执行其他操作，比如使用分类值对字符串字段进行重新编码。 有关更多信息，请参阅主题 第 120 页的『使用导出节点对值进行重新编码』。

正在读入混合数据。 请注意，读取数字存储格式（整数、实数、时间、时间戳或日期）的字段中的数据时，任何非数字值将被设置为空或系统缺失。 这时因为 IBM SPSS Modeler 与某些应用程序不同，它不允许字段中含有混合存储类型。 为了避免出现混合存储类型，必须根据需要更改源节点中或外部应用程序中的存储类型，从而将任何具有混合数据的字段以字符串的格式读入。

字段输入格式（仅限“固定文件”节点、“变量文件”节点和“用户输入”节点）

对于除字符串和整数以外的所有存储类型，都可以使用下拉列表为选定的字段指定格式选项。 例如，从不同的语言环境中合并数据时，可能需要为一个字段指定句号 (.) 作为小数分隔符，而为另一个字段指定逗号分隔符。

在源节点中指定的输入选项会覆盖在流属性对话框中指定的格式选项；但是这些指定的输入选项不会在流的其他位置中保留。 根据所掌握的数据知识，这些选项可用来正确地解析输入数据。 指定的格式可用作将数据读取到 IBM SPSS Modeler 时解析数据的指导，但不能确定将数据读取到 IBM SPSS Modeler 之后应对其格式化的方式。 要在流的其他位置处基于每个字段指定格式，可使用“类型”节点的“格式”选项卡。 有关更多信息，请参阅主题 第 111 页的『字段格式设置选项卡』。

选项根据存储类型的变化而变化。 例如，对于实数存储类型，可以选择 **句号 (.)** 或 **逗号 (,)** 作为小数分隔符。 对于时间戳字段，从下拉列表中选择 **指定** 将打开一个单独的对话框。 有关更多信息，请参阅主题 第 112 页的『设置字段格式选项』。

对于所有存储类型，也可以选择 **流缺省值** 以便导入时使用流缺省设置。 流设置可在流属性对话框中指定。

其他选项

使用“数据”选项卡可指定其他几个选项：

- 要查看不再通过当前节点连接的数据（例如，训练数据）的存储设置，可选择 **查看未使用的字段设置** 。可通过单击 **清除** 清除遗产字段。
- 在此对话框中操作的任何时刻，都可单击 **刷新** 以从数据源重新加载字段。 在更改到源节点的数据连接时，或在对话框的选项卡之间进行操作时，此操作都非常有用。

列表存储以及相关联的测量级别

为了处理新测量级别“地理空间”和“集合”，在 SPSS Modeler V17 中引入了“列表”存储字段，对于单个记录，此字段包含多个值。 列表括在方括号 ([]) 中。列表示例为：`[1,2,4,16]` 和 `["abc", "def"]`。

可以通过三个源节点（“Analytic Server”、“地理空间”或“变量文件”）中的某一个将列表导入到 SPSS Modeler 中，可以在流中使用“派生”或“填充”字段操作节点创建列表，或者在使用“排名式条件”合并方法时由“合并”节点生成该列表。

列表被视为具有深度；例如，将各个项括在单个方括号内的简单列表（格式为 `[1,3]`）在 IBM SPSS Modeler 中记录为深度为零。 除了深度为零的简单列表以外，您还可以使用嵌套列表，这种列表中的每个值本身也是列表。

嵌套列表的深度取决于相关联的测量级别。 对于“无类型”，未设置深度限制；对于“集合”，深度为零；对于“地理空间”，深度必须介于 0 与 2（含首尾值）之间，具体取决于嵌套项数。

对于零深度列表，可以将测量级别设置为“地理空间”或“集合”。 这两个级别都是父测量级别，您可以在“**值**”对话框中设置测量子级别信息。 “集合”的测量子级别确定该列表中的元素的测量级别。 除“无类型”和“地理空间”以外的所有测量级别都可以用作“集合”的子级别。 “地理空间”测量级别有 6 个子级别，分别为“点”、“线串”、“多边形”、“多点”、“多线串”和“多多边形”；有关更多信息，请参阅第 105 页的『地理空间测量子级别』。

注: “集合”测量级别只能与深度为 0 的列表配合使用，“地理空间”测量级别只能与最大深度为 2 的列表配合使用，“无类型”测量级别可以与任意深度的列表配合使用。

以下示例使用“地理空间”测量子级别“点”和“线串”的结构来显示深度为零的列表与嵌套列表之间的差别：

- “地理空间”测量子级别“点”的字段深度为 0：

  `[1,3]` 两个坐标

  `[1,3,-1]` 三个坐标
- “地理空间”测量子级别“线串”的字段深度为 1：

  `[ [1,3], [5,0] ]` 两个坐标

  `[ [1,3,-1], [5,0,8] ]` 三个坐标

“点”字段（深度为零）是正常列表，其中的每个值都由两个或三个坐标构成。 “线串”字段（深度为 1）是点列表，其中的每个点都由进一步的列表值序列构成。

有关创建列表的更多信息，请参阅第 118 页的『派生列表或地理空间字段』。

不受支持的控制字符

SPSS Modeler 中的某些过程无法处理包含多种控制字符的数据。 如果您的数据使用这些字符，可能会看到错误消息，例如以下示例：

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

不支持的字符包括：`0x0` 到 `0x3F`（不包括 `0x0` 和 `0x3F`）以及 `0x7F`；但是，Tab 键 (`0x9(\t)`)、新行 (`0xA(\n)`) 以及回车符 (`0xD(\r)`) 不会导致出现问题。

如果看到与不受支持的字符有关的错误消息，请在流中“源”节点之后使用“填充”节点和 CLEM 表达式 **`stripctrlchars`** 替换这些字符。

Analytic Server 源节点

Analytic Server 源使您可以在 Hadoop 分布式文件系统 (HDFS) 上运行流。 Analytic Server 数据源中的信息可来源于各个位置，其中包括：

- HDFS 上的文本文件
- 数据库
- HCatalog

通常，将在 HDFS 上执行具有 Analytic Server 源的流；但是，如果流包含不支持在 HDFS 上执行的节点，那么系统会将流尽可能多的部分“推送回”Analytic Server，然后 SPSS Modeler Server 将尝试处理剩余的流。 您将需要对超大数据集进行二次抽样；例如，通过将“样本”节点放在流中进行二次抽样。

如果要使用您自己的 Analytic Server 连接而不是管理员定义的缺省连接，请取消选择**使用缺省 Analytic Server**，并选择您的连接。

数据源。 假定您或者 SPSS Modeler Server 管理员已建立连接，您可以选择包含要使用的数据的数据源。 数据源包含与该源关联的文件和元数据。 单击**选择**以显示可用数据源的列表。 有关更多信息，请参阅主题 第 9 页的『选择数据源』。

如果您需要创建新数据源或编辑现有数据源，请单击**启动数据源编辑器 ...**。

请注意，使用多个 Analytic Server 连接在控制数据流方面十分有用。 例如，使用 Analytic Server 源节点和导出节点时，您可能希望在流的不同分支中使用不同的 Analytic Server 连接，以便在每个分支运行时，使用自己的 Analytic Server，并且不会将任何数据拉取到 IBM SPSS Modeler Server。 请注意，如果某个分支包含多个 Analytic Server 连接，那么会将数据从 Analytic Server 拉取到 IBM SPSS Modeler Server。

选择数据源

“数据源”表显示了可用数据源的列表。 选择要使用的源，然后单击**确定**。

单击**显示所有者**以显示数据源所有者。

过滤依据使您能够按**关键字**对数据源列表进行过滤，这将根据数据源名称和数据源描述或者**所有者**检查过滤条件。 您可以输入下面描述的字符串、数字或通配符的组合作为过滤条件。 搜索字符串区分大小写。 单击**刷新**以更新“数据源”表。

_
可以使用下划线来表示搜索字符串中的任意单个字符。

%
可以使用百分号来表示搜索字符串中零个或零个以上字符的任意序列。

修改凭证

如果用于访问 Analytic Server 的凭证与用于访问 SPSS Modeler Server 的凭证不同，那么在 Analytic Server 上运行流时，您将需要输入 Analytic Server 凭证。 如果您不知道您的凭证，请联系服务器管理员。

受支持的节点

支持在 HDFS 上执行许多 SPSS Modeler 节点，但是某些节点的执行过程可能存在一些差异，而且一些节点当前不受支持。 本主题详细介绍了当前支持级别。

常规

- Analytic Server 将不接受加引号的 Modeler 字段名称中通常接受的某些字符。
- 要在 Analytic Server 中运行 Modeler 流，该流必须以一个或多个 Analytic Server“源”节点开头，并以单个建模节点或 Analytic Server“导出”节点结尾。
- 建议您将连续目标的存储设置为实数而不是整数。 评分模型始终会将实数值写入连续目标的输出数据文件，而评分的输出数据模型将遵循目标存储设置。 因此，如果连续目标具有整数存储，那么写入值与评分的数据模型将出现不匹配，并且此不匹配会在您尝试读取已评分数据时导致发生错误。
- 如果字段测量是“地理空间”，那么不支持 @OFFSET 的函数。

源

- 以 Analytic Server 源节点之外的任何内容开始的流将在本地运行。

记录操作

支持所有记录操作，但不支持“流式 TS”和“空间时间限制”节点。 下面提供了有关受支持节点功能的进一步说明。

选择

- 支持“派生”节点所支持的一组功能。

样本

- 不支持块级别抽样。
- 不支持复杂抽样方法。
- 不支持使用 `"Discard sample"` 的首个 n 抽样。
- 不支持使用 `N>20000` 的首个 n 抽样。
- 如果未设置 `"Maximum sample size"`，那么不支持“n 中取 1”抽样。
- `N * "Maximum sample size" > 20000` 时不支持“n 中取 1”抽样。
- 不支持随机 % 块级别抽样。
- 随机 % 当前支持提供种子值。

汇总

- 不支持连续键。 如果您要复用对数据进行排序而设置的现有流，然后在“汇总”节点中使用此设置，请更改此流以移除“排序”节点。
- 顺序统计（中位数、第一个四分位数和第三个四分位数）以近似方式计算，并支持“优化”选项卡。

排序

- 不支持“优化”选项卡。

在分布式环境中，存在有限数目的操作，这些操作将保留“排序”节点所确定的记录顺序。

- 后跟“导出”节点的“排序”将生成已排序的数据源。
- 后跟“样本”节点（包含**第一个**记录抽样）的“排序”将返回前 *N* 条记录。

通常，放置“排序”节点的位置应该尽可能靠近需要已排序记录的操作。

合并

- 不支持按顺序合并。

- 不支持“优化”选项卡。
- 合并操作的速度相对较慢。 如果 HDFS 中有可用空间，那么与合并以下每个流中的数据源相比，合并数据源一次并在这些流中使用合并后的源的速度更快。

R 转换

此节点中的 R 语法应该包含一次记录操作。

字段操作

支持所有字段操作，但不支持“匿名化”、“转置”、“时间间隔”和“历史记录”节点。 下面提供了有关受支持节点功能的进一步说明。

自动数据准备

- 不支持对节点进行训练。 支持将经过训练的“自动数据准备”节点中的变换应用于新数据。

导出

- 支持除序列函数以外的所有派生函数。
- 派生新字段以作为“计数”本质是上一个序列操作，因此不受支持。
- 不能在使用分割字段作为分割的同一流中派生这些字段；您将需要创建两个流，一个用于派生分割字段，另一个使用该字段作为分割。

填充

- 支持“派生”节点所支持的一组功能。

离散化

不支持以下功能。

- 最优分级
- 等级
- 分位数 -> 分位：值的总和
- 分位数 -> 结：保留在当前分级中并随机分配
- 分位数 -> 定制 N：超过 100 的值，以及任何 100 % N 不等于 0 的 N 值。

RFM 分析

- 不支持用于处理结的“保留在当前分级中”选项。 RFM 近因、频率和货币评分并非始终与 Modeler 根据同一数据计算得出的评分相匹配。 评分范围相同，但评分分配（分级数）可能相差 1。

图形

支持所有“图形”节点。

建模

支持下列“建模”节点：时间序列、TCM、保序 AS、扩展模型、树 AS、C&R 树、Quest、CHAID、线性、线性 AS、神经网络、GLE、LSVM、二阶 AS、随机树、STP、关联规则、XGBoost-AS、随即林和 K-Means-AS。 下面提供了有关这些节点的功能的进一步说明。

线性

根据大数据构建模型时，您通常希望将目标更改为“超大数据集”或指定分割。

- 不支持对现有 PSM 模型进行持续训练。
- 仅在定义了分割字段的情况下才建议使用标准模型构建目标，以免每个分割中的记录数过大，其中“过大”的定义取决于 Hadoop 集群中各个节点的能力。 相比之下，您还必须非常小心，以确保不要过于精细地定义分割，从而避免由于记录过少而无法构建模型。
- 不支持 Boosting 目标。
- 不支持 Bagging 目标。
- 记录较少时，建议不要使用超大数据集目标；通常，此目标将不会构建模型或者将构建降级模型。
- 不支持自动数据准备。 尝试根据具有多个缺失值的数据构建模型时，这可能会产生问题；通常，将在自动数据准备过程中插补这些缺失值。 变通方法是将树模型或神经网络与高级设置配合使用，以插补所选的缺失值。
- 对于分割模型，未计算准确性统计量。

神经网络

根据大数据构建模型时，您通常希望将目标更改为“超大数据集”或指定分割。

- 不支持对现有标准或 PSM 模型进行持续训练。
- 仅在定义了分割字段的情况下才建议使用标准模型构建目标，以免每个分割中的记录数过大，其中“过大”的定义取决于 Hadoop 集群中各个节点的能力。 相比之下，您还必须非常小心，以确保不要过于精细地定义分割，从而避免由于记录过少而无法构建模型。
- 不支持 Boosting 目标。
- 不支持 Bagging 目标。
- 记录较少时，建议不要使用超大数据集目标；通常，此目标将不会构建模型或者将构建降级模型。
- 如果数据中存在多个缺失值，请使用高级设置来插补缺失值。
- 对于分割模型，未计算准确性统计量。

C&R 树、CHAID 和 Quest

根据大数据构建模型时，您通常希望将目标更改为“超大数据集”或指定分割。

- 不支持对现有 PSM 模型进行持续训练。
- 仅在定义了分割字段的情况下才建议使用标准模型构建目标，以免每个分割中的记录数过大，其中“过大”的定义取决于 Hadoop 集群中各个节点的能力。 相比之下，您还必须非常小心，以确保不要过于精细地定义分割，从而避免由于记录过少而无法构建模型。
- 不支持 Boosting 目标。
- 不支持 Bagging 目标。
- 记录较少时，建议不要使用超大数据集目标；通常，此目标将不会构建模型或者将构建降级模型。
- 不支持交互式会话。
- 对于分割模型，未计算准确性统计量。
- 出现拆分字段时，在 Modeler 本地构建的树模型与 Analytic Server 所构建的树模型略有不同，从而生成不同的评分。 在这两种情况下，算法是有效算法；Analytic Server 所使用的算法更新。 假定树算法倾向于具有许多启发式规则，两个组件之间的差异是正常的。

模型评分

所有受建模功能支持的模型也受评分功能支持。 另外，对于下列节点，支持对以局部方式构建的模型块进行评分：C&RT、Quest、CHAID、线性和神经网络（无论是标准模型、促进式袋装模型还是数据集非常大型均如此）、回归、C5.0、Logistic、Genlin、GLMM、Cox、SVM、贝叶斯网络、二阶、KNN、决策列表、判别式、自学、异常检测、Apriori、Carma、K-Means、Kohonen、R 和文本挖掘。

- 不会对原始倾向或调整后的倾向进行评分。 作为变通方法，可以通过使用具有以下表达式的“派生”节点手动计算原始倾向来获得相同的结果：if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif

R

模型块中的 R 语法应该包含一次记录操作。

输出

支持下列节点：矩阵、分析、数据审核、变换、设置全局量、统计信息、均值和表。 下面提供了有关受支持节点功能的进一步说明。

数据审核

“数据审核”节点无法生成连续字段的方式。

平均值

“均值”节点无法生成标准误差或 95% 置信区间。

表

通过写入包含上游操作结果的临时 Analytic Server 数据源来支持“表”节点。 然后，“表”节点将分页读取该数据源的内容。

导出

流可以开始于 Analytic Server 源节点，并以 Analytic Server 导出节点之外的导出节点结束，但数据将从 HDFS 移至 SPSS Modeler Server，并最终移至导出位置。

“数据库源”节点

“数据库源”节点可用于使用 ODBC（开放数据库连接）从多种其他数据包中导入数据，这些数据包包括 Microsoft SQL Server、Db2 和 Oracle 等。

要对数据库进行读或写操作，您必须具有为相关数据库安装和配置的 ODBC 数据源，并根据需要具有读或写许可权。 IBM SPSS Data Access Pack 包含可用于此用途的 ODBC 驱动程序集，并且这些驱动程序可从下载站点获取。 如果您有关于创建或设置 ODBC 数据源权限方面的疑问，请与数据库管理员联系。

支持的 ODBC 驱动程序

有关使用 IBM SPSS Modeler 支持和测试的数据库和 ODBC 驱动程序的最新信息，请参阅公司支持站点上的产品兼容性矩阵 (http://www.ibm.com/support)。

在何处安装驱动程序

注: 必须在可能发生处理的每台计算机上安装和配置 ODBC 驱动程序。

- 如果您以本地（独立）模式运行 IBM SPSS Modeler，必须在本地计算机上安装驱动程序。
- 如果您以分布式模式针对远程 IBM SPSS Modeler Server 运行 IBM SPSS Modeler，需要在安装 IBM SPSS Modeler Server 的计算机上安装 ODBC 驱动程序。 对于 UNIX 系统中的 IBM SPSS Modeler Server，另请参阅本节中随后的“在 UNIX 系统中配置 ODBC 驱动程序”。
- 如果您需要从 IBM SPSS Modeler 和 IBM SPSS Modeler Server 中访问相同数据源，必须在两台计算机上都安装 ODBC 驱动程序。
- 如果您通过终端服务运行 IBM SPSS Modeler，需要在安装 IBM SPSS Modeler 的终端服务服务器上安装 ODBC 驱动程序。

访问数据库中的数据

要访问数据库中的数据，请完成下列步骤。

- 为要使用的数据库安装 ODBC 驱动程序并配置数据源。
- 在“**数据库**”节点对话框中，使用“表”方式或“SQL 查询”方式连接到数据库。
- 从数据库中选择表。
- 通过使用“**数据库**”节点对话框中的选项卡，您可以更改用法类型和过滤数据字段。

在相关文档主题中提供了更多有关上述步骤的详细信息。

注: 如果从 SPSS Modeler 中调用数据库存储过程 (SP)，那么您可能会看到返回了名为 `RowsAffected` 的单个输出字段，而不是返回期望的 SP 输出。 当 ODBC 未返回足够的信息，导致无法确定 SP 的输出数据模型时，将发生这种情况。 SPSS Modeler 对返回了输出的 SP 仅提供了有限的支持，因此建议您从 SP 中抽取 SELECT 并使用下列其中一项操作，而不要使用 SP。

- 创建基于 SELECT 的视图并在“数据库源”节点中选择该视图
- 直接在“数据库源”节点中使用 SELECT。

设置数据库节点选项

可以使用“数据库源”节点对话框的“数据”选项卡中的选项，来获取对数据库的访问并从选定的表中读取数据。

方式。 选择**表**可通过对话框控件连接到表。

选择 **SQL 查询** 以查询下面的使用 SQL 选择的数据库。 有关更多信息，请参阅主题 第 19 页的『查询数据库』。

数据源。 对于表方式和 SQL 查询方式，都可以在数据源字段中输入名称或者从下拉列表中选择**添加新的数据库连接**。

下列选项用于连接到数据库和选择表（使用对话框）：

表名称。 如果知道要访问的表的名称，那么可在表名字段中输入此名称。 否则，可单击 **选择** 按钮打开列出了可用的表的对话框。

将表和列名加上引号。 在数据库中进行查询时，指定是否要将表名和列名括入引号内（例如，这些名称包含空格或标点）。

- 选中 **需要时** 选项将 仅 在表名和字段名包括非标准字符时引用它们。 非标准字符包括非 ASCII 字符、空格字符和除全角句点 (.) 以外的所有非字母数字字符。
- 如果想给所有表名和字段名加引号，则选中**始终**。
- 如果从不想给表名和字段名加引号，则选中**从不**。

去除开头和结尾的空格。 选中选项以废弃字符串中开头和结尾的空格。

注: 在使用与不使用 SQL 回送的字符串之间的对比可能生成存在尾部空格的不同结果。

从 Oracle 读取空字符串。 在 Oracle 数据库中进行值的读写时，要注意，与 IBM SPSS Modeler 及大多数其他数据库不同，Oracle 将字符型空值等同于空值对待并存储。 这表示同样的数据从 Oracle 数据库中提取和从文件或其他数据库中提取其表现可能有所不同，可能会返回不同的结果。

添加数据库连接

要打开数据库，请先选择要连接的数据源。 从“数据”选项卡的数据源下拉列表中选择**添加新的数据库连接**。

此时将打开“数据库连接”对话框。

注: 要获取打开此对话框的替代方法，请从主菜单中选择：**工具 > 数据库...**

数据源。 列出可用的数据源。 如果看不到所需数据库，请向下滚动。 一旦选择了数据源并输入了任何密码，即可单击 **连接** 。 单击 **刷新** 以更新此列表。

mode. 请选择下列其中一种方式：

- **用户名和密码**。 如果数据源受密码保护，请输入用户名和关联密码。
- **存储的凭证**。 如果在 IBM SPSS 协作和部署服务 中已经配置了凭证，那么您可以选择此选项以便在存储库中浏览到该凭证。 此凭证的用户名和密码必须与访问数据库所需的用户名和密码匹配。

连接。 显示当前连接的数据库。

- **缺省值**。 可选择性地选择一个连接作为缺省值。 执行此操作将导致数据源或导出节点使用此预定义连接作为其数据源，然而，可根据需要对此连接进行编辑。
- **保存**。 选择性地选择一个或多个希望在后续会话中重新显示的连接。
- **数据源**。 当前连接的数据库的连接字符串。
- **预设**。 指示（使用一个 * 字符）是否为数据库连接指定了预设值。 要指定预设值，请单击此列中对应于数据库连接的行，然后从列表中选择“指定”。 有关更多信息，请参阅主题 第 17 页的『为数据库连接指定预设值』。

要删除连接，可从列表选择一个连接，然后单击 **删除** 。

驱动程序。 如果为方式选择**驱动程序**而不是**数据源**，请从列表中选择所需的驱动程序。

- 在**属性**字段中，输入数据库连接字符串。 请参阅数据库文档以获取要使用的正确字符串。
- 输入任何显示名称。
- 输入数据库用户名和密码并单击**连接**。

要对数据库进行读或写操作，您必须具有为相关数据库安装和配置的 ODBC 数据源，并根据需要具有读或写许可权。 IBM SPSS Data Access Pack 包含可用于此用途的 ODBC 驱动程序集，并且这些驱动程序可从下载站点获取。 如果您有关于创建或设置 ODBC 数据源权限方面的疑问，请与数据库管理员联系。

支持的 ODBC 驱动程序

有关使用 IBM SPSS Modeler 支持和测试的数据库和 ODBC 驱动程序的最新信息，请参阅公司支持站点上的产品兼容性矩阵 (http://www.ibm.com/support)。

在何处安装驱动程序

注: 必须在可能发生处理的每台计算机上安装和配置 ODBC 驱动程序。

- 如果您以本地（独立）模式运行 IBM SPSS Modeler，必须在本地计算机上安装驱动程序。
- 如果您以分布式模式针对远程 IBM SPSS Modeler Server 运行 IBM SPSS Modeler，需要在安装 IBM SPSS Modeler Server 的计算机上安装 ODBC 驱动程序。 对于 UNIX 系统中的 IBM SPSS Modeler Server，另请参阅本节中随后的“在 UNIX 系统中配置 ODBC 驱动程序”。
- 如果您需要从 IBM SPSS Modeler 和 IBM SPSS Modeler Server 中访问相同数据源，必须在两台计算机上都安装 ODBC 驱动程序。
- 如果您通过终端服务运行 IBM SPSS Modeler，需要在安装 IBM SPSS Modeler 的终端服务服务器上安装 ODBC 驱动程序。

在 UNIX 系统中配置 ODBC 驱动程序

缺省情况下，DataDirect 驱动程序管理器尚未配置 IBM SPSS Modeler Server 在 UNIX 中的使用。 要配置 UNIX 加载 DataDirect 驱动程序管理器，输入如下命令：

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

然后，如果要使用 UTF8 驱动程序包装器，请运行此命令：

```
ln -s libspssodbc_datadirect.so libspssodbc.so
```

或者，如果要使用 UTF16 驱动程序包装器，请运行此命令：

```

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

此命令可删除缺省链接并新建至 DataDirect 驱动程序管理器的链接。

注: 对于某些数据库，UTF16 驱动程序包装器是使用 SAP HANA 或 IBM Db2 CLI 驱动程序所必需的。 DashDB 需要 IBM Db2 CLI 驱动程序。

要配置 SPSS Modeler Server，请执行以下操作：

1. 通过将以下行添加到 `modelersrv.sh`，配置 SPSS Modeler Server 启动脚本 `modelersrv.sh` 以找出 IBM SPSS Data Access Pack `odbc.sh` 环境文件的来源：

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

   其中，`<pathtoSDAPinstall>` 是 IBM SPSS Data Access Pack 安装的完整路径。
2. 重新启动 SPSS Modeler Server。

此外，仅对于 SAP HANA 和 IBM Db2，在 `odbc.ini` 文件中，向 DSN 添加以下参数定义，以避免连接期间发生缓冲区溢出：

```
DriverUnicodeType=1
```

注: `libspssodbc_datadirect_utf16.so` 包装器还与其他 SPSS Modeler Server 支持的 ODBC 驱动程序兼容。

配置雪花或 Big Query 数据库

如果使用 SDAP 驱动程序 (使用需要从 JVM 实例开始的 Excel 节点或 XML 节点) 连接 Snowflake 或 BigQuery，那么可能会收到以下错误：

```
Internal Error. Failed to open Java VM.
```

要解决此问题，请编辑这些数据库的 SDAP ODBC 配置，以便 Modeler 和 SDAP 使用相同的 JVM 实例。 以下是如何为 Snowflake 配置 JVM 实例的示例。

Windows

1. 在 " **Progress DataDirect Snowflake ODBC 驱动程序设置** " 对话框中，转至 **SQL 引擎** 选项卡。
2. 将 **SQL Engine Modeler** 更改为 **1-Server**。
3. 单击 **编辑服务器设置**，然后更改 **Java 路径** 以指向随 Modeler 一起安装的 Java™ 。

   例如，*[INSTALLDIR_MODELER]*\jre\bin\java.exe
4. 在 **JVM 参数**中的缺省设置之后添加 -cp com.ddtek.snowflake.phoenix.sql.server.Server 。
5. 单击**应用**以保存所作的更改。
6. 编辑 Windows 的环境变量，并在变量路径中添加 *[INSTALLDIR_MODELER]*\jre\bin 和 *[INSTALLDIR_MODELER]*\lib 。
7. 启动 Modeler 客户机，然后重新连接到雪花驱动程序。

Linux®

1. 运行以下命令：
   a. export JAVA_HOME=/*[INSTALLDIR_MODELERSERVER]*/jre/
   b. export PATH=$JAVA_HOME/bin:$PATH
2. 在 *[INSTALLDIR_SDAP]*\odbc.ini 中，针对雪花驱动程序将 **SQLEngineMode=0** 更改为 **SQLEngineMode=1** 。
3. 在后台运行以下命令以启动 SQL 引擎：

```
java -Xmx1024m -cp /[INSTALLDIR_SDAP]/java/lib/snowflake.jar
com.ddtek.snowflake.phoenix.sql.server.Server -port 19947 &
```

4. 在同一控制台中重新启动 Modeler 服务器。

潜在数据库问题

根据您使用的数据库，可能存在一些您应该知道的潜在问题。

IBM Db2

当尝试在流中缓存从 Db2 数据库读取数据的节点时，可能会看到以下错误消息：

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

要配置 Db2 以使数据库内缓存能够在 SPSS Modeler 中正常工作，数据库管理员应创建“用户临时”表空间，并将该表空间的访问权授予相关 Db2 帐户。

我们建议在新表空间中使用页面大小 32768，因为这会增加对能够成功缓存的字段数量的限制。

IBM Db2 for z/OS

- 在启用置信度的情况下，使用生成的 SQL 对部分算法评分可能会在执行时返回错误。 该问题特定于 Db2 for z/OS；要解决此问题，请在 z/OS 上使用 SPSS Modeler Server Scoring Adapter for Db2。
- 当对 Db2 for z/OS 运行流时，如果空闲数据库连接的超时已启用并设置得较低，那么可能会遇到数据库错误。 在 Db2 for z/OS V8 中，缺省值从无超时更改为 2 分钟。 该解决方案是为了增加 Db2 系统参数 IDLE THREAD TIMEOUT （IDTHTOIN）的值，或将该值重置为 0。

Oracle

如果运行包含“汇总”节点的流，那么将 SQL 回送到 Oracle 数据库时，针对第一个和第三个四分位返回的值可能不同于以本机方式返回的值。

为数据库连接指定预设值

对于某些数据库，可以指定用于数据库连接的一些缺省设置。 这些设置将应用于数据库导出。

支持此功能的数据库类型如下。

- SQL Server Enterprise 和 Developer 版本。 有关更多信息，请参阅主题 第 17 页的『用于 SQL Server 的设置』。
- Oracle Enterprise 或 Personal 版本。 有关更多信息，请参阅主题 第 17 页的『用于 Oracle 的设置』。
- IBM Db2 for z/OS 和 Teradata 全都以类似方式连接到数据库或模式。 有关更多信息，请参阅主题 第 18 页的『用于 IBM Db2 for z/OS、IBM Db2 LUW 和 Teradata 的设置』。

如果连接到不支持此功能的数据库或模式，则会提示消息**无法为此数据库连接配置预设**。

用于 SQL Server 的设置

这些设置针对 SQL Server Enterprise 和 Developer 版本显示。

使用压缩。 如选中，使用压缩为导出创建表格。

压缩对象。 选择压缩层级。

- **行**。 启用行级别压缩（例如，SQL 中 `CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = ROW);` 的等效项）。
- **页面**。 启用页面级压缩（例如，SQL 中的 `CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = PAGE);`）。

用于 Oracle 的设置

Oracle 设置 -“基本”选项

这些设置针对使用“基本”选项的 Oracle Enterprise 或 Personal 版本显示。

使用压缩。 如选中，使用压缩为导出创建表格。

压缩对象。 选择压缩层级。

- **缺省值**。 启用缺省压缩（例如，SQL 中的 `CREATE TABLE MYTABLE(…) COMPRESS;`）。 在此情况下，它与**基本**选项的效果相同。
- **基本**。 启用基本压缩（例如，SQL 中的 `CREATE TABLE MYTABLE(…) COMPRESS BASIC;`）。

Oracle 设置 -“高级”选项

这些设置针对使用“高级”选项的 Oracle Enterprise 或 Personal 版本显示。

使用压缩。 如选中，使用压缩为导出创建表格。

压缩对象。 选择压缩层级。

- **缺省值**。 启用缺省压缩（例如，SQL 中的 `CREATE TABLE MYTABLE(…) COMPRESS;`）。 在此情况下，它与**基本**选项的效果相同。
- **基本**。 启用基本压缩（例如，SQL 中的 `CREATE TABLE MYTABLE(…) COMPRESS BASIC;`）。
- **OLTP**。 启用 OLTP 压缩（例如，SQL 中的 `CREATE TABLE MYTABLE(…)COMPRESS FOR OLTP;`）。
- **查询低/高**。 （仅限 Exadata 服务器）对查询启用混合列式压缩（例如，SQL 中的 `CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY LOW;` 或 `CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY HIGH;`）。 查询压缩非常适合用在数据仓储环境中；HIGH 提供比 LOW 更高的压缩比。
- **归档低/高**。 （仅限 Exadata 服务器）对归档启用混合列式压缩（例如，SQL 中的 `CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE LOW;` 或 `CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE HIGH;`）。 归档压缩非常适合用于压缩那些需要长时间存储的数据；HIGH 提供比 LOW 更高的压缩比。

用于 IBM Db2 for z/OS、IBM Db2 LUW 和 Teradata 的设置

在为 IBM Db2 for z/OS、IBM Db2 LUW 或 Teradata 指定预设值时，系统将提供您选择以下各项：

使用服务器评分适配器数据库或**使用服务器评分适配器模式**。 如果选中，那么将启用**服务器评分适配器数据库**或**服务器评分适配器模式**选项。

服务器评分适配器数据库或**服务器评分适配器模式** 请从下拉列表中选择所需的连接。

另外，对于 Teradata，还可以设置查询分段详细信息以提供附加的元数据，从而帮助完成工作负载管理、查询排序、识别查询和解析查询以及跟踪数据库使用情况等任务。

拼写查询分段。 选择是为您使用 Teradata 数据库连接的整个时间段设置一次查询分段（**针对会话**），还是每次运行流时都进行此设置（**针对事务**）。

注: 如果对流设置了查询分段，那么将该流复制到另一台机器时，该分段将丢失。 为了避免发生这种情况，您可以使用脚本来运行流，并在脚本中使用关键字 *querybanding* 来应用所需的设置。

需要的数据库权限

为使 SPSS Modeler 数据库功能正常运行，请向使用的任何用户标识授予对下列项的访问权：

Db2 LUW

SYSIBM.SYSDUMMY1
SYSIBM.SYSFOREIGNKEYS
SYSIBM.SYSINDEXES
SYSIBM.SYSKEYCOLUSE
SYSIBM.SYSKEYS
SYSIBM.SYSPARMS
SYSIBM.SYSRELS
SYSIBM.SYSROUTINES
SYSIBM.SYSROUTINES_SRC
SYSIBM.SYSSYNONYMS
SYSIBM.SYSTABCONST
SYSIBM.SYSTABCONSTPKC
SYSIBM.SYSTABLES
SYSIBM.SYSTRIGGERS
SYSIBM.SYSVIEWDEP
SYSIBM.SYSVIEWS
SYSCAT.TABLESPACES
SYSCAT.SCHEMATA

Db2/z

SYSIBM.SYSDUMMY1
SYSIBM.SYSFOREIGNKEYS
SYSIBM.SYSINDEXES
SYSIBM.SYSKEYCOLUSE
SYSIBM.SYSKEYS
SYSIBM.SYSPARMS
SYSIBM.SYSRELS
SYSIBM.SYSROUTINES
SYSIBM.SYSROUTINES_SRC
SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSIBM.SYSDUMMYU

SYSIBM.SYSPACKSTMT

Teradata

DBC.Functions

DBC.USERS

选择数据库表

已连接到数据源后，可以选择从特定的表或视图中导入字段。 在“数据库”对话框的“数据”选项卡上，可以在表名字段中输入表的名称，也可以单击**选择**以打开“选中表/视图”对话框，此对话框将列出可用的表和视图。

显示表所有者。 如果数据源要求在访问表之前必须指定表的所有者，请选中此选项。 如果数据源没有此要求，则取消选中此选项。

注: SAS 和 Oracle 数据库通常会要求显示表所有者。

表/视图。 选择要导入的表或视图。

显示。 列出当前连接到的数据源中的列。 单击下列选项之一可以自定义对可用表的查看：

- 单击 **用户表** 查看由数据库用户创建的普通数据库表。
- 单击 **系统表** 查看系统拥有的数据库表（例如，可提供有关数据库的信息的表，如索引的详细信息）。 此选项可用于查看 Excel 数据库中使用的选项卡。 （注意，也可以使用单独的 Excel 源节点。 请参阅主题第 35 页的『Excel 源节点』，以获取更多信息。）
- 单击 **视图** 基于查询查看包括一个或多个普通表的虚表。
- 单击 **同义名** 查看在数据库中创建的所有现有表的同义名。

名称/所有者过滤器。 使用这些字段可以按名称或所有者过滤显示的表的列表。 例如，键入 SYS 仅列出具有该所有者的表。 使用通配符搜索时，可使用下划线 (_) 表示所有的单字符，使用百分比符号 (%) 表示以任何顺序排列的零个或多个字符。

设为缺省值。 为当前用户保存当前设置作为缺省值。 当用户将来打开新的表选择器对话框时（仅在数据源名称和用户登录名相同的情况下），可恢复使用这些设置。

查询数据库

连接到数据源之后，可以选择使用 SQL 查询来导入字段。 从主对话框中，选择 **SQL 查询** 作为连接模式。 这将在对话框中添加一个查询编辑器窗口。 使用查询编辑器可创建或加载一个或多个 SQL 查询，其结果集将被读取到数据流中。

如果指定多个 SQL 查询，请使用分号 (;) 进行分隔，并确保不存在多 SELECT 语句。

要取消和关闭查询编辑器窗口，可选择 **表** 作为连接模式。

可以在 SQL 查询中包含 SPSS Modeler 流参数（一种用户定义变量）。 有关更多信息，请参阅第 20 页的『在 SQL 查询中使用流参数』。

加载查询。 单击此选项可打开文件浏览器，并装入先前保存的查询。

保存查询。 单击此选项可打开“保存查询”对话框，以保存当前的查询。

导入缺省值。 单击以导入示例 SQL SELECT 语句，该语句是使用对话框中选择的表和列自动构造的。

清除。 清除工作区的内容。 当想要重新开始编辑时，可使用此选项。

拆分文本。 缺省选项**从不**表示查询将完整地发送到数据库。 另外，您还可以选择**根据需要**，这表示 SPSS Modeler 会尝试解析查询，并确定是否存在应该逐条发送到数据库的 SQL 语句。

要点: 根据您所使用的数据库不同，SPSS Modeler 可能会尝试运行您输入的定制 SQL，以获取该 SQL 所生成的数据模型。 例如，在使用 MySQL 或 Google BigQuery 时，缺省情况下，在获取表模式时将调用 `SQLExecute()`；因此 SPSS Modeler 将运行 SQL 并获取数据模型。 对于大部分数据库驱动程序而言，情况并非如此。

如果您想避免这种情况，请参阅第 20 页的『使用定制数据库配置文件』，以详细了解在此类情况下定制 SPSS Modeler 处理 SQL 的方式。

在 SQL 查询中使用流参数

在编写 SQL 查询来导入字段时，可以包含之前定义的 SPSS Modeler 流参数。 支持所有类型的流参数。

下表显示如何在 SQL 查询中对流属性的一些示例进行解释。

表 *3:* 流参数示例

流参数名称（示例）	存储	流参数值	解释为
PString	String	ss	'ss'
PInt	整数	5	5
PReal	实数	5.5	5.5
PTime	时间	23:05:01	t{'23:05:01'}
PDate	日期	2011-03-02	d{'2011-03-02'}
PTimeStamp	TimeStamp	2011-03-02 23:05:01	ts{'2011-03-02 23:05:01'}
PColumn	未知	IntValue	IntValue

在 SQL 查询中，指定流参数的方式与在 CLEM 表达式中的指定方式相同，即通过 `'$P-<parameter_name>'` 来指定，其中 `<parameter_name>` 是已为流参数定义的名称。

引用字段时，存储类型必须定义为“未知”，并且参数值必须根据需要用引号括起。 因此，如果使用表中所显示的示例输入了 SQL 查询:

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

那么此查询将被视为:

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

如果要使用 `PColumn` 参数来引用 `IntValue` 字段，那么需要以如下方式指定查询以获取同一结果:

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

使用定制数据库配置文件

如果您需要定制 SPSS Modeler 处理 SQL 的方式，您可使用定制数据库配置文件。 这样，您就可以根据具体情况，进行一系列可能需要的定制。 此功能的完整文档尚未发布。 请与您的数据库管理员或 IBM 支持人员联系，以帮助您满足有关此功能的特定需要，但请注意，IBM 不支持通过这些并非由 IBM 提供的配置文件进行的定制。

实现定制数据库配置文件

1. 创建用于特定数据库的 `.cfg` 文件。

2. 根据需要添加选项。 可以进行广泛的定制。 请注意，该文件必须根据您的特定数据库正确命名并正确编排格式。
3. 将文件放置在 SPSS Modeler 服务器和/或 SPSS Modeler 客户端安装目录的 config 文件夹中。

该 .cfg 文件名必须采用 odbc-<db>-custom-properties.cfg 格式，其中 <db> 是下列其中一个数据库名称：

- bigquery
- db2
- greenplum
- hana
- hive
- impala
- informix
- mssql
- mysql
- neoview
- netezza
- oracle
- postgresql
- redshift
- soliddb
- sybase
- teradata
- vertica

示例：execute_while_getting_schema

如果将 MySQL 数据库与“数据库”源节点中的定制 SQL 配合使用，那么缺省情况下，在获取表模式时，MySQL 数据库驱动程序会调用 SQLExecute()。 因此，SPSS Modeler 需要运行 SQL 并获取数据模型。 如果您不想让 SPSS Modeler 获取数据模型，请完成下列步骤：

1. 创建名为 odbc-mysql-custom-properties.cfg 的文件。
2. 添加下一行，并将其设置为 N，以覆盖 MySQL 的缺省行为：

```
execute_while_getting_schema, N
```

3. 将文件复制到 SPSS Modeler 服务器和 SPSS Modeler 客户端安装的 config 目录中。

示例：sqlmx_sort_by

缺省情况下，在 SQL 回送期间，嵌套的 SQL 中不允许使用 order_by。 通过启用 order_by，可以让 SQL 回送强制应用于“样本”节点，前提是在它之前存在“排序”节点。 例如，要对 Google BigQuery 数据库启用 order_by，请完成下列步骤：

1. 创建名为 odbc-bigquery-custom-properties.cfg 的文件。
2. 添加下一行，并将其设置为 Y，以覆盖 Google BigQuery 的缺省行为：

```
sqlmx_sort_by, Y
```

3. 将文件复制到 SPSS Modeler 服务器和 SPSS Modeler 客户端安装的 config 目录中。

示例：uda_list_sql_basic and uda_list_sql_parameter

您可能想使用定制 SQL 来检索数据库函数和聚集函数。 例如，在 Oracle 数据库上，请完成下列步骤：

1. 创建名为 odbc-oracle-custom-properties.cfg 的文件。
2. 在该文件中添加以下行：

```
#Define the UDA (database window aggregates) sqls

uda_list_sql_basic, "SELECT '<src_database_name>',OBJECT_NAME,OWNER,'', '', CASE WHEN
OWNER='SYS' THEN 1 ELSE 0 END BUILTIN FROM ALL_ARGUMENTS WHERE OBJECT_ID IN (SELECT
OBJECT_ID FROM ALL_PROCEDURES WHERE AGGREGATE = 'YES') AND OBJECT_ID NOT IN (SELECT DISTINCT
OBJECT_ID FROM ALL_ARGUMENTS WHERE PLS_TYPE IS NULL) AND ARGUMENT_NAME IS NULL ORDER BY
OBJECT_ID"

uda_list_sql_parameter, "SELECT POSITION,
DATA_PRECISION,DATA_SCALE,DATA_TYPE,'',DATA_TYPE,'',0 FROM ALL_ARGUMENTS WHERE OBJECT_ID IN
(SELECT OBJECT_ID FROM ALL_PROCEDURES WHERE AGGREGATE = 'YES') AND OBJECT_ID NOT IN (SELECT
DISTINCT OBJECT_ID FROM ALL_ARGUMENTS WHERE PLS_TYPE IS NULL) ORDER BY OBJECT_ID,POSITION"
```

3. 将文件复制到 SPSS Modeler 服务器和 SPSS Modeler 客户端安装的 config 目录中。

uda_list_sql_basic 和 uda_list_sql_parameter 可以使用其他定制 SQL，前提是它们符合下列表模式（步骤 2 是示例）。

```
#table schema for uda_list_sql_basic
databaseName,function,schema,catalog,description,isBuiltIn

#table schema for uda_list_sql_parameter
position,precision,scale,returnType,returnTypeName,parameterTypes,parameterTypeNames,isVarArg
```

“变量文件”节点

可以使用“变量文件”节点从自由字段文本文件（其记录包含的字段数不变，但包含的字符数可改变）中读取数据，该文件又称为分隔文本文件。 此类型的节点也可用于具有固定长度的页眉文本和特定类型的注解的文件。 每次读取一条记录，并将这些记录传递到流中，直到读完整个文件。

有关读取地理空间数据的说明

如果此节点包含地理空间数据，并且此节点是作为平面文件的导出而创建，那么您必须执行一些额外的步骤以设置地理空间元数据。 有关更多信息，请参阅 第 24 页的『将地理空间数据导入到“变量文件”节点中』。

有关读取定界文本数据的说明

- 必须在每行末尾处用换行符分隔记录。 换行符不可转作他用（例如，包含在任何字段名称或字段值内）。 最好删除开头和结尾处的空格以节省空间（尽管这不是必需步骤）。 此节点可以选择性地去除这些空格。
- 必须使用逗号或其他字符（最好是仅用作分隔符，即该字符不能出现在字段名称或字段值中）分隔字段。 如果做不到这一点，那么可以将所有文本字段都括在双引号内，前提是所有字段名称或文本值均不包含双引号。 如果字段名称或字段值包含双引号，那么可以改为将文本字段括在单引号内，前提同样是字段值中的其他位置未使用单引号。 如果单引号和双引号都不能使用，那么需要对文本值进行修改以移除或替换定界字符或者单引号/双引号。
- 每一行（包括标题行）都应包含相同的字段数。
- 第一行应包含字段名称。 如果不是这种情况，请取消选中**从文件中读取字段名**以便对每个字段指定一个通用名称，例如 Field1 和 Field2 等等。
- 第二行必须包含数据的第一条记录。 不得存在空行或注释。
- 数字值不能包括千位分隔符或分组符号，例如，3,000.00 中不能使用逗号。 小数指示符（美式英语或英式英语中的句点或句号）只能在适当的情况下使用。
- 日期值和时间值应该采用“流选项”对话框中可以识别的某种格式，例如 DD/MM/YYYY 或 HH:MM:SS。 文件中的所有日期字段和时间字段最好采用同一种格式，并且任何包含日期的字段内的所有值必须采用同一格式。

设置“变量文件”节点的选项

请在““变量文件”节点”对话框的“文件”选项卡上设置选项。

文件 指定文件名。 可以输入文件名或单击省略按钮 (**...**) 来选择文件。 您一旦选择了文件，文件路径就会显示，并且文件内容将与定界符一起显示在下面的面板中。

您可以复制所显示的来自数据源的样本文本，并将其粘贴到下列控件中：EOL 注释字符和用户指定的定界符。 使用 Ctrl-C 和 Ctrl-V 进行复制和粘贴。

从文件中读取文件名 此选项在缺省情况下处于选中状态，用于将数据文件中的第一行作为列的标签进行处理。 如果第一行不是标题，则取消选中此选项，针对数据集中的字段数为每个字段自动分配一个一般名称，例如 *Field1, Field2* 。

指定字段数。 指定每个记录中的字段数。 只要记录以新行结束，就可以自动检测字段数。 也可以手动设置字段数。

跳过标题字符。 指定要忽略第一个记录的开头处的多少个字符。

EOL 注释字符。 指定字符，例如 # 或 !， 以指示数据中的注释。 无论这些字符之一出现在数据文件的何处，从该字符起直到下一个新行字符（不包括）之前的所有字符都将被忽略。

去除开头和结尾的空格。 选中选项可废弃导入字符串中开头和结尾的空格。

注: 在使用与不使用 SQL 回送的字符串之间的对比可能生成存在尾部空格的不同结果。

无效字符。 选择**丢弃**以删除数据源中的无效字符。 选择 **替换为** 用指定的符号（仅含一个字符）替换无效字符。 无效字符为空字符或指定的编码方法中不存在的任何字符。

编码。 指定使用的文本编码方法。 您可以选择系统缺省值、流缺省值或 UTF-8。

- 系统缺省值在 Windows 控制面板中指定，如果以分布式模式运行，则在服务器计算机上指定。
- 流缺省值在“流属性”对话框中指定。

小数符号 请选择数据源中使用的小数分隔符类型。 **流缺省值**是从流属性对话框的“选项”选项卡中选择的字符。 否则，在此对话框中选择**句号 (.)** 或**逗号 (,)** 作为小数分隔符读取所有的数据。

行定界符是换行符 要将换行符用作行定界符，而非用作字段分隔符，请选中此选项。 例如，如果由于行中的分隔符数为奇数而导致换行，那么此选项非常有用。 请注意，选择此选项表示您将无法选择“分隔符”列表中的**换行**。

注: 如果选中此选项，那么将去除数据行末尾的所有空白值。

定界符。 通过使用针对此控件列出的复选框，可以指定哪些字符（例如逗号 (,)）定义文件中的字段边界。 也可以为使用多个定界符的记录指定一个以上的定界符，例如“, |”。 缺省的定界符是逗号。

注: 如果逗号还定义为**小数符号**，那么此处的缺省设置不会起作用。 如果逗号既是**字段定界符**又是**小数符号**，请在**字段定界符**列表中选择**其他**。 然后在输入字段中手动指定逗号。

选择 **允许使用多个空白定界符** 可将多个相邻的空白定界符字符看作一个定界符。 例如，如果在一个数据值之后隔四个空格又有一个数据值，则这组数据将被看作是两个而不是五个字段。

要在其中扫描列和类型的行数 请指定要在其中扫描所指定数据类型的行数和列数。

自动识别日期和时间 要使 IBM SPSS Modeler 能够自动尝试将数据条目识别为日期或时间，请选中此复选框。 例如，这意味着 07-11-1965 之类的条目将被识别为日期，而 02:35:58 之类的条目将被识别为时间；然而，不明确的条目（例如 07111965 或 023558）由于数字之间没有分隔符而将显示为整数。

注: 为了避免使用来自先前 IBM SPSS Modeler 版本的数据文件时出现潜在的数据问题，缺省情况下，对于在 V13 以前的版本中保存的信息，未选中此复选框。

将方括号视为列表 如果选中此复选框，那么会将括在左右方括号之间的数据视为单个值，即使该内容包含逗号和双引号之类的定界字符也是如此。 例如，这可能包括两个或三个维度地理空间数据，在这些数据中，括在方括号内的坐标作为单个列表项进行处理。 有关更多信息，请参阅 第 24 页的『将地理空间数据导入到“变量文件”节点中』

引号。 通过使用下拉列表，可以指定导入时如何处理单引号和双引号。 可以选择 **丢弃** 所有引号，选择 **包含为文本** 将这些引号包括在字段值内，或选择 **成对丢弃** 匹配成对引号然后删除它们。 如果引号不匹配，则将收到错误消息。 选择 **丢弃** 和 **成对丢弃** 都会将字段值（不带引号）按一个字符串存储。

注: 使用**成对丢弃**时，将保留空格。 使用**丢弃**时，将除去引号内外的尾部空格（例如，**' " ab c" , "d ef " , " gh i " '** 将导致 **'ab c, d ef, gh i'**）。 使用**包含为文本**时，引号将视为常规符号，因此将自然去除开头和结尾的空格。

在此对话框中操作的任何时刻，都可单击**刷新**以从数据源重新加载字段。 在更改到源节点的数据连接时，或在对话框的选项卡之间进行操作时，此操作都非常有用。

将地理空间数据导入到“变量文件”节点中

如果节点包含地理空间数据，作为平面文件的导出而创建，并且在创建该节点的流中使用，那么该节点将保留地理空间元数据，并且无需执行更多配置步骤。

但是，如果将该节点导出并在另一个流中使用，那么地理空间列表数据将自动转换为字符串格式；您必须执行一些额外的步骤以复原列表存储类型及相关地理空间元数据。

有关列表的更多信息，请参阅第 8 页的『列表存储以及相关联的测量级别』。

有关可以设置为地理空间元数据的详细信息的更多信息，请参阅第 105 页的『地理空间测量子级别』。

要设置地理空间元数据，请完成下列步骤。

1. 在“变量文件”节点的“文件”选项卡上，选中**将方括号作为列表进行处理**复选框。 选中此复选框意味着将括在左右方括号之间的数据视为单个值，即使该内容包含逗号和双引号之类的定界字符也是如此。 未选中此复选框意味着将数据作为字符串存储类型进行读取，字段中的所有逗号都将作为定界符进行处理，这将导致不正确地解释数据结构。
2. 如果数据包含单引号或双引号，请相应地在**单引号**和**双引号**字段中选中**配对并废弃**选项。
3. 在“变量文件”节点的“数据”选项卡上，对于地理空间数据字段，请选中**覆盖**复选框，并将**存储**类型由字符串更改为列表。
4. 缺省情况下，列表**存储**类型设置为实数列表，并且列表字段的底层值存储类型设置为实数。 要更改底层值存储类型或深度，请单击**指定...** 以显示“**存储**”子对话框。
5. 在“**存储**”子对话框中，可以修改下列设置：
   - **存储** 指定数据字段的整体存储类型。 缺省情况下，存储类型设置为“列表”；但是，下拉列表包含所有其他存储类型（“字符串”、“整数”、“实数”、“日期”、“时间”和“时间戳记”）。 如果您选择了除“列表”以外的任何存储类型，那么**值存储**和**深度**选项不可用。
   - **值存储** 指定列表中的元素的存储类型，而不是字段的整体存储类型。 导入地理空间字段时，相关的存储类型只有“实数”和“整数”；缺省设置为“实数”。
   - **深度** 指定列表字段的深度。 所需的深度取决于地理空间字段的类型并遵循下列条件：
     - 点 - 0
     - 线串 - 1
     - 多边形 - 1
     - 多点 - 1
     - 多线串 - 2
     - 多多边形 - 2

     警告: 您必须了解要重新转换为列表的地理空间字段的类型以及该类字段的所需深度。 如果未正确设置此信息，那么将无法使用此字段。
6. 在“变量文件”节点的“类型”选项卡上，对于地理空间数据字段，请确保**测量**单元格包含正确的测量级别。 要更改测量级别，请在**测量**单元格中单击**指定...** 以显示“**值**”对话框。
7. 在“**值**”对话框中，将显示该列表的**测量**、**存储**和**深度**。 选择**指定值和标签**选项，并从**类型**下拉列表中为**测量**选择正确的类型。 根据**类型**不同，系统可能会提示您输入更多详细信息，例如数据是否表示 2 个或 3 个维以及要使用的坐标系。

固定文件节点

可以使用“固定文件”节点从固定字段文本文件（其字段没有被分隔，但开始位置相同且长度固定）中导入数据。 机器生成的数据或遗存数据通常以固定字段格式存储。 使用固定文件节点的“文件”选项卡，可以轻松地指定数据中列的位置和长度。

设置“固定文件”节点的选项

使用固定文件节点的“文件”选项卡能够将数据导入 IBM SPSS Modeler，并指定列的位置和记录的长度。 使用位于对话框中心的数据预览窗格，可以单击以添加箭头用来指定字段间的断点。

文件。 指定文件的名称。 可以输入文件名或单击省略按钮 (**...**) 来选择文件。 一旦选定了一个文件，即可显示此文件的路径，并且文件的内容将使用定界符分隔显示在下面的面板中。

数据预览窗格可用来指定列的位置和长度。 预览窗口顶部的标尺有助于测量变量的长度并指定变量间的断点。 通过单击字段上方的标尺区域可以指定断点线。 通过拖动可移动断点，而将其拖动到数据预览区域之外则可废弃断点。 标尺旨在处理 ASCII 字符。

- 每个断点线会自动将一个新字段添加到下面的字段表中。
- 由箭头表示的开始位置会被自动添加到下表中的开始列中。

面向行。 如果要跳过每个记录末尾的新行字符，请选中此选项。

跳过标题行。 指定要忽略第一个记录的开头处的行数。 这对忽略列标题非常有用。

记录长度。 指定每个记录中的字符数。

字段。 已为此数据文件定义的所有字段都在此处列出。 有以下两种定义字段的方式：

- 使用上述数据预览窗格交互指定字段。
- 通过向下面的表添加空字段行手动指定字段。 单击字段窗格右侧的按钮添加新字段。 然后在空字段中输入字段名、开始位置和长度。 这些选项会自动在数据预览窗格中添加箭头，并且可以轻松地调整这些箭头。

要删除以前定义的字段，可在列表选择该字段，然后单击红色的删除按钮。

启动。 指定字段中第一个字符的位置。 例如，如果记录的第二个字段开始于第十六个字符，则可以输入 16 作为起点。

长度。 为每个字段指定最长值中的字符数。 该值可为下一个字段确定截止点。

去除开头和结尾的空格。 选中此选项以废弃导入时字符串的开头和结尾的空格。

注: 在使用与不使用 SQL 回送的字符串之间的对比可能生成存在尾部空格的不同结果。

无效字符。 选择**废弃**可从数据输入中移除无效字符。 选择 **替换为** 用指定的符号（仅含一个字符）替换无效字符。 无效字符是空字符 (0) 或所有当前编码中不存在的字符。

编码。 指定使用的文本编码方法。 您可以选择系统缺省值、流缺省值或 UTF-8。

- 系统缺省值在 Windows 控制面板中指定，如果以分布式模式运行，则在服务器计算机上指定。
- 流缺省值在“流属性”对话框中指定。

小数符号。 选择在数据源中使用的小数分隔符类型。 **流缺省值**是从流属性对话框的“选项”选项卡中选择的字符。 否则，在此对话框中选择**句号 (.)** 或**逗号 (,)** 作为小数分隔符读取所有的数据。

自动识别日期和时间。 要使 IBM SPSS Modeler 能够自动尝试将数据条目识别为日期或时间，请选中此复选框。 例如，这意味着 07-11-1965 之类的条目将被识别为日期，而 02:35:58 之类的条目将被识别为时间；然而，不明确的条目（例如 07111965 或 023558）由于数字之间没有分隔符而将显示为整数。

注: 为了避免使用来自先前 IBM SPSS Modeler 版本的数据文件时出现潜在的数据问题，缺省情况下，对于在 V13 以前的版本中保存的信息，未选中此复选框。

类型的扫描行数。 指定对于指定的数据类型要扫描的行数。

在此对话框中操作的任何时刻，都可单击 **刷新** 以从数据源重新加载字段。 在更改到源节点的数据连接时，或在对话框的选项卡之间进行操作时，此操作都非常有用。

Statistics 文件节点

可以使用 Statistics 文件节点从已保存的 IBM SPSS Statistics 文件（*.sav* 或 *.zsav*）中直接读取数据。 现在可使用该格式替换 IBM SPSS Modeler 早期版本中的缓存文件。 如果想要导入已保存的缓存文件，则应使用 IBM SPSS Statistics 文件节点。

导入文件。 指定文件名。 可以输入文件名或单击省略按钮 (**...**) 来选择文件。 一旦选定了一个文件，即可显示此文件的路径。

文件受密码加密。 如果您知道该文件受密码保护，请选中此框；系统将提示您输入**密码**。 如果该文件受密码保护，但您未输入密码，那么在尝试切换至另一选项卡、刷新数据、预览节点内容或尝试执行包含节点的流时，将显示一条警告消息。

注: 受密码保护的文件只能由 IBM SPSS Modeler V16 或更高版本打开。

变量名称。 选择从 IBM SPSS Statistics *.sav* 或 *.zsav* 文件导入变量名称和标签时的处理方法。 在您使用 IBM SPSS Modeler 的整个过程中，所选的包含在此处的元数据会保留，并且可以再次导出以在 IBM SPSS Statistics 中使用。

- **读取名称和标签**。 选中此选项可将变量名称和标签同时读入 IBM SPSS Modeler。 缺省情况下将选中此选项，并且变量名称将显示在“类型”节点中。 根据流属性对话框中指定的选项，标签将显示在图表、模型浏览器和其他类型的输出中。 缺省情况下，将禁止在输出中显示标签。
- **读取标签作为名称**。 选择从 IBM SPSS Statistics *.sav* 或 *.zsav* 文件中读取描述性变量标签（而不是短字段名称），并在 IBM SPSS Modeler 中将这些标签用作变量名称。

值。 选择从 IBM SPSS Statistics *.sav* 或 *.zsav* 文件导入值和标签时的处理方法。 在您使用 IBM SPSS Modeler 的整个过程中，所选的包含在此处的元数据会保留，并且可以再次导出以在 IBM SPSS Statistics 中使用。

- **读取数据和标签**。 选中此选项可将实际值和值标签同时读入 IBM SPSS Modeler。 缺省情况下将选中此选项，并且这些值本身将显示在“类型”节点中。 根据流属性对话框中指定的选项，值标签将显示在表达式构建器、图表、模型浏览器和其他类型的输出中。
- **读取标签作为数据**。 如果要使用 *.sav* 或 *.zsav* 文件中的值标签而不是用于表示值的数字或符号代码，请选中此选项。 例如，对于含性别字段（其值 1 和 2 实际上分别代表男性和女性）的数据，选中此选项可将该字段转换为字符串，并将男性和女性作为实际值导入。

  选中此选项前考虑 IBM SPSS Statistics 数据中的缺失值非常重要。 例如，如果数字字段仅对缺失值使用标签 (0 = *No Answer* , -99 = *Unknown*)，则选中上述选项将仅导入值标签 *No Answer* 和 *Unknown* ，并会将字段转换为字符串。 在这种情况下，应在类型节点中导入值本身并设置缺失值。

使用字段格式信息来确定存储。 如果取消选中此复选框，那么将使用整数存储导入 *.sav* 文件中格式化为整数的字段值（即，在 IBM SPSS Statistics 的“变量视图”中指定为 F*n*.0 的字段）。 将除字符串外的所有其他字段值作为实数导入。

如果选中了此框（缺省情况），那么无论是否已在 *.sav* 文件中格式化为整数，除字符串以外的所有字段值都将作为实数导入。

多重响应集。 导入文件后，IBM SPSS Statistics 文件中定义的任何多重响应集都将自动被保留。 借助“过滤器”选项卡，您可以查看和编辑任意节点的多重响应集。 有关更多信息，请参阅主题 第 114 页的『编辑多重响应集』。

数据收集 节点

数据收集 源节点根据随附于 数据收集 产品的 Survey Reporter Developer Kit 导入调查数据。 此格式将案例数据（对调查期间所收集的问题的实际响应）与元数据（描述如何收集和组织案例数据）进行区分。 元数据包括问题文本、变量名称和描述、多响应变量定义、文本字符串的变换以及观测值数据结构的定义等信息。

注: 此节点需要随 数据收集 产品分发的 Survey Reporter Developer Kit。 除安装 Developer Kit 外，不需要其他配置。

注释

- 可以从平面或表格 VDATA 格式或者从分层 HDATA 格式中的源（如果这些源包含元数据源）读取调查数据。
- 通过使用元数据信息，可以自动实例化类型。
- 将调查数据导入到 SPSS Modeler 时，问题将转变为字段，同时每个被调查者对应一个记录。

数据收集导入文件选项

使用 数据收集 节点的“文件”选项卡可为要导入的元数据和观测值数据指定选项。

元数据设置

注: 要查看可用提供程序文件类型的完整列表，您需要安装随附于 数据收集 软件产品的 Survey Reporter Developer Kit。

元数据提供者。 可从 数据收集 Survey Reporter Developer Kit 所支持的多种格式导入调查数据。 包括下列可用的提供者类型:

- **DataCollectionMDD**。 从调查表定义文件 (*.mdd*) 中读取元数据。 这是标准的 数据收集 数据模型格式。
- **ADO 数据库**。 从 ADO 文件中读取观测值数据和元数据。 指定包含元数据的 *.adoinfo* 文件的名称和位置。 此 DSC 的内部名称是 *mrADODsc* 。
- **In2data 数据库**。 读取 In2data 观测值数据和元数据。 此 DSC 的内部名称是 *mrI2dDsc*。
- **数据收集日志文件**。 从标准 数据收集 日志文件中读取元数据。 通常，日志文件具有 *.tmp* 文件扩展名。 但是，某些日志文件可能具有其他文件扩展名。 如果必要，可以重命名该文件使其具有 *.tmp* 文件扩展名。 此 DSC 的内部名称是 *mrLogDsc* 。
- **Quancept 定义文件**。 将元数据转换为 Quancept 脚本。 指定 Quancept *.qdi* 文件的名称。 此 DSC 的内部名称是 *mrQdiDrsDsc* 。
- **Quanvert 数据库**。 读取 Quanvert 观测值数据和元数据。 指定 *.qvinfo* 或 *.pkd* 文件的名称和位置。 此 DSC 的内部名称是 *mrQvDsc* 。
- **数据收集参与数据库**。 读取工程的“样本和历史记录表”表并创建与这些表中的列相对应的派生分类变量。 此 DSC 的内部名称是 *mrSampleReportingMDSC* 。
- **统计文件**。 从 IBM SPSS Statistics *.sav* 文件中读取观测值数据和元数据。 将观测值数据写入 IBM SPSS Statistics *.sav* 文件以便在 IBM SPSS Statistics 中分析。 将 IBM SPSS Statistics *.sav* 文件中的元数据写入 *.mdd* 文件。 此 DSC 的内部名称是 *mrSavDsc* 。
- **Surveycraft 文件**。 读取 SurveyCraft 观测值数据和元数据。 指定 SurveyCraft *.vq* 文件的名称。 此 DSC 的内部名称是 *mrSCDsc* 。
- **数据收集脚本文件**。 从 mrScriptMetadata 文件中读取元数据。 通常，这些文件具有 *.mdd* 或 *.dms* 文件扩展名。 此 DSC 的内部名称是 *mrScriptMDSC* 。
- **Triple-S XML 文件**。 从 XML 格式的 Triple-S 文件中读取元数据。 此 DSC 的内部名称是 *mrTripleSDsc* 。

元数据属性。 这是可选选项，选择**属性**可指定要导入的调查版本及要使用的语言、环境和标签类型。 有关更多信息，请参阅主题 第 28 页的『数据收集导入元数据属性』。

观测值数据设置

注: 要查看可用提供程序文件类型的完整列表，您需要安装随附于 数据收集 软件产品的 Survey Reporter Developer Kit。

获取案例数据设置。 仅从 *.mdd* 文件中读取元数据时，单击**获取观测值数据设置**可确定哪些观测值数据源与选定的元数据关联，并确定访问给定的源所需的特定设置。 此选项仅用于 *.mdd* 文件。

案例数据提供者。 支持下列提供者类型:

- **ADO 数据库**。 使用 Microsoft ADO 接口读取观测值数据。 选择 OLE-DB UDL 作为观测值数据类型，并在观测值数据 UDL 字段中指定连接字符串。 有关更多信息，请参阅主题 第 29 页的『数据库连接字符串』。 此组件的内部名称是 *mrADODsc*。

- **定界文本文件 (Excel)**。 从以逗号分隔的 (.CSV) 文件中读取观测值数据，例如可通过 Excel 输出的文件。 内部名称是 *mrCsvDsc* 。
- **数据收集数据文件**。 从本机 数据收集 数据格式文件中读取观测值数据。 内部名称是 *mrDataFileDsc* 。
- **In2data 数据库**。 从 In2data 数据库 (*.i2d*) 文件中读取观测值数据和元数据。 内部名称是 *mrI2dDsc* 。
- **数据收集日志文件**。 从标准 数据收集 日志文件中读取观测值数据。 通常，日志文件具有 *.tmp* 文件扩展名。 但是，某些日志文件可能具有其他文件扩展名。 如果必要，可以重命名该文件使其具有 *.tmp* 文件扩展名。 内部名称是 *mrLogDsc* 。
- **量子数据文件**。 从任何 Quantum 格式的 ASCII 文件 (*.dat*) 中读取观测值数据。 内部名称是 *mrPunchDsc* 。
- **Quancept 数据文件**。 从 Quancept *.drs*、*.drz* 或 *.dru* 文件中读取观测值数据。 内部名称是 *mrQdiDrsDsc* 。
- **Quanvert 数据库**。 从 Quanvert *qvinfo* 或 *.pkd* 文件中读取观测值数据。 内部名称是 *mrQvDsc* 。
- **数据收集数据库 (MS SQL Server)**。 从 Microsoft SQL Server 数据库中读取观测值数据。 有关更多信息，请参阅主题 第 29 页的『数据库连接字符串』。 内部名称是 *mrRdbDsc2*。
- **统计文件**。 从 IBM SPSS Statistics *.sav* 文件中读取观测值数据。 内部名称是 *mrSavDsc* 。
- **Surveycraft 文件**。 从 Surveycraft *.qdt* 文件中读取观测值数据。 *.vq* 文件和 *.qdt* 文件必须在同一个目录下，并且都具有读写权限。 缺省情况下，这两个由 SurveyCraft 创建的文件位于不同目录之下，因此需要将一个文件移动至另一个文件所在目录下以便能够导入 SurveyCraft 数据。 内部名称是 *mrScDsc* 。
- **Triple-S 数据文件**。 以长度固定的格式或逗号分隔的格式从 Triple-S 数据文件中读取观测值数据。 内部名称是 *mr TripleDsc* 。
- **数据收集 XML**。 从 数据收集 XML 数据文件中读取观测值数据。 通常，此格式可用于将观测值数据从一个位置传输到另一个位置。 内部名称是 *mrXmlDsc* 。

案例数据类型。 指定从文件、文件夹、OLE-DB UDL 或 ODBC DSN 中读取观测值数据，并相应地更新对话框选项。 有效选项取决于提供者的类型。 对于数据库提供者，可以为 OLE-DB 或 ODBC 连接指定选项。 有关更多信息，请参阅主题 第 29 页的『数据库连接字符串』。

案例数据项目。 从 数据收集 数据库中读取观测值数据时，可以输入工程的名称。 对于所有其他的观测值数据类型，应将此设置留空。

变量导入

导入系统变量。 指定是否导入系统变量，其中包括表示访问状态的变量（进行中、已完成和完成日期等）。 您可以选择 **无** 、**所有** 或 **通用** 。

导入“Codes”变量。 控制导入代表代码（用于分类变量的开放式“其他”响应）的变量。

导入“SourceFile”变量。 控制导入包含已扫描响应图像文件名的变量。

将多响应变量导入为。 多响应变量可作为多标志字段（一个多二分法集）导入，此方法是用于新流的缺省方法。 在 12.0 之前的 IBM SPSS Modeler 版本中创建的流已将多个响应导入用逗号分隔值的单个字段。 仍然支持使用旧方法，以允许现有流按照之前的方式运行，但建议更新旧流以使用新方法。 有关更多信息，请参阅主题 第 29 页的『导入多重响应集』。

数据收集导入元数据属性

导入 数据收集 调查数据时，您可以在“元数据属性”对话框中指定要导入的调查版本及要使用的语言、环境和标签类型。 请注意，一次只能导入一种语言、环境和标签类型。

版本。 每个调查版本都可看作是用于收集观测值数据特定集合的元数据的一个快照。 随着调查表的更改，可创建多个调查版本。 可以导入最新版本、所有版本或特定的版本。

- **所有版本**。 如果要使用所有可用版本的组合（父集），请选中此选项。 （该父集有时称作父版本）。 版本之间存在冲突时，最新的版本通常优先于较早的版本。 例如，如果类别标签在所有的版本中各不相同，那么将使用最新版本中的文本。
- **最新版本**。 如果要使用最新版本，请选中此选项。
- **指定版本**。 如果要使用特定的调查版本，请选中此选项。

选择所有版本非常有用，例如，当您要为一个以上的版本导出观测值数据，且变量和类别定义已发生更改（这意味着在一个版本中收集的观测值数据在另一个版本中无效）时，即可选中此选项。 选择要为其导出观测值数据的所有版本意味着，在不发生因版本间差异而导致的有效性错误的情况下，通常可同时导出在不同版本中收集的观测值数据。 但是，因为版本有所更改，某些有效性错误仍可能发生。

语言。 问题和关联的文本可以多种语言存储在元数据中。 对于调查，可使用缺省语言，也可指定某种特定的语言。 如果某个项目在指定的语言中不可用，那么将使用缺省语言。

上下文。 选择要使用的用户上下文。 用户上下文可控制显示哪些文本。 例如，选择 **问题** 可显示问题文本，选择 **分析** 可显示适合在分析数据时显示的较短文本。

标签类型。 列出已定义的标签类型。 缺省类型为 **标签**，它可用于问题用户环境中的问题文本和分析用户环境中的变量说明。 针对指导、说明等等，可定义其他标签类型。

数据库连接字符串

使用 数据收集 节点通过 OLE-DB 或 ODBC 从数据库中导入观测值数据时，选择“文件”选项卡上的**编辑**可访问连接字符串对话框，通过此对话框可以自定义传递到提供者的连接字符串以便对连接进行微调。

高级属性

使用 数据收集 节点从需要显式登录的数据库中导入观测值数据时，选择**高级**以提供用户标识和密码来访问数据源。

导入多重响应集

通过对每个可能的变量值使用一个单独的标志字段，多响应变量可作为多二分法集从 数据收集 导入。 例如，如果要求响应者从列表中选择他们已经参观过的博物馆，那么该集合中就会包含与每个列出的博物馆一一对应的单独标志字段。

导入数据后，您可以从包含“过滤器”选项卡的任意节点添加或编辑多重响应集。 有关更多信息，请参阅主题第 114 页的『编辑多重响应集』。

将多个响应导入单个字段（适用于在前发行版中创建的流）

在旧版本的 SPSS Modeler 中，并不是按以上方式导入多个响应，实际上是将它们导入到单独的字段中，并且用逗号分隔值。 为支持现有的流，该方法仍旧适用，但是建议更新所有这种流以使用新的方法。

数据收集列导入说明

数据收集 数据中的列按照在下表中汇总的方式读入 SPSS Modeler。

表 *4:* 数据收集 列名导入摘要

数据收集 列类型	**SPSS Modeler 存储器**	**测量级别**
布尔标志 (yes/no)	String	标记（值为 0 和 1）
分类	String	名义
日期或时间戳	时间戳记	连续
双精度值（指定范围内的浮点值）	实数	连续
长整型（指定范围内的整数值）	整数	连续
文本（自由文本描述）	String	无类型
等级（指示问题内的网格或循环）	不发生在 VDATA 中且不导入到 SPSS Modeler	
对象（二元数据，例如显示不规则文字的传真或声音记录）	不导入到 SPSS Modeler	

表 4: 数据收集 列名导入摘要 (继续)		
数据收集 列类型	**SPSS Modeler 存储器**	**测量级别**
无（未知类型）	不导入到 SPSS Modeler	
Respondent.Serial 列（为每个被调查者关联一个唯一的标识）	整数	无类型

为避免从元数据中读取和从实际值中读取的值标签之间可能出现的不一致现象，可将所有元数据值转换为小写。 例如，可将值标签 *E1720_years* 转换为 *e1720_years* 。

IBM Cognos 源节点

使用 IBM Cognos 源节点可将 Cognos 数据库数据或单列表报告导入到数据挖掘会话中。 这样，就可以将 Cognos 的商业智能功能与 IBM SPSS Modeler 的预测性分析能力融为一体。 您可导入关系数据、维度建模关系 (DMR) 数据和 OLAP 数据。

从 Cognos 服务器连接中，首先选择从中导入数据或报告的位置。 位置包含 Cognos 模型以及所有与该模型关联的文件夹、查询、报告、视图、快捷方式、URL 和作业定义。 Cognos 模型定义业务规则、数据描述、数据关系、业务元素和层次结构以及其他管理任务。

如果要导入数据，那么可以从所选数据包中选择要导入的对象。 可以导入的对象包括查询主体（表示数据库表）或各个查询项（表示表列）。 有关更多信息，请参阅第 30 页的『Cognos 对象图标』。

如果数据包定义了过滤器，那么可以导入一个或多个这些过滤器。 如果导入的过滤器与导入的数据关联，那么该过滤器在导入数据之前应用。 要导入的数据必须为 UTF-8 格式。

如果要导入报告，请选择包含一个或多个报告的数据包或数据包内的文件夹。 接着，可以选择要导入的单个报告。 只能导入单列表报告；不支持导入多个列表。

如果为数据对象或报告定义了参数，那么您可以先指定这些参数的值，然后再导入对象或报告。

注: Cognos 源节点仅支持 Cognos CQM 数据包。 不支持 DQM 数据包。

Cognos 对象图标

可从 Cognos Analytics 数据库导入的各种对象类型以不同的图标表示，如下表所述。

表 5: *Cognos* 对象图标	
图标	**对象**
	打包
	名称空间
	查询对象
	查询项目
	测量维度
	测量
	维度
	层级层次
	级别

表 5: Cognos 对象图标 (继续)	
图标	**对象**
	过滤
	报告
	独立计算

导入 Cognos 数据

要从 IBM Cognos Analytics 数据库导入数据，请确保 IBM Cognos 对话框的“数据”选项卡上的**方式**设置为**数据**。

连接。 单击**编辑**将显示一个对话框，可在其中定义从其导入数据或报告的全新 Cognos 连接的详细信息。 如果已经通过 IBM SPSS Modeler 登录 Cognos 服务器，也可编辑当前连接的明细。 有关更多信息，请参阅第 32 页的『Cognos 连接』。

位置。 建立 Cognos 服务器连接后，单击此字段旁边的**编辑**可显示从中导入内容的可用数据包列表。 有关更多信息，请参阅第 32 页的『Cognos 位置选择』。

内容。 显示所选数据包的名称，以及与该数据包关联的名称空间。 双击名称空间即可显示可以导入的对象。 各种对象类型以不同的图标表示。 有关更多信息，请参阅第 30 页的『Cognos 对象图标』。

要选择所要导入的对象，请选中该对象，并单击两个向右箭头中位于上方的箭头，将该对象移入**要导入的字段**窗格。 选择查询主体会导入它的所有查询项。 双击查询主体会将其展开，以便您选择其中的一个或多个查询项。 通过按住 Ctrl 并单击（选中个别的项）、按住 Shift 并单击（选中项块）以及按 Ctrl-A（选中所有的项），可以执行多选。

要选择所要应用的过滤器（如果已经为数据包定义过滤器），请在“内容”窗格中浏览到过滤器，选中该过滤器，并单击两个向右箭头中位于下方的箭头，将该过滤器移入**要应用的过滤器**窗格。 通过按住 Ctrl 并单击（选中个别过滤器）和按住 Shift 并单击（选中过滤器块），可以执行多选。

要导入的字段。 列出您已选定要导入到 IBM SPSS Modeler 以供处理的数据库对象。 如果不再需要特定对象，请将其选中，并单击向左箭头，将其移回到**内容**窗格中。 您可采用与**内容**相同的方式执行多选。

要应用的过滤器。 列出您已选定要在导入数据前应用于数据的过滤器。 如果不再需要特定过滤器，请将其选中，并单击向左箭头，将其移回到**内容**窗格中。 您可采用与**内容**相同的方式执行多选。

参数。 如果此按钮已启用，那么表明所选对象已定义有参数。 您可以先使用参数进行调整（例如，执行参数化计算），然后再导入数据。 如果已定义参数，但未提供缺省值，那么此按钮会显示警告三角形。 单击此按钮即可显示参数并选择性地进行编辑。 如果此按钮已禁用，那么表明该报告尚未定义任何参数。

在导入之前聚集数据。 如果要导入汇总数据而非原始数据，请选中此框。

导入 Cognos 报告

要从 IBM Cognos 数据库导入预定义报告，请在 IBM Cognos 对话框的“数据”选项卡上，确保**方式**设置为**报告**。 只能导入单列表报告；不支持导入多个列表。

连接。 单击**编辑**将显示一个对话框，可在其中定义从其导入数据或报告的全新 Cognos 连接的详细信息。 如果已经通过 IBM SPSS Modeler 登录 Cognos 服务器，也可编辑当前连接的明细。 有关更多信息，请参阅第 32 页的『Cognos 连接』。

位置。 建立 Cognos 服务器连接后，单击此字段旁边的**编辑**可显示从中导入内容的可用数据包列表。 有关更多信息，请参阅第 32 页的『Cognos 位置选择』。

内容。 显示报告所在的选定数据包或文件夹的名称。 请浏览到特定的报告，将其选中，并单击向右箭头，以将其移动到**要导入的报告**字段中。

要导入的报告。 指示您已选定要导入到 IBM SPSS Modeler 的报告。 如果不再需要该报告，请将其选中，并单击向左箭头以将其移回到**内容**窗格，或者将其他报告移动到此字段中。

参数。 如果此按钮已启用，那么表明所选报告已定义有参数。 您可以先使用参数进行调整（例如，指定报告数据的开始日期和结束日期），然后再导入报告。 如果已定义参数，但未提供缺省值，那么此按钮会显示警告三角形。 单击此按钮即可显示参数并选择性地进行编辑。 如果此按钮已禁用，那么表明该报告尚未定义任何参数。

Cognos 连接

在“Cognos 连接”对话框中，您可以选择在其中导入或导出数据库对象的 Cognos Analytics 服务器。

Cognos 服务器 URL 请输入要在其中进行导入或导出的 Cognos Analytics 服务器的 URL。 这是 Cognos 服务器上 IBM Cognos 配置的“外部分派器 URI”环境属性值。 如果不确定要使用哪个 URL，请与 Cognos 系统管理员联系。

方式 如果您希望使用特定的 Cognos 名称空间、用户名和密码（例如，以管理员身份）登录，请选择**设置凭证**。 选择**使用匿名连接**登录而不使用用户凭证，在这种情况您不需要填写其他字段。

另外，如果您有存储在 IBM SPSS 协作和部署服务 存储库中的 IBM Cognos 凭证，您可使用此凭证，而不必输入用户名和密码信息或创建匿名连接。 要使用现有凭证，请选择**存储凭证**，然后输入**凭证名称**或浏览凭证名称。

Cognos 名称空间由 IBM SPSS 协作和部署服务 中的域建模。

名称空间标识 指定用于登录服务器的 Cognos 安全认证提供程序。 认证服务提供者用于定义和维护用户、组和角色，并控制认证过程。 请注意，这是名称空间标识，而不是名称空间名称（标识并非始终与名称相同）。

用户名 输入用于登录服务器的 Cognos 用户名。

密码 输入与指定用户名关联的密码。

另存为缺省值 单击此按钮可以将这些设置存储为缺失值，以避免在您每次打开节点时重新输入这些设置。

Cognos 位置选择

使用“指定位置”对话框可以选择要从中导入数据的 Cognos 数据包，或要从中导入报告的数据包或文件夹。

公共文件夹。 如果要导入数据，这将列出选定服务器上的可用数据包和文件夹。 选择希望使用的数据包，然后单击**确定**。 对于每个 Cognos 源节点，只能选择一个数据包。

如果要导入报告，这将列出选定服务器上可用的包含报告的文件夹和数据包。 选择数据包或报告文件夹并单击**确定**。 对于每个 Cognos 源节点，您只能选择一个数据包或报告文件夹，但是报告文件夹中可以包含其他报告文件夹和各个报告。

指定数据或报告参数

如果在 Cognos Analytics 中已经为数据对象或报告定义参数，您可以先指定这些参数的值，再导入对象或报告。 报告的参数示例可以是报告内容的开始日期和结束日期。

名称。 在 Cognos 数据库中指定的参数名称。

类型。 参数的描述。

值。 分配给参数的值。 要输入或编辑值，请双击表中该值的单元格。 未在此处对值进行验证，因此，运行时将检测到任何无效值。

自动从表中除去无效参数。 缺省选择此选项，将会自动删除任何在数据对象或报告中找到的无效参数。

IBM Cognos TM1 源节点

通过 IBM Cognos TM1 源节点，您可以将 Cognos TM1 数据引入数据挖掘会话中。 这样，可将 Cognos 的企业规划功能与 IBM SPSS Modeler 的预测性分析功能结合起来。 可以导入多维 OLAP 多维数据集数据的序列化版本。

注: TM1 用户需要下列许可权：对多维数据集的写特权、对维度的读特权以及对维度元素的写特权。 此外，需要 IBM Cognos TM1 10.2 FP3 或更高版本，SPSS Modeler 才能导入和导出 Cognos TM1 数据。 基于先前版本的现有流仍能够正常运行。

该节点不需要管理员凭证。 但如果仍要使用 17.1 之前的旧的遗存 TM1 节点，仍需要管理员凭证。

SPSS Modeler 仅支持通过 IntegratedSecurityMode 1、4 和 5 使用 Cognos TM1。

您需要先修改 TM1 中的数据，然后再导入数据。 要导入的数据必须为 UTF-8 格式。

从 IBM Cognos TM1 管理主机连接，先选择要从中导入数据的 TM1 服务器；一台服务器包含一个或多个 TM1 多维数据集。 然后，选择所需多维数据集，并在该多维数据集内选择要导入的列和行。

注: 必须先验证 `tm1s.cfg` 文件中的某些设置，然后才能在 SPSS Modeler 中使用 TM1“源”或“导出”节点；此文件是 TM1 服务器根目录中的 TM1 服务器配置文件。

- HTTPPortNumber - 设置有效的端口号；通常介于 1 到 65535 之间。 请注意，这不是后续在节点中的连接内指定的端口号；这是 TM1 使用的内部端口，缺省情况下处于禁用状态。 如果需要，请联系 TM1 管理员以确认该端口的有效设置。
- UseSSL - 如果将此项设置为 *True*，那么将使用 HTTPS 作为传输协议。 在这种情况下，您必须将 TM1 证书导入 SPSS Modeler Server JRE。

导入 IBM Cognos TM1 数据

要从 IBM Cognos TM1 数据库导入数据，请在 IBM Cognos TM1 对话框的**数据**选项卡上指定服务器连接详细信息，并选择多维数据集和数据详细信息。

注: 必须在 TM1 内执行一些处理操作以确保数据格式可由 IBM SPSS Modeler 识别，然后再导入数据。 这包括使用“子集编辑器”来过滤数据，以便将视图调整为适合于导入的正确大小和形状。

请注意，从 TM1 中导入的零 (0) 值将被视为“null”值（TM1 不区分空白值和零值）。 并且，请注意，来自常规维度的非数字数据（或元数据）可以导入到 IBM SPSS Modeler 中。 但是，目前不支持导入非数字度量。

连接类型。 选择**管理服务器**或 **TM1 服务器**。 请注意，已从 Planning Analytics on Cloud 中移除了管理服务器，因此如果您有连接到旧管理服务器的旧流，那么可以转而将其修改为指向 Planning Analytics on Cloud。 如果在此处选择**管理服务器**，那么必须输入服务器的 URL（REST API 的主机名）和服务器的名称。 如果选择 **TM1 服务器**，请继续到以下部分。

TM1 服务器 URL。 输入要连接的 TM1 服务器将要安装到的管理主机 URL。 管理主机定义为所有 TM1 服务器的单个 URL。 通过此 URL 可以发现并访问环境中安装并运行的所有 IBM Cognos TM1 服务器。 单击**登录**。 如果先前未连接值此服务器，那么系统将提示您输入**用户名**和**密码**；另外，您可以搜索先前输入的已保存为**已存储的凭证**的登录详细信息。

选择要导入的 TM1 立方体视图。 显示可以从中导入数据的 TM1 服务器内多维数据集的名称。 双击多维数据集，以显示可以导入的视图数据。

注:

只有具有维度的多维数据集才能导入到 IBM SPSS Modeler 中。

如果已为 TM1 多维数据集中的元素定义别名（例如，如果值 23277 具有别名 `Sales`），那么将导入该值而不是别名。

为了选择要导入的数据，请选中视图，然后单击右箭头，从而将该视图移至**要导入的视图**字段中。 如果未显示所需视图，请双击多维数据集以展开其视图列表。 可以选择公共或私有视图。

行维度 。 列出您已选择导入的数据中行维度的名称。 滚动级别列表并选择所需级别。

列维度。 列出您已选择导入的数据中列维度的名称。 滚动级别列表并选择所需级别。

上下文维度。 仅用于显示。 显示与所选列和行相关的上下文维度。

TWC 源节点

TWC 源节点从 The Weather Company, an IBM Business 导入天气数据。 您可以使用它来获取某一位置的历史或预测天气。 这可以帮助制定天气驱动型业务解决方案，以使用可用的最准确天气数据做出更明智的决定。

通过该节点，可以输入天气相关的数据，例如，latitude、longitude、time、day_ind（指示夜间或白天）、temp、dewpt（露点）、rh（相对湿度）、feels_like 气温、heat_index、wc（风寒）、wx_phrase（多云、局部多云等）、pressure、clds（云）、vis（能见度）、wspd（风速）、gust、wdir（风向）、wdir_cardinal（NW、NNW、N 等）、uv_index（紫外线指数）和 uv_desc（低、高等）。

TWC 源节点使用以下 API：

- TWC Historical Observations Airport (http://goo.gl/DplOKj)，用于获取历史天气数据
- TWC Hourly Forecast (http://goo.gl/IJhhvZ)，用于获取预测天气数据

位置

纬度。 以格式 [-90.0~90.0] 输入想要获取其天气数据的位置的纬度值。

经度。 以格式 [-180.0~180.0] 输入想要获取其天气数据的位置的经度值。

其他

许可证密钥。 需要许可证密钥。 输入您从 The Weather Company 获取的许可证密钥。 如果您没有密钥，请联系管理员或 IBM 代表。

您的管理员可能在 IBM SPSS Modeler Server 上的新 config.cfg 文件中指定了密钥，而不是向所有用户发放密钥，在这种情况下，您可以将此字段留空。 如果在两个位置中都进行了指定，该对话框中的密钥具有优先权。 管理员注意：要在服务器上添加许可证密钥，请创建名为 config.cfg 的新文件，其中的内容是 LicenseKey=<LICENCEKEY>（其中，<LICENSEKEY> 是许可证密钥），该文件位置为 <ModelerServerInstallation>\ext\bin\pasw.twcdata。

单位。 选择要使用的度量单位：**English**、**Metric** 或 **Hybrid**。 缺省值为 **Metric**。

时间格式

UTC。 如果要导入历史天气数据，而且您不想让 SPSS Modeler 访问 TWC Hourly Forecast API，请选择 UTC 时间格式。 选择此选项时，您的许可证密钥只需有权访问 TWC Historical Observations Airport API。

本地。 如果需要让 SPSS Modeler 访问 TWC Hourly Forecast API，以将时间从 UTC 时间转换为本地时间，请选择“本地”时间格式。 选择此选项时，您的许可证密钥必须有权访问这两个 TWC API。

数据类型

历史。 如果想要导入历史天气数据，请选择**历史**，然后指定 YYYYMMDD 格式的开始和结束日期（例如，20120101 表示 2012 年 1 月 1 日）。

预测。 如果想要导入预测天气数据，请选择**预测**，然后指定要预测的小时数。

SAS 源节点

SPSS Modeler Professional 和 SPSS Modeler Premium 中提供了此功能。

使用 SAS 源节点可以将 SAS 数据导入数据挖掘会话中。 可以导入以下四种类型文件：

- 适用于 Windows/OS2 的 SAS (*.sd2*)
- 适用于 UNIX 的 SAS (*.ssd*)
- SAS 传输文件 (*.tpt*)
- SAS 版本 7/8/9 (*.sas7bdat*)

导入数据时，将保留所有变量，并且不会更改任何变量类型。 将选定所有观测值。

为 SAS 源节点设置选项

IMPORT。 选择要传输的 SAS 文件的类型。 选择 **适用于 Windows/OS2 的 SAS (.sd2)** 、 **适用于 UNIX 的 SAS (.SSD)** 、 **SAS 传输文件 (.tpt)** 或 **SAS 版本 7/8/9 (.sas7bdat)** 。

导入文件。 指定文件的名称。 可以输入文件名或单击省略号按钮 (**...**) 来浏览到文件所在位置。

成员。 从上面选定的 SAS 传输文件中选择要导入的成员。 可以输入成员名或单击 **选择** 浏览文件中的所有成员。

从 SAS 数据文件读取用户格式。 选中此选项以读取用户格式。 SAS 文件将数据和数据格式（例如变量标签）存储在不同的文件中。 在很多时候，可能希望同时导入格式。 不过，如果拥有的数据集较大，则可能希望取消选中此选项以节省内存。

格式文件。 如果需要格式化文件，请激活此文本框。 可以输入文件名或单击省略号按钮 (**...**) 来浏览到文件所在位置。

变量名称。 选择从 SAS 文件中导入时所使用的处理变量名称和标签的方法。 选择在此处包括的元数据会保留在 IBM SPSS Modeler 的整个过程中，并且可以再次导出以在 SAS 中使用。

- **读取名称和标签**。 选中此选项将变量名称和标签同时读入 IBM SPSS Modeler。 缺省情况下将选中此选项，并且变量名称将显示在“类型”节点中。 根据流属性对话框中指定的选项，标签将显示在表达式构建器、图表、模型浏览器和其他类型的输出中。
- **读取用作名称的标签**。 选择从 SAS 文件中读取说明性的变量标签而不是短字段名，并将这些标签作为变量名称在 IBM SPSS Modeler 中使用。

Excel 源节点

Excel 源节点允许您从 Microsoft Excel 导入 `.xlsx` 文件格式的数据。

文件类型。 选择要导入的 Excel 文件类型。

导入文件。 指定要导入的电子表格文件的名称和位置。

使用指定范围。 选中此选项可以指定在 Excel 工作表中定义的单元格的指定范围。 单击省略按钮 (**...**) 从可用范围列表中进行选择。 如果使用指定范围，则其他工作表和数据范围设置将不再可用并最终被禁用。

选择工作表。 按索引或者按名称指定要导入的工作表。

- **按索引**。 指定要导入的工作表的索引值，开头的 0 表示第一个工作表，1 表示第二个工作表，依此类推。
- 按名称。 指定要导入的工作表的名称。 单击省略按钮 (**...**) 从可用工作表列表中进行选择。

工作表上的范围。 可以第一个非空行作为开始导入数据，也可通过单元格的显式范围导入数据。

- **范围从第一个非空行开始**。 找到第一个非空单元格，并将此单元格作为数据范围的左上角单元格。
- **单元格的显式范围**。 选中此选项可按行和列指定显式范围。 例如，要指定 Excel 范围 `A1:D5`，您可以在第一个字段中输入 `A1`，在第二个字段中输入 `D5`，（或，`R1C1` 和 `R5C4`）。 指定范围内的所有行都将返回，包括空行。

在空白行上。 如果遇到多个空行，则可选择**停止读取**，或选择**返回空行**以继续读取所有数据（包括空行）直到工作表的末尾。

第一行具有列名。 表示指定范围中的第一行应作为字段（列）名使用。 如果未选中此选项，那么将自动生成字段名称。

列和类型的扫描行数。 如果希望 IBM SPSS Modeler 扫描更多 Excel 数据行以确定列类型和存储类型，那么可以增大此值。 缺省值是 200 行。 请注意，此设置可能会影响性能。

字段存储和测量级别

从 Excel 中读取值时，缺省情况下将按连续的测量级别读取以数值存储的字段，按名义读取以字符串存储的字段。 可以在“类型”选项卡上手动更改测量级别（连续和名义），但存储类型是自动确定的（虽然必要时可

在过滤节点或导出节点中使用转换函数，例如 to_integer，来更改此类型）。 有关更多信息，请参阅主题 第 6 页的『设置字段存储类型和格式』。

缺省情况下，将按数字类型读取以数字和字符串值混合存储的字段，这意味着在 IBM SPSS Modeler 中所有字符串值都将被设置为空（系统缺失）值。 这是因为与 Excel 不同，IBM SPSS Modeler 不允许字段中存在混合的存储类型。 为避免这种情况，您可以在 Excel 电子表格中手动将单元格格式设置为 Text，这会导致所有值（包括数字）都作为字符串读入。

XML 源节点

SPSS Modeler Professional 和 SPSS Modeler Premium 中提供了此功能。

使用 XML 源节点可将 XML 格式文件中的数据导入到 IBM SPSS Modeler 流中。 XML 是用于数据交换的标准语言，对许多组织而言，它是用于此目的所选择的格式。 例如，政府税务机构可能希望分析在线提交的纳税申报单中的数据，并且这些申报单的数据处于 XML 格式（请参阅 http://www.w3.org/standards/xml/）。

通过将 XML 数据导入 IBM SPSS Modeler 流，允许您针对数据源执行多种预测分析功能。 系统将 XML 数据解析为表格格式，在此格式中，列对应于 XML 元素和属性的不同嵌套级别。 XML 项目以 XPath 格式进行显示（参阅 http://www.w3.org/TR/xpath20/）。

要点: “XML 源”节点不会考虑名称空间声明。 因此，（例如）您的 XML 文件不能在 name 标记中包含冒号 (:) 字符。 如果包含，在执行期间将会收到与无效字符有关的错误。

读取单个文件。 缺省情况下，SPSS Modeler 读取您在 **XML 数据源**字段中指定的单个文件。

读取目录中的所有 XML 文件。 如果您要读取某个特定目录中的所有 XML 文件，请选择此选项。 在显示的**目录**字段中指定位置。 选择**包含子目录**复选框，以另外读取指定目录的所有子目录中的 XML 文件。

XML 数据源。 输入您要导入的 XML 源文件的完整路径和文件名，或使用“浏览”按钮查找文件。

XML 模式。 （可选）指定您要从中读取 XML 结构的 XSD 或 DTD 文件的完整路径和文件名，或使用“浏览”按钮查找此文件。 如果您保留此字段为空，那么将从 XML 源文件中读取结构。 XSD 或 DTD 文件可以具有多个根元素。 在这种情况下，将光标移动到另一字段时，将显示一个对话框，您可以在此对话框中选择要使用的根元素。 有关更多信息，请参阅主题 第 36 页的『从多个根元素中选择』。

注: SPSS Modeler 将忽略 XSD 指示符

XML 结构。 显示 XML 源文件（或架构，如果您在 **XML 模式**字段中进行了指定）结构的层次结构树。 要定义记录边界，选择某个元素，并单击右方向方向按钮将此项目复制到**记录**字段。

显示属性。 在 **XML 结构**字段中显示或隐藏 XML 元素的属性。

记录（XPath 表达式）。 显示从 XML 结构字段复制的元素的 XPath 语法。 随后，此元素将在 XML 结构中突出显示，并定义记录边界。 每次在源文件中遇到此元素时，都将创建新的记录。 如果此字段为空，那么将使用根元素下的第一个子元素作为记录边界。

读取所有数据。 缺省情况下，源文件中的所有数据都将读取到流中。

指定要读取的数据。 如果您要导入单独元素和/或属性，请选择此选项。 选择此选项将启用“字段”表，您可以在该表中指定要导入的数据。

字段。 如果您选择了**指定要读取的数据**选项，此表将列出选择用于导入的元素与属性。 可以直接在 XPath 列中输入元素或属性的 XPath 语法，也可以在 XML 结构中选择元素或属性，然后单击右方向按钮将其复制到表中。 要复制元素的所有子元素和属性，请在 XML 结构中选中此元素并单击双向方向按钮。

- **XPath**。 要导入的项目的 XPath 语法。
- **位置**。 要导入的项目在 XML 结构中的位置。 **固定路径**显示相对于在 XML 结构中突出显示的元素（或者如果没有突出显示的元素，则为根下面的第一个子元素）的项目路径。 **任何位置**表示在 XML 结构中任何位置上给定名称的项目。 如果您在 XPath 列中直接输入位置，则显示**自定义**。

从多个根元素中选择

格式正确的 XML 文件仅能具有单一根元素，而 XSD 或 DTD 文件可包含多个根元素。 如果其中一个根元素与 XML 源文件中的根元素匹配，那么将使用此根元素，否则，您需要选择一个要使用的根元素。

选择要显示的根。 选择要使用的根元素。 缺省根元素是 XSD 或 DTD 结构中的第一个根元素。

从 XML 源数据中移除不需要的空格

XML 源数据中的换行符可以通过 [CR][LF] 字符组合实现。 在某些情况下，这些换行符可以出现在文本字符串中间，例如：

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

在某些程序（例如，Web 浏览器）中打开此文件时，这些换行符可能不可见。 但是，通过 XML 源节点将数据读取到流后，换行符将转换为一系列空格字符。

可以使用“过滤”节点移除这些不需要的空格来进行更正：

下面是如何实现此操作的示例：

1. 将“填充”节点附加到 XML 源节点。
2. 打开“填充”节点并使用字段选择器选择带有不需要空格的字段。
3. 将**替换**设置为**根据以下条件**，并将**条件**设置为 **true**。
4. 在**替换为**字段中，输入 `replace(" ","",@FIELD)` 并单击“确定”。
5. 将“表”节点附加到“过滤”节点，然后运行流。

在“表”节点的输出中，此时显示的文本不包含额外的空格。

用户输入节点

“用户输入”节点提供了创建综合数据的简便方式 - 从头开始创建综合数据，或通过更改现有的数据创建综合数据。 此节点非常有用，例如，在希望为建模创建测试数据集时，即可使用此节点。

从零开始创建数据

可在源选项板中找到用户输入节点，并将此节点直接添加到流工作区中。

1. 单击节点选项板的 **源** 选项卡。
2. 使用拖放或双击操作将用户输入节点添加到流工作区中。
3. 双击以打开此节点的对话框并指定字段和值。

注意：在“源”选项板中选择的“用户输入”节点将是完全空白的，不含字段也不含任何数据信息。 您可以完全从头开始创建综合数据。

从现有的数据源生成数据

还可以从流的任何非终端节点生成用户输入节点：

1. 确定要在流的哪个点上替换节点。
2. 右键单击可将其数据提供给“用户输入”节点的节点，然后从菜单中选择**生成用户输入节点**。
3. 此时将出现“用户输入”节点，其所有下游过程都将附加到此节点上，从而可取代数据流的该点上现有的节点。 此节点生成时可从元数据中继承所有的数据结构和字段类型信息（如果可用）。

注意：如果数据尚未在流的所有节点中从头到尾地运行，那么这些节点未完全实例化，这表示当使用“用户输入”节点替换原来的节点时，存储类型和数据值可能不可用。

为用户输入节点设置选项

通过使用“用户输入”节点对话框中包含的几个工具，可为综合数据输入值并定义数据结构。 对于生成的节点，“数据”选项卡上的表包含来自原始数据源的字段名称。 对于从“源”选用板中添加的节点，该表是空的。 通过使用表中的选项可执行下列任务：

- 使用表右侧的“添加新字段”按钮添加新的字段。
- 重命名现有字段。
- 为每个字段指定数据存储类型。
- 指定值。

- 更改字段显示的顺序。

输入数据

可使用表右侧的值选取器按钮从原始数据集中为每个字段指定值或插入值。 有关指定值的详细信息，请参阅下面说明的规则。 您也可以选择将字段留空 - 留空的字段会使用系统 null 值 ($null$) 进行填充。

要指定字符串值，仅需要在“值”列中输入这些值并以空格进行分隔：

```
Fred Ethel Martin
```

含有空格的字符串可以用双引号括起来：

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

对于数字字段，可以按照同样的方式（以空格作为间隔列出）输入多个值：

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

也可以通过设置上述值序列的界限 (10, 20) 及其间隔值 (2) 来指定相同的值序列。 使用此方法，可键入：

```
10,20,2
```

这两种方法也可以通过相互嵌套而组合使用，例如：

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

此输入将生成下列值：

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

使用在“流属性”对话框中选定的当前缺省格式输入日期值和时间值。

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

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

timestamp 值既包含日期组件又包含时间组件，所以必须对其使用双引号：

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

有关其他详细信息，请参阅下面数据存储的相关注释。

生成数据。 通过此选项可指定运行流时生成记录的方式。

- **所有组合**。 生成包含字段值的各种可能组合的记录，此时每个字段值将出现在几个记录中。 选中此选项有时可使生成的数据比希望生成的更多，所以通常可能要在此节点后附加一个样本节点。
- **依照顺序**。 按指定的数据字段值的顺序生成记录。 每个字段值仅出现在一个记录中。 记录的总数与单个字段值的最大数相等。 如果字段包含的记录数小于最大记录数，则插入未定义的 ($null$) 值。

显示示例

例如，下列条目将生成以下两个表示例中列出的记录。

- **年龄**。 30,60,10
- **血压** 低
- **胆固醇** 正常高值
- **药物**。 （留空）

表 *6:* 生成设置为“所有组合”的数据字段

年龄	BP	Cholesterol	Drug
30	LOW	NORMAL	$null$
30	LOW	HIGH	$null$
40	LOW	NORMAL	$null$
40	LOW	HIGH	$null$
50	LOW	NORMAL	$null$
50	LOW	HIGH	$null$
60	LOW	NORMAL	$null$
60	LOW	HIGH	$null$

表 *7:* 生成设置为“依照顺序”的数据字段

年龄	BP	Cholesterol	Drug
30	LOW	NORMAL	$null$
40	$null$	HIGH	$null$
50	$null$	$null$	$null$
60	$null$	$null$	$null$

数据存储

存储格式描述数据在某个字段中的存储方式。 例如，值为 1 和 0 的字段存储整型数据。 这点与测量级别明显不同，测量级别描述的是数据的使用方法，而且不影响存储。 例如，您可能希望将值为 1 和 0 的某个整数字段的测量级别设置为标志。 这通常表明 1 = 真 ，0 = 假 。 存储格式必须在数据源中确定，而测量级别可以使用“类型”节点在流中的任意点上进行更改。 有关更多信息，请参阅主题 第 104 页的『测量级别』。

可用存储类型有：

- **字符串** 用于包含非数字数据（也称为字母数字数据）的字段。 字符串可以包含任何字符序列，比如 *fred* 、 *Class 2* 或 *1234* 。 注意：字符串中的数字不能用于计算。
- **整数** 值为整数的字段。
- **实数** 值为可能包含小数（不限于整数）的数字。 显示格式在“流属性”对话框中指定，并且可以被“类型”节点（“格式”选项卡）中的各个字段覆盖。
- **日期** 以标准格式指定的日期值，例如年月日（例如 2007-09-26）。 具体格式在“流属性”对话框中指定。
- **时间** 以持续时间形式测量的时间。 例如，某个服务电话持续 1 小时 26 分 38 秒，该时间可以根据“流属性”对话框中指定的当前时间格式表示为：01:26:38。
- **时间戳记** 同时包含日期和时间组成部分的值，例如 2007-09-26 09:04:00，具体同样取决于“流属性”对话框中的当前日期和时间格式。 请注意，需要用双引号将时间戳值括起来，以确保将此值解释为单一值而非单独的日期和时间值。 （同样适用于在用户输入节点中输入值时的情况。）
- **列表** 在 SPSS Modeler V17 中，随新测量级别“地理空间”和“集合”一起引入了“列表”存储字段，对于单个记录，此字段包含多个值。 存在所有其他存储类型的列表版本。

表 *8:* 列表存储类型图标

图标	存储类型
[]	字符串列表
[]	整数列表

表 8: 列表存储类型图标 (继续)	
图标	**存储类型**
[]	实数列表
[]	时间列表
[]	日期列表
[]	时间戳记列表
[[]]	深度大于零的列表

另外，为了与“集合”测量级别配合使用，提供了下列测量级别的列表版本。

表 9: 列表测量级别图标	
图标	**测量级别**
[]	连续值列表
[]	分类值列表
[]	标志列表
[]	名义值列表
[]	有序值列表

可以通过三个源节点（“Analytic Server”、“地理空间”或“变量文件”）中的某一个将列表导入到 SPSS Modeler 中，也可以在流中使用“派生”或“填充”字段操作节点创建列表。

有关“列表”及其与“集合”和“地理空间”测量级别的交互的更多信息，请参阅第 8 页的『列表存储以及相关联的测量级别』。

存储转换。 用户可使用各种转换函数来转换某个字段的存储格式，比如“填充”节点中的 `to_string` 和 `to_integer`。 有关更多信息，请参阅主题 第 120 页的『使用填充节点进行存储类型转换』。 请注意，转换函数（以及需要特定类型输入（如日期或时间值）的任何其他函数）取决于“流属性”对话框中指定的当前格式。 例如，如果要将值为 *Jan 2018*、*Feb 2018* 等等的字符串字段转换为日期存储格式，请选择 **MON YYYY** 作为流的缺省日期格式。 “派生”节点中也有可用的转换函数，用于派生计算过程中的临时转换。 另外，还可以使用“派生”节点来执行其他操作，比如使用分类值对字符串字段进行重新编码。 有关更多信息，请参阅主题 第 120 页的『使用导出节点对值进行重新编码』。

正在读入混合数据。 请注意，读取数字存储格式（整数、实数、时间、时间戳或日期）的字段中的数据时，任何非数字值将被设置为空或系统缺失。 这时因为 IBM SPSS Modeler 与某些应用程序不同，它不允许字段中含有混合存储类型。 为了避免出现混合存储类型，必须根据需要更改源节点中或外部应用程序中的存储类型，从而将任何具有混合数据的字段以字符串的格式读入。

注意：在生成的“用户输入”节点中可能已包含了从源节点（如果已实例化）获取的存储类型信息。 未实例化的节点不包含存储类型信息或使用类型信息。

用于指定值的规则

对于符号字段，应在多个值之间保留空格，例如：

```
HIGH MEDIUM LOW
```

对于数字字段，可以按照同样的方式（以空格作为间隔列出）输入多个值：

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

也可以通过设置上述值序列的界限 (10, 20) 及其间隔值 (2) 来指定相同的值序列。 使用此方法，可键入：

```
10,20,2
```

这两种方法也可以通过相互嵌套而组合使用，例如：

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

此输入将生成下列值：

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

“模拟生成”节点

“模拟生成”节点提供了一种生成模拟数据的简单方法，即在没有历史数据的情况下，使用用户指定的统计分布生成数据；或者使用对现有历史数据运行“模拟拟合”节点而获取的分布自动生成数据。 如果模型输入中存在不确定性，那么在您要对预测模型的结果进行评估时，生成模拟数据非常有用。

在没有历史数据的情况下创建数据

在“源”选用板中找到“模拟生成”节点，并且可以将其直接添加到流画布中。

1. 单击节点选项板的 **源** 选项卡。
2. 使用拖放或双击操作将“模拟生成”节点添加到流画布中。
3. 双击以打开此节点的对话框，并指定字段、存储类型、统计分布和分布参数。

注：在“源”选用板中选择的“模拟生成”节点将是完全空白的，不包含任何字段和任何分布信息。 这使您能够在完全没有历史数据的情况下创建模拟数据。

使用现有历史数据创建模拟数据

还可以通过执行“模拟拟合”终端节点来创建“模拟生成”节点：

1. 右键单击“模拟拟合”节点，然后从菜单中选择**运行**。
2. “模拟生成”节点将显示在包含指向“模拟拟合”节点的更新链接的流画布中。
3. 生成“模拟生成”节点后，此节点将从“模拟拟合”节点中继承所有字段、存储类型和统计分布信息。

定义指向“模拟拟合”节点的更新链接

您可以在“模拟生成”节点与“模拟拟合”节点之间创建一个链接。 如果要使用最佳拟合分布（由到历史数据的拟合确定）的信息更新一个或多个字段，那么此链接很有用。

1. 右键单击“模拟生成”节点。
2. 从菜单中选择**定义更新链接**。 光标将更改为链接光标。
3. 单击其他节点。 如果此节点是“模拟拟合”节点，那么将建立链接。 如果此节点不是“模拟拟合”节点，那么不会建立链接，并且光标将更改回正常光标。

如果“模拟拟合”节点中的字段不同于“模拟生成”节点中的字段，那么将显示一条消息，告知您存在差异。

使用“模拟拟合”节点更新已链接的“模拟生成”节点时，结果取决于是否在这两个节点中显示了相同的字段，以及是否在“模拟生成”节点中解锁了这些字段。 下表显示了“模拟拟合”节点的更新结果。

表 *10:* “模拟拟合”节点的更新结果

	“模拟拟合”节点 中的字段	
“模拟生成”节点中的字段	**存在**	**缺失**
存在并已解锁。	已覆盖字段。	已删除字段。
缺失。	已添加字段。	未进行更改。
存在并已锁定。	未覆盖字段的分布。 **“拟合详细信息”**对话框中的信息以及相关性已更新。	未覆盖字段。 相关性设置为 0。
已选中**请不要在重新拟合时清除最小值/最大值**复选框。	已覆盖字段，	“最小值”和“最大值”列中的值除外。
已选中**请不要在重新拟合时重新计算相关性**复选框。	如果字段已解锁，那么将覆盖该字段。	未覆盖相关性。

移除指向“模拟拟合”节点的更新链接

通过完成下列步骤，可以移除“模拟生成”节点与“模拟拟合”节点之间的链接：

1. 右键单击“模拟生成”节点。
2. 从菜单中选择**移除更新链接**。 将移除此链接。

为“模拟生成”节点设置选项

使用“模拟生成”节点对话框的“数据”选项卡上的选项可以执行下列操作：

- 查看、指定和编辑字段的统计分布信息。
- 查看、指定和编辑字段之间的相关性。
- 指定要模拟的迭代数和观测值数。

选择一个项目。 使您可以在“模拟生成”节点的以下三个视图之间进行切换：模拟字段、相关性和高级选项。

“模拟字段”视图

如果使用历史数据根据“模拟拟合”节点生成或更新了“模拟生成”节点，那么您可以在“模拟字段”视图中查看并编辑每个字段的统计分布信息。 以下有关每个字段的信息将从“模拟拟合”节点复制到“模拟生成”节点的**类型**选项卡中：

- 测量级别
- 值
- 丢失
- 检查(E)
- 角色

如果您没有历史数据，那么可以通过选择存储类型，选择分布类型并输入必需参数来定义字段并指定这些字段的分布。 以这种方式生成数据意味着，在对数据进行实例化之前（例如，在**类型**选项卡或“类型”节点中对数据进行实例化），有关每个字段的测量级别的信息将不可用。

“模拟字段”视图包含一些工具，您可以使用这些工具来执行下列任务：

- 添加和移除字段。
- 更改字段显示的顺序。
- 为每个字段指定存储类型。
- 为每个字段指定统计分布。
- 为每个字段的统计分布指定参数值。

模拟字段。 如果已将“模拟生成”节点从“源”选用板添加到流工作区中，那么此表将包含一个空行。 编辑此行时，将在表的底部添加一个新的空行。 如果已根据“模拟拟合”节点创建了“模拟生成”节点，那么此表将为历史数据的每个字段提供一行。 通过单击**添加新字段**图标，您可以向表中添加额外的行。

“模拟字段”表包含以下列：

- **字段**。 包含字段的名称。 可以通过在单元格中输入内容来编辑字段名称。
- **存储器**。 此列中的单元格包含存储类型的下拉列表。 可用的存储类型为**字符串**、**整数**、**实数**、**时间**、**日期**和**时间戳记**。 存储类型的选择确定了“分布”列中可用的分布。 如果已根据“模拟拟合”节点创建了“模拟生成”节点，那么将从“模拟拟合”节点复制存储类型。

  注: 对于存储类型为日期时间的字段，必须将分布参数指定为整数。 例如，要指定平均日期 1970 年 1 月 1 日，请使用整数 0。 带符号整数表示自 1970 年 1 月 1 日午夜（或之前）以来经过的秒数。
- **状态**。 “状态”列中的图标指示每个字段的拟合状态。

未对字段指定分布或者缺少一个或多个分布参数。 要运行模拟，您必须为此字段指定分布并输入参数的有效值。

字段设置为最接近的拟合分布。

注: 只有在根据“模拟拟合”节点创建了“模拟生成”节点的情况下，才会显示此图标。

已将最接近的拟合分布替换为“**拟合详细信息**”子对话框中的替代分布。 有关更多信息，请参阅主题 第 46 页的『拟合详细信息』。

已手动指定或编辑了分布，该分布可能包含在多个级别指定的参数。

- **已锁定**。 通过选中带有锁定图标的列中的复选框来锁定模拟字段，可以阻止已链接的“模拟拟合”节点自动更新该字段。 如果您手动指定分布并希望确保在执行已链接的“模拟拟合”节点时自动分布拟合不会影响该分布，那么此列非常有用。
- **分布**。 此列中的单元格包含统计分布的下拉列表。 存储类型的选择决定了指定字段的此列中可用的分布。 有关更多信息，请参阅主题 第 48 页的『分布』。

  注: 无法为每个字段指定固定分布。 如果希望生成的数据中每个字段都是固定字段，那么您可以使用后跟“平衡”节点的“用户输入”节点。
- **参数**。 此列中显示与每个已拟合的分布关联的分布参数。 参数的多个值之间以逗号分隔。 为参数指定多个值将为模拟生成多个迭代。 有关更多信息，请参阅主题 第 48 页的『迭代』。 如果缺少参数，那么“状态”列中显示的图标将反映此情况。 要为参数指定值，请在对应于相关字段的行中单击此列，然后从列表中选择**指定**。 这将打开“**指定参数**”子对话框。 有关更多信息，请参阅主题 第 47 页的『指定参数』。 如果在“分布”列中选择了“经验”，那么此列将处于禁用状态。
- **最小值、最大值**。 对于某些分布，您可以在此列中指定模拟数据的最小值和/或最大值。 小于最小值以及大于最大值的模拟数据将被拒绝，即使这些数据对于指定分布有效也是如此。 要指定最小值和最大值，请在对应于相关字段的行中单击此列，然后从列表中选择**指定**。 这将打开“**指定参数**”子对话框。 有关更多信息，请参阅主题 第 47 页的『指定参数』。 如果在“分布”列中选择了“经验”，那么此列将处于禁用状态。

使用最接近的拟合。 只有在已使用历史数据根据“模拟拟合”节点自动创建了“模拟生成”节点，并且在“模拟字段”表中选择了单个行的情况下，才会启用此按钮。 此按钮用于将所选行中字段的信息替换为该字段的最佳拟合分布信息。 如果对所选行中的信息进行了编辑，那么单击此按钮会将信息重置为根据“模拟拟合”节点确定的最佳拟合分布。

拟合详细信息。 只有在根据“模拟拟合”节点自动创建了“模拟生成”节点的情况下，才会启用此列。 此列用于打开“**拟合详细信息**”子对话框。 有关更多信息，请参阅主题 第 46 页的『拟合详细信息』。

使用“模拟字段”视图右侧的图标可以执行一些有用的任务。 下表描述了这些图标。

表 *11:* “模拟字段”视图上的图标

图标	工具提示	描述
	**编辑分布参数**	只有在“模拟字段”表中选择了单个行的情况下，才会启用此图标。 用于打开所选行的**“指定参数”**子对话框。 有关更多信息，请参阅主题 第 47 页的『指定参数』。
	**添加新字段**	只有在“模拟字段”表中选择了单个行的情况下，才会启用此图标。 用于向“模拟字段”表的底部添加新的空行。
	**创建多个副本**	只有在“模拟字段”表中选择了单个行的情况下，才会启用此图标。 用于打开**“克隆字段”**子对话框。 有关更多信息，请参阅主题 第 46 页的『克隆字段』。
	**删除所选字段**	用于从“模拟字段”表中删除所选行。
	**移动到顶部**	只有在所选行不在“模拟字段”表的顶部时，才会启用此图标。 用于将所选行移动到“模拟字段”表的顶部。 此操作将影响模拟数据中字段的顺序。
	**向上移动**	只有在所选行不在“模拟字段”表的顶部时，才会启用此图标。 用于在“模拟字段”表中将所选行上移一个位置。 此操作将影响模拟数据中字段的顺序。
	**向下移动**	只有在所选行不在“模拟字段”表的底部时，才会启用此图标。 用于在“模拟字段”表中将所选行下移一个位置。 此操作将影响模拟数据中字段的顺序。
	**移动到底部**	只有在所选行不在“模拟字段”表的底部时，才会启用此图标。 用于将所选行移动到“模拟字段”表的底部。 此操作将影响模拟数据中字段的顺序。

在重新拟合时不清除最小值和最大值。 选择此图标后，通过执行已连接的“模拟拟合”节点来更新分布时，将不会覆盖最小值和最大值。

“相关性”视图

我们知道，预测模型的输入字段通常具有相关性 - 例如，身高和体重。 必须考虑将模拟的字段之间的相关性，以确保模拟值保留这些相关性。

如果使用历史数据根据“模拟拟合”节点生成或更新了“模拟生成”节点，那么您可以在“相关性”视图中查看并编辑字段对之间的已计算相关性。 如果您没有历史数据，那么可以根据您对字段相关方式的了解来手动指定相关性。

注: 在生成任何数据之前，将自动检查相关性矩阵以确认它是否为半正定矩阵，并且是否能因此进行反转。 如果矩阵的列线性独立，那么可以将该矩阵反转。 如果无法反转相关性矩阵，那么它将自动进行调整以实现可反转。

您可以选择以矩阵格式或列表格式显示相关性。

相关性矩阵。 用于显示矩阵中字段对之间的相关性。 字段名称按字母顺序列出，从矩阵的左上方向下列出。 只能编辑对角线下方的单元格；必须输入介于 -1.000（含本数）到 1.000（含本数）之间的值。 对角线上方的单元格将在焦点离开对角线下方该单元格的镜像单元格时进行更新；这两个单元格将显示同一值。 对角线单元格始终处于禁用状态，并始终具有相关性 1.000。 所有其他单元格的缺省值为 0.000。 值 0.000 指定关联的字段对之间不存在相关性。 矩阵中仅包括连续字段和有序字段。 名义、分类和标志字段以及分配了“固定”分布的字段不会显示在表中。

相关性列表。 用于显示表中字段对之间的相关性。 表中的每行显示字段对之间的相关性。 无法添加或删除行。 标题为“字段 1”和“字段 2”的列包含无法编辑的字段名称。 “相关性”列包含可以编辑的相关性；必须输入介于 -1.000（含本数）到 1.000（含本数）之间的值。 所有单元格的缺省值为 0.000。 列表中仅包括连续字段和有序字段。 名义、分类和标志字段以及分配了“固定”分布的字段不会显示在列表中。

重置相关性。 用于打开“**重置相关性**”对话框。 如果历史数据可用，那么您可以选择以下三个选项中的一个：

- **已拟合**。 用于将当前相关性替换为那些使用历史数据计算出的相关性。
- **零**。 用于将当前相关性替换为 0。
- **取消**。 用于关闭对话框。 相关性保持不变。

如果历史数据不可用，但您已更改相关性，那么可以选择将当前相关性替换为 0，或者选择取消。

显示为。 选择**表**可将相关性显示为矩阵。 选择**列表**可将相关性显示为列表。

在重新拟合时不重新计算相关性。 如果您要手动指定相关性，并阻止使用“模拟拟合”节点和历史数据自动拟合分布时覆盖这些相关性，请选择此选项。

对于具有分类分布的输入，使用拟合的多路列联表。 缺省情况下，所有具有分类分布的字段将包括在列联表或多路列联表中，具体取决于具有分类分布的字段的数目。 执行“模拟拟合”节点时，将构造列联表（与相关性相似）。 无法查看列联表。 如果选择了此选项，那么将使用列联表中的实际百分比来模拟具有分类分布的字段。 也就是说，将在新的模拟数据中重新创建名义字段之间的关联。 如果取消选择此选项，那么将使用列联表中的期望百分比来模拟具有分类分布的字段。 如果修改了某个字段，那么将从列联表中移除该字段。

“高级选项”视图

要模拟的个案数。 此选项显示用于指定要模拟的观测值数以及任何迭代的命名方式的选项。

- **最大个案数**。 此选项指定要生成的模拟数据的最大观测值数以及关联的目标值。 缺省值为 100,00，最小值为 1000，最大值为 2,147,483,647。
- **迭代**。 此数字由系统自动计算，并且无法进行编辑。 每次对分布参数指定多个值时，将自动创建迭代。
- **总行数**。 仅当迭代次数大于 1 时才启用。 该数字是自动计算的，使用所显示的等式，并且无法编辑。
- **创建迭代字段**。 仅当迭代次数大于 1 时才启用。 选中时，将启用**名称**字段。 有关更多信息，请参阅主题 第 48 页的『迭代』。
- **名称**。 仅当选中**创建迭代字段**复选框且迭代次数大于 1 时，才启用此选项。 通过在此文本字段中输入内容来编辑迭代字段的名称。 有关更多信息，请参阅主题 第 48 页的『迭代』。

随机种子值。 设置随机种子使您可以复制模拟。

- **复制结果**。 选择此选项后，将启用**生成**按钮和**随机种子**字段。
- **随机种子值**。 只有在选中**复制结果**复选框的情况下，才会启用此选项。 您可以在此字段中指定要用作随机种子的整数。 缺省值为 629111597。
- **生成**。 只有在选中**复制结果**复选框的情况下，才会启用此选项。 用于在**随机种子**字段中创建介于 1（含本数）到 999999999（含本数）之间的伪随机整数。

克隆字段

您可以在**“克隆字段”**对话框中指定要创建的所选字段副本数以及每个副本的命名方式。 调查复合效应（例如一些连续时间段内的利率或增长率）时，具有字段的多个副本非常有用。

此对话框的标题栏包含所选字段的名称。

要制作的副本数。 包含要创建的字段副本数。 单击箭头可选择要创建的副本数。 最小副本数为 1，最大数为 512。 副本数最初设置为 10。

复制后缀字符。 包含添加到每个副本的字段名称末尾的字符。 这些字符用于分隔字段名称与副本编号。 可以通过在此字段中输入内容来编辑后缀字符。 可以将此字段留空；在这种情况下，字段名称与副本编号之间将不存在任何字符。 缺省字符为下划线。

初始副本号。 包含第一个副本的后缀编号。 单击箭头可选择初始副本编号。 最小初始副本编号为 1，最大编号为 1000。 缺省初始副本编号为 1。

副本号步长。 包含后缀编号的增量。 单击箭头可选择增量。 最小增量为 1，最大增量为 255。 增量最初设置为 1。

字段。 包含副本的字段名称预览，对**“克隆字段”**对话框的任何字段进行编辑时，此预览将进行更新。 此文本由系统自动生成，并且无法进行编辑。

确定。 按照对话框中指定的内容生成所有副本。 这些副本将添加到“模拟生成”节点对话框的“模拟字段”表中，位于包含已复制字段的行的正下方。

取消。 用于关闭对话框。 将废弃所有已执行的更改。

拟合详细信息

只有在已通过执行“模拟拟合”节点来创建或更新“模拟生成”节点的情况下，**“拟合详细信息”**对话框才可用。 此对话框显示所选字段的自动分布拟合结果。 分布按拟合度进行排序，最接近的拟合分布首先列出。 您可以在此对话框中执行下列任务：

- 检查拟合到历史数据的分布。
- 选择其中一个已拟合的分布。

字段。 包含所选字段的名称。 无法编辑此文本。

视为（度量）。 显示所选字段的度量类型。 此类型来自“模拟生成”节点对话框中的“模拟字段”表。 可以通过单击箭头并从下拉列表中选择度量类型来更改此度量类型。 提供了以下三个选项：**连续**、**名义**和**有序**。

分布。 “分布”表显示适合于此度量类型的所有分布。 已拟合到历史数据的分布将按拟合度从最佳到最差的顺序进行排序。 拟合度由“模拟拟合”节点中选择的拟合统计量确定。 未拟合到历史数据的分布按字母顺序列示在表中已拟合的分布下方。

“分布”表包含以下列：

- **使用**。 所选单选按钮指示当前为字段选择的分布。 通过在“使用”列中选择与所需分布对应的单选按钮，您可以覆盖最接近的拟合分布。 在“使用”列中选择单选按钮还将显示所选字段的历史数据直方图（或条形图）上叠加的分布图。 一次只能选择一个分布。
- **分布**。 包含分布的名称。 无法编辑此列。
- **拟合统计量**。 包含针对分布计算的拟合统计量。 无法编辑此列。 此单元格的内容取决于字段的度量类型：
  - **连续**。 包含 Anderson-Darling 检验和 Kolmogorov-Smirnoff 检验的结果。 还将显示与这些检验关联的 p 值。 最先显示的是选择作为“模拟拟合”节点中的拟合度标准的拟合统计量，它用于对分布进行排序。 Anderson-Darling 统计量显示为 A=*aval* P=*pval*。 Kolmogorov-Smirnoff 统计量显示为 K=*kval* P=*pval*。 如果无法计算某个统计量，那么将显示一个点来代替数字。
  - **名义和有序**。 包含卡方检验的结果。 还将显示与此检验关联的 p 值。 统计量显示为 Chi-Sq=*val* P=*pval*。 如果未拟合分布，那么将显示**未拟合**。 如果无法以数学方法拟合分布，那么将显示**无法拟合**。

  注：对于经验分布，此单元格始终为空。

- **参数**。 包含与每个已拟合分布关联的分布参数。 这些参数显示为 *parameter_name* = *parameter_value*，参数之间以单空格分隔。 对于分类分布，参数名是类别，而参数值是关联的概率。 如果分布未拟合到历史数据，那么此单元格为空。 无法编辑此列。

直方图缩略图。 显示所选字段的历史数据直方图上叠加的所选分布图。

分布缩略图。 显示所选分布的说明和图示。

确定。 用于关闭对话框，并使用所选分布中的信息对所选字段的“模拟字段”表中以下列的值进行更新：度量、分布、参数以及最小值和最大值。 另外，还将更新“状态”列中的图标，以反映所选分布是否为与数据拟合程度最近的分布。

取消。 用于关闭对话框。 将废弃所有已执行的更改。

指定参数

您可以在“**指定参数**”对话框中手动为所选字段的分布指定参数值。 另外，也可以为所选字段选择其他分布。

“指定参数”对话框可通过以下三种方式打开：

- 双击“模拟生成”节点对话框中“模拟字段”表内的字段名称。
- 单击“模拟字段”表的“参数”或“最小值，最大值”列，然后从列表中选择**指定**。
- 在“模拟字段”表中选择一行，然后单击**编辑分布参数**图标。

字段。 包含所选字段的名称。 无法编辑此文本。

分布。 包含所选字段的分布。 此信息来自“模拟字段”表。 可以通过单击箭头并从下拉列表中选择分布来更改分布。 可用分布取决于所选字段的存储类型。

面数。 只有在从**分布**字段中选择了骰子分布的情况下，此选项才可用。 单击箭头可选择要将字段分割为的面数（类别）。 最小面数为 2，最大面数为 20。 面数最初设置为 6。

分布参数。 分布参数表为所选分布的每个参数都提供了一行。

注: 分布使用比率参数，并且形状参数为 α = k，逆尺度参数为 β = 1/θ。

该表包含两列：

- **参数**。 包含参数的名称。 无法编辑此列。
- **值**。 包含参数的值。 如果已根据“模拟拟合”节点创建或更新了“模拟生成”节点，那么此列中的单元格包含已通过将分布拟合到历史数据来确定的参数值。 如果已将“模拟生成”节点添加到“源”节点选用板内的流画布中，那么此列中的单元格将为空。 可以通过在单元格中输入内容来编辑值。 请参阅主题第 48 页的『分布』以获取有关每个分布所需要的参数以及可接受的参数值的更多信息。

  参数的多个值之间必须以逗号进行分隔。 为参数指定多个值将定义模拟的多个迭代。 只能为一个参数指定多个值。

  注: 对于存储类型为日期时间的字段，必须将分布参数指定为整数。 例如，要指定平均日期 1970 年 1 月 1 日，请使用整数 0。

注: 如果选择了骰子分布，那么分布参数表将略有不同。 此表为每个面（或类别）都提供了一行。 此表包含“值”列和“概率”列。 “值”列包含每个类别的标签。 标签的缺省值为 1 - N 之间的整数，其中 N 是面数。 可以通过在单元格中输入内容来编辑标签。 可以在单元格中输入任意值。 如果希望使用非数字值，那么在存储类型未设置为字符串的情况下，必须将数据字段的存储类型更改为字符串。 “概率”列包含每个类别的概率。 无法编辑概率，并且这些概率按照 1/N 进行计算。

预览。 根据指定的参数显示分布的样本图。 如果为一个参数指定了两个或两个以上的值，那么将显示此参数的每个值的样本图。 如果历史数据可用于所选字段，那么分布图将叠加在历史数据直方图上。

可选设置。 使用这些选项可以指定模拟数据的最小值和/或最大值。 小于最小值以及大于最大值的模拟数据将被拒绝，即使这些数据对于指定分布有效也是如此。

- **指定最小值**。 选择此选项可启用**拒绝小于最小值的值**字段。 如果选择了经验分布，那么此复选框将处于禁用状态。
- **拒绝低于以下值的值**。 只有在选择了**指定最小值**的情况下，才会启用此字段。 请输入模拟数据的最小值。 任何小于此值的模拟值将被拒绝。

- **指定最大值**。 选择此选项可启用**拒绝大于最大值的值**字段。 如果选择了经验分布，那么此复选框将处于禁用状态。
- **拒绝高于以下值的值**。 只有在选择了**指定最大值**的情况下，才会启用此字段。 请输入模拟数据的最大值。任何大于此值的模拟值将被拒绝。

确定。 用于关闭对话框，并对所选字段的“模拟字段”表的以下列的值进行更新：分布、参数以及“最小值，最大值”。 另外，还将更新“状态”列中的图标以反映所选分布。

取消。 用于关闭对话框。 将废弃所有已执行的更改。

迭代

如果您已经为固定字段或分布参数指定了多个值，那么将针对每个指定值生成一组独立的模拟观测值（有效的独立模拟）。 这使您可以调查更改字段或参数所产生的影响。 每个模拟观测值集称为迭代。 在模拟数据中，迭代是重叠的。

如果选中了“模拟生成”节点对话框的“高级选项”视图中的**创建迭代字段**复选框，那么迭代字段将作为具有数字存储的名义字段添加到模拟数据中。 通过在“高级选项”视图中的**名称**字段内输入内容，可以对此字段的名称进行编辑。 此字段包含一个标签，指示每个模拟观测值所属的迭代。 标签的格式取决于迭代的类型：

- **对固定字段进行迭代**。 标签包含：字段名称，后跟等号，然后是该迭代的字段值，即

  field_name = *field_value*
- **对分布参数进行迭代**。 标签包含：字段名称，后跟冒号，其次是迭代参数的名称，然后是等号，最后是该迭代的参数值，即

  field_name:*parameter_name* = *parameter_value*
- **对分类分布或范围分布的分布参数进行迭代**。 标签包含：字段名称，后跟冒号，然后是“Iteration”，最后是迭代编号，即

  field_name: Iteration *iteration_number*

分布

通过打开任何字段的“**指定参数**”对话框，并从**分布**列表中选择所需分布，然后在**分布参数**表中输入分布参数，您可以手动指定该字段的概率分布。 以下是有关特定分布的一些说明：

- **分类**。 分类分布描述数字值数目固定的输入字段，该输入字段称为类别。 每个类别都具有关联的概率，以使所有类别的概率总和等于 1。

  注: 如果您为类别指定的概率的总和不等于 1，那么将接收到警告。
- **负二项式 - 失败**。 描述在观察到指定次数的成功之前，一系列试验中失败次数的分布。 参数 *Threshold* 表示指定的成功次数，而参数 *Probability* 表示任何指定试验的成功概率。
- **负二项式 - 试验**。 描述在观察到指定次数的成功之前，所需失败次数的分布。 参数 *Threshold* 表示指定的成功次数，而参数 *Probability* 表示任何指定试验的成功概率。
- **范围**。 该分布由一组区间组成，每个区间都分配有一个概率，使得所有区间的概率之和等于 1。 给定区间中的值来自于在该区间上定义的均匀分布。 可以通过输入最小值、最大值以及关联的概率来指定区间。

  例如，您认为原料成本落入单位价格 $10 - $15 的范围内的概率为 40%，而落入单位价格 $15 - $20 的范围内的概率为 60%。 您将使用由两个区间 [10 - 15] 和 [15 - 20] 组成的“范围”分布对成本进行建模，并将与第一个区间关联的概率设置为 0.4，而将与第二个区间关联的概率设置为 0.6。 区间不必连续，它们甚至可以重叠。 例如，您可能指定了区间 $10 - $15 和 $20 - $25，或者指定了 $10 - $15 和 $13 - $16。
- **韦伯**。 参数 *Location* 是一个可选的位置参数，用于指定分布源所在的位置。

下表显示了可用于定制分布拟合的分布以及参数的可接受值。 其中某些分布可用于到特定存储类型的定制拟合，即使“模拟拟合”节点未自动将它们拟合到这些存储类型也是如此。

表 *12:* 可用于定制拟合的分布

分布	定制拟合支持的存储类型	参数	参数限制	注释
伯努利	整数、实数和日期时间	概率	0 ≤ *Probability* ≤ 1	
Beta	整数、实数和日期时间	Shape 1 Shape 2 最小值 最大值	≥ 0 ≥ 0 最大值 > 最小值	最小值和最大值是可选的。
二项式	整数、实数和日期时间	试验次数 (n) 概率 最小值 最大值	> 0, integer 0 ≤ *Probability* ≤ 1 最大值 > 最小值	试验次数必须是整数。 最小值和最大值是可选的。
分类	整数、实数、日期时间和字符串	类别名称（或标签）	0 ≤ *Value* ≤ 1	值表示类别的概率。 值的总和必须等于 1，否则将生成警告。
骰子	整数，字符串	Sides	2 ≤ *Sides* ≤ 20	每个类别（面）的概率按照 1/N 进行计算，其中 N 表示面数。 无法编辑概率。
经验	整数、实数和日期时间			无法编辑经验分布或者选择它作为某种类型。 只有在存在历史数据的情况下，经验分布才可用。
指数	整数、实数和日期时间	Scale 最小值 最大值	> 0 最大值 > 最小值	最小值和最大值是可选的。
固定	整数、实数、日期时间和字符串	值		无法为每个字段指定固定分布。 如果希望生成的数据中每个字段都是固定字段，那么您可以使用后跟“平衡”节点的“用户输入”节点。
伽玛	整数、实数和日期时间	形状 Scale 最小值 最大值	≥ 0 ≥ 0 最大值 > 最小值	最小值和最大值是可选的。 分布使用比率参数，并且形状参数为 α = k，逆尺度参数为 β = 1/θ。

表 *12:* 可用于定制拟合的分布 *(*继续*)*

分布	定制拟合支持的存储类型	参数	参数限制	注释
对数正态	整数、实数和日期时间	Shape 1 Shape 2 最小值 最大值	≥ 0 ≥ 0 最大值 > 最小值	最小值和最大值是可选的。
负二项式 - 失败	整数、实数和日期时间	Threshold 概率 最小值 最大值	≥ 0 0 ≤ *Probability* ≤ 1 最大值 > 最小值	最小值和最大值是可选的。
负二项式 - 试验	整数、实数和日期时间	Threshold 概率 最小值 最大值	≥ 0 0 ≤ *Probability* ≤ 1 最大值 > 最小值	最小值和最大值是可选的。
正态	整数、实数和日期时间	平均值 标准差 最小值 最大值	≥ 0 > 0 最大值 > 最小值	最小值和最大值是可选的。
泊松	整数、实数和日期时间	平均值 最小值 最大值	≥ 0 最大值 > 最小值	最小值和最大值是可选的。
范围	整数、实数和日期时间	Begin(X) End(X) Probability(X)	0 ≤ *Value* ≤ 1	X 是每个分级的指数。 概率值的总和必须等于 1。
三角	整数、实数和日期时间	方式 最小值 最大值	最小值 ≤ 值 ≤ 最大值 最大值 > 最小值	
均匀	整数、实数和日期时间	最小值 最大值	最大值 > 最小值	
韦伯	整数、实数和日期时间	Rate Scale 位置 最小值 最大值	> 0 > 0 ≥ 0 最大值 > 最小值	位置、最大值和最小值是可选的。

“扩展导入”节点

通过“扩展导入”节点，您可以运行 R 或 Python for Spark 脚本来导入数据。

“扩展导入”节点 -“语法”选项卡

选择语法类型 - **R** 或 **Python for Spark**。 然后，为导入数据输入或粘贴定制脚本。 语法就绪时，您可以单击**运行**来执行“扩展导入”节点。

R 示例

```
# import R demo data cars to modeler
modelerData <- cars

# write the data model that matches the data
var1<-c(fieldName="speed",fieldLabel="",fieldStorage="integer",fieldMeasure="",fieldFormat="",
fieldRole="")
var2<-c(fieldName="dist",fieldLabel="",fieldStorage="integer",fieldMeasure="",fieldFormat="",
fieldRole="")
modelerDataModel<-data.frame(var1, var2)
```

Python for Spark 示例

```
import spss.pyspark.runtime
from pyspark.sql import SQLContext
from pyspark.sql.types import *

cxt = spss.pyspark.runtime.getContext()
if cxt.isComputeDataModelOnly():
    _schema = StructType([StructField("Age", LongType(), nullable=True), \
                        StructField("Sex", StringType(), nullable=True), \
                        StructField("BP", StringType(), nullable=True), \
                        StructField("Cholesterol", StringType(), nullable=True), \
                        StructField("Na", DoubleType(), nullable=True), \
                        StructField("K", DoubleType(), nullable=True), \
                        StructField("Drug", StringType(), nullable=True)])
    cxt.setSparkOutputSchema(_schema)
else:
    sqlContext = cxt.getSparkSQLContext()
    # the demo data is in modeler installation path
    df = sqlContext.read.option("inferSchema", "true").option("header", "true").csv("/opt/IBM/
SPSS/ModelerServer/Cloud/demos/DRUG1n")
    cxt.setSparkOutputData(df)
    df.show()
    # print (df.dtypes[:])
```

“扩展导入”节点 -“控制台输出”选项卡

控制台输出选项卡包含当“语法”选项卡上的 R 脚本或 Python for Spark 脚本运行时接收到的任何输出（例如，如果使用 R 脚本，当执行**语法**选项卡上的 **R 语法**字段中的 R 脚本时，它显示从 R 控制台接收到的输出）。 此输出可能包括执行 R 或 Python 脚本时生成的 R 或 Python 错误消息或警告。 输出可主要用于调试脚本。 **控制台输出**选项卡还包含 **R 语法**或 **Python 语法**字段中的脚本。

每次执行“扩展导入”脚本时，都会使用从 R 控制台或 Python for Spark 接收到的输出来覆盖**控制台输出**选项卡的内容。 输出不能编辑。

过滤或重命名字段

您可以在流中的任意时间点上重命名或排除字段。 例如，作为医学研究人员，您可能不关心患者（记录级别数据）的钾水平（字段级别数据）；因此，您可以过滤掉 K（钾）字段。 也可以使用单独的“过滤”节点，或者源节点或输出节点上的“过滤”选项卡实现此操作。 无论使用哪种节点，结果都是一样的。

- 可以在将数据从源节点（如变量文件、固定文件、统计信息 文件、XML 或扩展导入）读入 IBM SPSS Modeler 时对字段进行重命名或过滤。
- 使用过滤节点，可以在流的任何位置对字段进行重命名或过滤。
- 通过 统计信息 导出、统计信息 变换、统计信息 模型和 统计信息 输出节点，可以对字段进行过滤或重命名，使之符合 IBM SPSS Statistics 命名标准。
- 您可以使用以上任何节点中的“过滤器”选项卡来定义或编辑多响应集。
- 最后，可以使用“过滤”节点将一个源节点中的字段映射至另一个源节点。

“地理空间”源节点

您可以使用“地理空间”源节点将地图或空间数据引入到数据挖掘会话中。 您可以通过两种方法中的一种来导入数据：

- 通过形状文件 (.shp) 进行导入
- 通过连接到包含地图文件的分层文件系统所在的 ESRI 服务器进行导入。

  注： 您只能连接到公共映射服务。

空间-时间预测 (STP) 模型可以将地图或空间元素包括在其预测中。 有关这些模型的更多信息，请参阅《Modeler 建模节点》指南 (ModelerModelingNodes.pdf) 的『时间序列模型』部分中标题为『空间-时间预测建模节点』的主题。

设置“地理空间”源节点的选项

数据源类型 您可以从**形状文件** (.shp) 导入数据，也可以连接到**地图服务**。

如果您使用的是**形状文件**，请输入该文件的文件名和文件路径，或者进行浏览以选择文件。 该文件必须位于本地目录上或从映射驱动器进行访问；您无法使用统一命名约定 (UNC) 路径访问该文件。

注： 形状数据同时需要 .shp 和 .dbf 文件。 这两个文件必须同名并且在同一个文件夹中。 您选择 .shp 文件时，将自动导入 .dbf 文件。 另外，可以存在一个 .prj 文件，用于指定形状数据的坐标系。

如果您使用的是**地图服务**，请输入该服务的 URL，并单击**连接**。 连接到该服务后，该服务中的层将显示在对话框底部**可用地图**窗格中的树结构中；请展开该树并选择所需的层。

注： 您只能连接到公共映射服务。

自动定义地理空间数据

缺省情况下，在有可能时，SPSS Modeler 将使用正确的元数据自动定义源节点中的所有地理空间数据字段。 元数据可能包括地理空间字段的测量级别（例如“点”或“多边形”）以及这些字段所使用的坐标系，其中包括原点（例如纬度 0 经度 0）和计量单位之类的详细信息。 有关测量级别的更多信息，请参阅第 105 页的『地理空间测量子级别』。

构成形状文件的 .shp 和 .dbf 文件包含用作键的公共标识字段。 例如，.shp 文件可能包含国家或地区，并将国家或地区名称字段用作标识，而 .dbf 文件可能包含有关这些国家或地区的信息，并且同样将国家或地区名称用作标识。

注： 如果坐标系与缺省的 SPSS Modeler 坐标系不同，那么您可能必须重新投影数据以使用所需的坐标系。有关更多信息，请参阅 第 138 页的『“重新投影”节点』。

JSON 源节点

使用 JSON 源节点以使用 UTF-8 编码将数据从 JSON 文件导入到 SPSS Modeler 流。 JSON 文件中的数据的格式可以为对象、数组或值。 此 JSON 源节点仅支持读取对象数组，并且无法嵌套对象。

示例 JSON 数据：

```
[
    {
        "After": 122762,
        "Promotion": 1467,
        "Cost": 23.99,
        "Class": "Confection",
        "Before": 114957
    },
    {
        "After": 137097,
        "Promotion": 1745,
        "Cost": 79.29,
        "Class": "Drink",
        "Before": 123378
```

```
    }
]
```

在 SPSS Modeler 从 JSON 文件读取数据时，其执行以下转换。

表 *13: JSON* 数据存储转换

JSON 值	**SPSS Modeler 数据存储**
string	String
number(int)	整数
number(real)	实数
true	1(Integer)
false	0(Integer)
null	缺失值

JSON 源节点对话框提供以下选项。

JSON 数据源。 选择要导入的 JSON 文件。

JSON 字符串格式。 指定 JSON 字符串格式。 如果 JSON 文件是名称和值对的集合，那么选择**记录**。 JSON 源节点导入名称作为 SPSS Modeler 中的字段名称。 或者，如果 JSON 数据仅使用值（无名称），那么选择**值**。

公共源节点选项卡

通过单击相应的选项卡可为所有源节点指定下列选项：

- **“数据”选项卡**。 用于更改缺省存储类型。
- **“过滤”选项卡**。 用于删除或重命名数据字段。 此选项卡所提供的功能与“过滤”节点相同。 有关更多信息，请参阅主题 第 113 页的『设置过滤选项』。
- **“类型”选项卡**。 用于设置测量级别。 此选项卡所提供的功能与类型节点相同。
- **“注解”选项卡**。 用于所有节点，此选项卡提供的选项可用于重命名节点、提供定制的工具提示及存储长的注解。

在源节点中设置测量级别

字段属性可在源节点中指定也可在单独的“类型”节点中指定。 两种节点的功能相似。 可用的属性如下：

- **字段** 双击任何字段名均可指定 IBM SPSS Modeler 中的数据的值标签和字段标签。 例如，从 IBM SPSS Statistics 导入的字段元数据可在此处查看或修改。 与之相似，您也可以为字段及其值创建新的标签。 您在此处指定的标签将根据您在“流属性”对话框中的选项显示在整个 IBM SPSS Modeler 中。
- **测量** 这是测量级别，用于描述给定字段中的数据的特征。 如果已经了解某个字段的所有详细信息，则称为**已完全实例化** 。 有关更多信息，请参阅 第 104 页的『测量级别』。

  注: 字段的测量级别与字段的存储类型不同，后者指示数据是作为字符串、整数、实数、日期、时间、时间戳记还是列表进行存储。
- **值** 此列使您能够指定用于从数据集中读取数据值的选项，或者使用**指定**选项在单独的对话框中指定测量级别和值。 您还可以选择遍历字段，而不读取它们的值。 有关更多信息，请参阅 第 107 页的『数据值』。

  注: 如果相应的**字段**条目包含列表，那么您无法对此列中的单元格进行修改。
- **缺失** 用于指定如何处理此字段的缺失值。 有关更多信息，请参阅 第 110 页的『定义缺失值』。

  注: 如果相应的**字段**条目包含列表，那么您无法对此列中的单元格进行修改。
- **检查** 在此列中，您可以设置选项，以确保字段值符合指定的值或范围。 有关更多信息，请参阅 第 110 页的『检查类型值』。

  注: 如果相应的**字段**条目包含列表，那么您无法对此列中的单元格进行修改。

- **角色** 用于向建模节点指示字段将成为用于机器学习过程的**输入**（预测变量字段）还是**目标**（预测的字段）。 **两者**、**无**以及**分区**也是可用角色，最后一个可用角色表明字段用于将记录分区到不同的样本中，以用于进行训练、检验和验证。 值 **分割** 指定将为字段的每个可能值构建单独的模型。 有关更多信息，请参阅 第 111 页的『设置字段角色』。

有关更多信息，请参阅主题 第 103 页的『类型节点』。

何时在源节点上进行实例化

可使用两种方法了解数据存储类型和字段值。 实例化 可在第一次将数据导入 IBM SPSS Modeler 时在源节点上进行，也可以在将类型节点插入数据流时进行。

在下列情况下，在源节点上进行实例化非常有用：

- 数据集较小。
- 计划使用表达式构建器派生新字段（实例化可使字段值在表达式构建器中可用）。

通常，如果数据集不是非常大，并且不打算稍后在流中添加字段，那么在源节点上进行实例化是最方便的方法。

注: 如果要在数据库导出节点中导出数据，那么该数据必须完全实例化。

从源节点中过滤字段

使用源节点对话框上的“过滤”选项卡可以根据对数据的初始检查排除下游操作中的字段。 此功能非常有用，例如，如果数据中存在重复的字段，或假设您已非常熟悉数据并能够排除不相关的字段，那么可选择此功能。 此外，还可以稍后在流中添加一个单独的“过滤”节点。 此节点的功能与上述两种情况下所使用的功能相似。 有关更多信息，请参阅主题 第 113 页的『设置过滤选项』。

第 3 章 记录操作节点

记录操作概述

记录操作节点用于在记录级别对数据进行更改。 这些操作在数据挖掘的 **数据理解** 和 **数据准备** 阶段非常重要，因为通过这些操作可以根据您的特定业务需要裁剪数据。

例如，根据使用“数据审核”节点（“输出”选用板）执行的数据审核结果，您可能决定合并过去三个月的客户购买记录。 使用合并节点，可以基于某个关键字段（如 客户标识 ）的值合并记录。 您还可能会发现无法管理一个包含超过一百万条网站点击信息记录的数据库。 使用“样本”节点，可以选择要用于建模的数据子集。

记录操作选项板包含下列节点：

“选择”节点根据特定条件从数据流中选择或废弃一部分记录。 例如，可以选择与特定销售区域相关的记录。

“样本”节点用于选择一部分记录。 支持各种样本类型，包括分层、聚类和非随机（结构化）样本。 采样对于提高性能以及选择相关记录组或事务组进行分析十分有用。

“均衡”节点用于纠正数据集中的不平衡，以使其遵循指定的条件。 “均衡”伪指令根据指定系数调整条件成立的记录所占的比例。

“汇总”节点将一系列输入记录替换为经过摘要和汇总的输出记录。

“近因、频率和货币 (RFM) 总量”节点使您能够接受客户的历史交易数据、剥离任何未使用的数据，并将所有余下的交易数据合并到一行中，其中列出客户上次与您交易的时间、进行的交易数量以及这些交易的总货币价值。

“排序”节点根据一个或多个字段的值按升序或降序对记录进行排序。

“合并”节点使用多个输入记录，并创建包含某些或全部输入字段的单个输出记录。 这对于合并来源不同的数据 非常有用，例如内部客户数据和已购买人群统计数据。

“追加”节点用于连接多组记录。 另外，也可以用于将结构类似但内容不同的数据集组合到一起。

“区分”节点通过将第一个区分记录传递到数据流，或者通过丢弃第一个记录并将任何重复记录传递到数据流，移除重复的记录。

“流式时间序列”节点在一个步骤中对时间序列模型同时进行构建和评分。 您可以在本地或分布式环境中使用带有数据的节点；在分布式环境中，可以利用 IBM SPSS Analytic Server 的能力

Spectral Clustering© 算法使用多个特征向量将数据投影到维数更少的空间。 然后，将在新空间中应用 K-Means 聚类算法以将数据分隔为不同集群。 对于有许多字段的小型记录，此操作会比较快速，而对于大型数据集，则需要大量计算。 SPSS Modeler 中的 Spectral Clustering 节点公开 Spectral Clustering 库的核心特征和常用参数。 此节点使用 Python 进行实现。

空间时间限制 (STB) 是进行了 Geohash 计算的空间位置的扩展。 更具体地说，STB 是一个字母数字字符串，它表示形状规则的空间和时间区域。

“流式 TCM”节点在一个步骤中对时间因果模型同时进行构建和评分。

借助 CPLEX Optimization 节点，可以通过优化编程语言 (OPL) 模型文件来使用基于优化的复杂数学。 此功能可在不再受支持的 IBM Analytical Decision Management 产品中使用。 但是，您也可以在 SPSS Modeler 中使用 CPLEX 节点，而无需 IBM Analytical Decision Management。

记录操作选用板中的很多节点都需要使用 CLEM 表达式。 如果您熟悉 CLEM，则可以在字段中键入表达式。但是，所有表达式字段都提供了一个打开 CLEM 表达式构建器的按钮，可以帮助您自动创建此类表达式。

图 *1:* “表达式构建器”按钮

选择节点

您可以使用选择节点，根据某个特定的条件（如 BP （血压） = "HIGH" 选择或丢弃数据流中的部分记录。

方式。 指定将符合条件的记录包括还是不包括在数据流中。

- **包含**。 选择包括符合选择条件的记录。
- **废弃**。 选择排除符合选择条件的记录。

条件。 显示将要用于检验每个记录的选择条件，您可以使用 CLEM 表达式进行指定。 在窗口中输入表达式，或者单击窗口右侧的计算器（表达式构建器）按钮，使用表达式构建器。

如果您选择根据条件丢弃记录，例如以下条件：

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

缺省情况下，“选择”节点也会废弃所有选择字段均为空值的记录。 为了避免这种情况，将以下条件附加到原始条件：

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

“选择”节点还用于选择记录的比例。 通常情况下，对于此操作要使用另外一个节点，“样本”节点。 但如果您要指定的条件比提供的参数更复杂，那么可以使用“选择”节点创建自己的条件。 例如，您可以创建类似下面的条件：

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

此条件将选择大约 40% 显示高血压的记录，并向下游传递这些记录进行进一步分析。

样本节点

您可以使用“样本”节点来选择记录的子集进行分析，或指定要废弃的记录的比例。 支持各种样本类型，包括分层、聚类和非随机（结构化）样本。 需要使用抽样的原因有以下几点：

- 通过评估数据子集上的模型提高性能。 通过样本评估的模型通常与利用全部数据集得到的模型一样准确，并且如果提高的性能允许您体验尚未尝试的不同方法，那么所得的模型还有可能更为准确。
- 选择相关的记录或交易组来进行分析，例如选择在线购物车（或市场购物篮）中的所有项目，或特定近邻的所有属性。
- 指定单元或观测值以进行随机检查，从而确保质量、防止欺诈和保证安全。

注意：如果仅希望将数据分区到训练样本和检验样本以进行验证，那么可以改用“分区”节点。 有关更多信息，请参阅主题 第 131 页的『分区节点』。

样本的类型

聚类样本。 属于样本组或聚类，而不是单个单元。 例如，假设您有一个数据文件，其中每个学生对应一条记录。 如果按学校聚类并且样本大小为 50%，那么将选中一半的学校并从每所选定的学校中选出所有学生。 而去除未选中学校的学生。 一般而言，您可能期望选出大约一半的学生，但由于学校规模不同，百分比也可能不太准确。 同样，您可以按交易标识对购物车项目进行聚类，以确保保留所选交易的所有项目。 有关按镇对属性聚类的示例，请参阅 *complexsample_property.str* 样本流。

分层样本。 在总体或分层的没有重叠的子组中独立选择样本。 例如，您可以确保以同样的比例对男性和女性进行抽样，或者可以确保在城市总体中显示每个地区或社会经济群体。 还可以为每层指定一个不同的样本大小（例如，如果您认为一个组在原始数据中被低估了）。 有关按县对属性分层的示例，请参阅 *complexsample_property.str* 样本流。

系统抽样或 n 中取 1 抽样。 如果随机选择难以实现，那么可以系统（以固定间隔）或顺序方式抽取单元。

抽样权重。 在绘制复杂样本时会自动计算抽样加权，并且这些加权会与每个抽样单元在原始数据中所表示的“频率”大致对应。 因此，样本的加权总和应该可以估计原始数据的大小。

抽样框

抽样框定义将包含在样本或研究中的观测对象的潜在源。 在某些情况下，抽样框可以识别总体中的每个单独成员并且可以包含样本中的任何成员 - 例如，对来自某条产品线的产品进行抽样。 更普遍的情况是，您将无法访问每一个可能的观测对象。 例如，在选举之前，您无法确定谁将在选举中投票。 在这种情况下，您可以将选民名册作为抽样框，即使在下列情况下也是如此：有些注册人不会投票，而有些人在您停止注册时还尚未注册，但可能会投票。 您无法对抽样框之外的任何人进行抽样。 抽样框是否在本质上与您尝试评估的总体足够相似，是必须要为每个现实的观测对象解决的问题。

样本节点选项

您可以根据需要，选择 **简单** 或 **复杂** 方法。

简单抽样选项

通过“简单”方法，您可以选择记录的随机百分比、连续记录或所有第 *n* 条记录。

方式。 选择对于下面的模式遍历（包括）还是丢弃（排除）记录：

- **包含样本**。 包含数据流中的选定记录并废弃所有其他记录。 例如，如果您将模式设置为 **包含样本** 并将 **n 中取 1** 选项设置为 5，则每隔五个记录便有一个记录被包含进来，结果将生成大约为原大小五分之一的数据集。 此模式为对数据进行抽样的缺省模式，并且是使用复杂方法时的唯一模式。
- **废弃样本**。 排除选定记录并包含所有其他记录。 例如，如果您将模式设置为 **丢弃样本** 并将 **n 中取 1** 选项设置为 5，则每隔五条记录便有一条被丢弃（排除）。 此模式仅适用于简单方法。

样本。 从下列选项中选择抽样方法：

- **最前面**。 选择此选项将使用连续数据抽样。 例如，如果最大样本大小设置为 10000，则前 10000 条记录会被选中。
- **n 中取 1**。 选择此选项会按照这样的方式抽样数据：每隔 *n* 个记录进行一次遍历或废弃。 例如，如果 *n* 设为 5，则每隔五条记录便会选中一条。
- **随机 %**。 选择此选项将随机抽样指定百分比的数据。 例如，如果百分比设置为 20，那么根据选择的模式，将 20% 的数据传递到数据流或将其废弃。 使用该字段可指定抽样百分比。 您还可以使用 **设置随机种子** 控件指定一个种子值。

  使用块级别采样（仅限数据库内）。 在 Oracle 或 IBM Db2 数据库上执行数据库内挖掘时，只在您选择随机百分比抽样时才启用此选项。 在这些情况，块级别抽样的效率会更高。

  注: 每次运行相同的随机样本设置时，系统不会返回确切的行数。 这是因为每个输入记录包含在样本中的可能性为 N/100（其中，N 是您在节点中指定的**随机 %**），而且可能性是独立的；因此结果不是确切的 N%。

最大样本大小。 指定样本中所包含的最大记录数。 此选项为多余选项，因此在选定 **第一个** 和 **包括** 时会被禁用。 另外，当与 **随机 %** 选项结合使用时还请注意，此设置可能会阻止选中某些记录。 例如，如果数据集中有一千万条记录，而您选择了 50% 的记录且最大样本大小为三百万条记录，那么将选中前六百万条记录中的 50% 的记录，剩余的四百万条记录便不会再被选中。 为避免这种限制，请选择 **复杂** 抽样方法，然后对三百万条记录进行随机样本，无需指定聚类或分层变量。

复杂抽样选项

通过复杂样本选项，您可以与其他选项一起更好地控制样本，包括聚类样本、分层样本和加权样本。

聚类和分层。 允许您指定聚类和分层并根据需要输入权重字段。 有关更多信息，请参阅主题 第 59 页的『聚类和分层设置』。

样本类型。

- **随机**。 在每一层内随机选择聚类或记录。
- **系统化**。 以固定间隔选择记录。 除了会根据随机种子更改第一条记录的位置之外，此选项工作原理与 *n* 中取 *1* 方法基本相似。 *n* 的值会根据样本大小和比例自动确定。

采样单位。 可以选择比例或计数作为基本样本单元。

样本大小。 您可以按以下几种方式指定样本大小：

- **固定**。 允许您将样本总大小指定为计数或比例。
- **定制**。 允许您为每个子组或分层指定样本大小。 此选项只有在“聚类”和“分层”子对话框中指定了层字段时才可用。
- **变量**。 允许用户选取一个字段来为每个子组或层定义样本大小。 对于特定层内的每条记录，此字段应该都有相同的值；例如，如果样本按县分层，那么具有 *county = Surrey* 的所有记录必须具有相同值。 该字段必须为数值型并且它的值必须与所选样本单元相匹配。 比例的值应该大于 0 小于 1；计数的最小值为 1。

每层的最小样本。 指定记录的最小值（如果已指定了聚类字段，可指定聚类的最小值）。

每层的最大样本。 指定记录或聚类的最大值。 如果在没有指定聚类或分层字段的情况下选择了此选项，那么将选择指定大小的随机或系统化样本。

设置随机种子值。 根据随机数百分比对记录进行抽样或分区时，此选项允许在另一会话中复制相同的结果。 通过指定随机数生成器所使用的起始值，可以确保在每次执行节点时都会分配相同的记录。 输入所需的种子值，或单击 **生成** 按钮自动生成一个随机值。 如果未选中该选项，则每次执行节点时会生成不同的抽样。

注: 对从数据库中读取的记录使用**设置随机种子**选项时，可能需要在抽样前使用“排序”节点以确保每次执行节点时都获得相同的结果。 这是因为随机种子依赖于记录的顺序，而在关系数据库中不能保证记录具有这种顺序。 有关更多信息，请参阅主题 第 64 页的『排序节点』。

聚类和分层设置

通过“聚类”和“分层”对话框，您可以在绘制复杂样本时选择聚类、层和权重字段。

聚类。 指定用于聚类记录的分类字段。 会根据聚类成员资格对记录进行抽样，有些聚类包含在内，有些聚类不包含在内。 但是如果包含了指定聚类的所有记录，那么所有的聚类都将包含在内。 例如，当分析购物车中的产品关联时，您可以按交易标识对项目进行聚类从而确保可以维护选定交易中的所有项目。 如果改为对记录进行抽样，则将破坏一起销售的项目的信息，您可以对交易进行抽样以确保已保存选定交易的所有记录。

分层依据。 指定用于分层记录的分类字段，这样将在总体或层的没有重叠的子组中独立选择样本。 例如，如果选中一个按 50% 的比例抽取的按性别分层的样本，那么采用两个按 50% 的比例抽取样本，一个为男性，另一个为女性。 例如，层可以为社会经济学群体、工作类别、年龄组或种族组，从而可以确保所关注的子组有足够的样本大小。 如果在原始数据集中女性人数为男性人数的三倍，那么此比率将通过从每个组中抽样而得以保存。 您还可以指定多个层字段（例如，按各区域内的产品线或各个产品线所在的区域进行抽样）。

注意：如果按包含缺失值（空值或系统缺失值、空字符串、空白以及空值或用户定义的缺失值）的字段进行分层，那么无法为层指定定制样本大小。 当按包含缺失值或空白值的字段进行分层时，如果要使用定制样本大小，那么需要在上游进行填写。

使用输入权重。 指定在抽样之前加权记录的字段。 例如，如果权重字段值的范围为 1 到 5，那么权重为 5 的记录被选中的几率是其他记录的 5 倍。 该字段的值将被节点生成的最终输出加权覆盖（请参阅以下章节）。

新的输出加权。 指定在未指定输入权重字段的情况下，记录最终加权的字段的名称。 （如果已指定输入权重字段，则上述的最终加权将替换其值，并且无法创建独立的输出权重字段。） 输出加权值表示原始数据中每一个抽样记录所代表的记录数。 通过加权值总和，可以评估样本的大小。 例如，如果按 10% 的比例随机抽取样本，则所有记录的输出加权将为 10，表示每个抽取的记录大体上代表原始数据中的 10 条记录。 在分层或加权样本中，输出加权值可能会发生变化，具体取决于每层的样本比例。

注释

- 如果无法获得所需抽样的总体的完整列表，但可以得到某些组或聚类的完整列表，那么聚类抽样会非常有用。 如果随机样本生成一列检验主项，而要联系所有的对象又不切实际，那么可以使用聚类抽样。 例如，选择拜访一个县的所有农民要比选择全国范围内所有县的农民要容易的多。
- 为了在每个层内对聚类进行独立抽样，您可以同时指定聚类字段和分层字段。 例如，您可以在每个县内对按县分层、按镇聚类的属性值进行抽样。 这将确保从每个县所提取的城镇样本保持独立。 样本中将包含某些镇，而其他镇将不会包含在其中，但对于所包含的每个镇，镇内的所有属性也一定包含在其中。
- 要从每个聚类内选择单元的随机样本，您可以将两个“样本”节点联系起来。 例如，如上所述，您可以首先对按县分层的镇进行抽样。 然后附加另一个样本节点并选择 镇 作为分层字段，这样您可以在每一个镇中按一定比例对记录进行抽样。
- 如果需要使用字段组合来唯一标识聚类，那么可以使用“派生”节点生成新的字段。 例如，如果多个商店在交易时使用的是相同的数字系统，那么可以派生结合了商店标识和交易标识的新字段。

层的样本大小

提取分层样本时，缺省选项是对每个层中相同比例的记录和聚类进行抽样。 例如，如果某个组的数目超出另一个组数目的 3 倍，那么通常希望在样本中保留同一比率。 但如果不是这种情况，那么可以为每个层单独指定样本大小。

“层的样本大小”对话框列出了层字段的每个值，您可以覆盖层的缺省值。 如果选择了多个层字段，那么将列出每个可能的值组合，这样您可以指定具体的大小，例如每个城市内每一种族组的大小，或每个县内的每个镇的大小。 可以将大小指定为比例或计数，具体取决于“样本”节点中现有设置。

指定层的样本大小

1. 在“样本”节点，选择**复杂**，然后选择一个或多个层字段。 有关更多信息，请参阅主题 第 59 页的『聚类和分层设置』。
2. 选择 **自定义**，然后选择 **指定大小**。
3. 在“层的样本大小”对话框中，单击左下角的**读取值**按钮填充屏幕。 如有必要，您可能需要在上游源节点或“类型”节点中实例化值。 有关更多信息，请参阅主题 第 106 页的『什么是实例化？』。
4. 单击任意一行以覆盖该层的缺省大小。

有关样本大小的注意事项

例如，如果不同的层具有不同的方差，为了使样本大小与标准差成比例，定制样本大小可能会十分有用。（如果层中的观测值变化比较大，则需要抽样更多的观测值以获得具有代表性的样本。） 或者层比较小，而您可能想要使用更大的样本比例以确保将观测值的最小数包含在内。

注意：如果按包含缺失值（空值或系统缺失值、空字符串、空白以及空值或用户定义的缺失值）的字段进行分层，那么无法为层指定定制样本大小。 当按包含缺失值或空白值的字段进行分层时，如果要使用定制样本大小，那么需要在上游进行填写。

平衡节点

您可以使用 Balance 节点修正数据集中的不平衡，以便它们符合指定的检验标准。 例如，假设某个数据集只有两个值（*low* 或 *high*），并且 90% 的观测值为 *low*，而只有 10% 的观测值为 *high*。 很多建模技术处理此类偏倚数据都有困难，因为它们倾向于只学习这些 *low* 的结果，而忽略 *high* 的结果（因为这些结果少的可怜）。 如果数据平衡很好， *low* 和 *high* 结果具有大致相同的数量，那么模型将更有可能找出分辨这两个组的模式。 这种情况下，平衡节点对于创建平衡指令，从而减少带有 *low* 结果的观测值数量非常有用。

平衡是通过复制记录，然后根据指定的条件丢弃记录完成执行的。 将始终遍历不符合任何条件的记录。 因为此过程的工作模式为复制和/或废弃记录，所以在下游操作中将丢失数据的原始顺序。 在向数据流添加平衡节点之前，请确保派生任何与序列相关的值。

注意：Balance 节点可从条形图和直方图自动生成。 例如，您可以平衡数据以显示某一分类字段所有分类的相同比例，如分布图所示。

示例。 构建 RFM 流以识别积极响应以往营销活动的最新客户时，销售公司的市场营销部可以使用 Balance 节点来平衡数据中真假响应之间的差异。

为平衡节点设置选项

记录平衡指令。 列出当前平衡指令。 每个指令都包括一个因子和一个条件，该条件告知软件“在该条件为真的情况下以指定的因子值提高记录比例”。 如果因子小于 1.0，那么表示指定记录的比例要降低。 例如，如果您要减少治疗药为 drug Y 的记录数，则可以使用因子 0.7 和条件 `Drug = "drugY"` 创建一个平衡指令。 此指令表示对于所有下游操作，治疗药为 drug Y 的记录数将减少到 70%。

注意：用于减少的平衡因子可以指定为四位小数。 小于 0.0001 的因子设置会产生错误，因为这样的结果无法正确计算。

- **创建条件**，此操作通过单击该文本字段右侧的按钮完成。 此操作将插入一个用于输入新条件的空行。 要为条件创建 CLEM 表达式，请单击表达式构建器按钮。
- **删除指令**，此操作通过使用红色删除按钮完成。
- **对指令排序**，此操作通过上下方向按钮完成。

仅平衡训练数据。 如果流中存在分区字段，那么此选项仅平衡训练分区中的数据。 尤其是，当生成需要不平衡检验或验证分区的调整倾向评分时，此选项非常有用。 如果流中不存在分区字段（或已指定多个分区字段），那么将忽略此选项并平衡所有的数据。

“汇总”节点

汇总是一项数据准备任务，经常用于减小数据集的大小。 继续执行汇总之前，您应该花一些时间来清理数据，尤其要关注缺失值。 完成汇总后，或许会丢失可能有用的缺失值信息。

您可以使用“汇总”节点将一系列输入记录替换为摘要，即经过汇总的输出记录。 例如，您可能有一组输入销售记录，例如下表中所显示的记录。

表 *14:* 销售记录输入示例

年龄	性别	区域	分支	Sales
23	M	S	8	4
45	M	S	16	4
37	M	S	8	5
30	M	S	5	7
44	M	N	4	9
25	M	N	2	11
29 日	F	S	16	6
41	F	N	4	8
23	F	N	6	2
45	F	N	4	5
33	F	N	6	10

您可以将 *Sex* 和 *Region* 作为关键字段对这些记录进行汇总。 然后选择使用**平均值**模式汇总年龄，并使用**合计**模式汇总销售。 在“汇总”节点对话框中选择**在字段中包含记录计数**后，汇总的输出将显示在下表中。

表 *15:* 汇总记录的示例

年龄（均值）	性别	区域	销售量（总和）	记录计数
35.5	F	N	25	4
29 日	F	S	6	1
34.5	M	N	20	2
33.75	M	S	20	4

例如，您可从中了解到，北部区域四名女性销售人员的平均年龄为 35.5 岁，其销售量总和为 25 件产品。

注：如果未指定汇总方式，那么将自动废弃分支之类的字段。

设置“汇总”节点的选项

在“汇总”节点上，您可以指定以下内容。

- 一个或多个用作汇总类别的关键字段
- 一个或多个要为其计算汇总值的汇总字段
- 一种或多种汇总模式（汇总类型），用于每个汇总字段的输出

您还可以指定用于新添加字段的缺省汇总模式，并使用表达式（类似于公式）对汇总进行分类。

请注意，对于所增加的性能，启用并行处理可能会有益于汇总操作。

关键字段。 列出可用作汇总类别的字段。 连续（数字）字段和分类字段都可用作关键字段。 如果您选择多个关键字段，那么这些值将进行合并，以生成用于汇总记录的键值。 对于每个唯一的关键字段，将会生成一条汇总记录。 例如，如果**性别**和**区域**是关键字段，那么**男性**和**女性**与**北部**和**南部**区域的每个唯一组合（四个唯一组合）都将具有一条汇总记录。 要添加关键字段，请使用窗口右侧的字段选择器按钮。

对话框的剩余部分分为两个主要区域 - **基本汇总**和**汇总表达式**。

基本汇总

汇总字段。 列出将汇总其值的字段以及所选的汇总方式。 要向此列表中添加字段，请使用右侧的“字段选择器”按钮。 可用的汇总模式如下。

注: 某些模式不适用于非数字字段（例如，**总和**不适用于日期/时间字段）。 不能用于所选汇总字段的模式将被禁用。

- **总和** 选择此选项可返回每个关键字段组合的合计值。 总和是指所有具有非缺失值的观测值中值的总计。
- **平均值**。 选择此选项可返回每个关键字段组合的平均值。 该均值是对集中趋势的测量，它是算术平均值（总和除以观测值数）。
- **最短** 选择此选项可返回每个关键字段组合的最小值。
- **最大值**。 选择此选项可返回每个关键字段组合的最大值。
- **标准差**。 选择此选项可返回每个关键字段组合的标准差。 标准差是对围绕平均值的离差的测量，该值等于方差测量结果的平方根。
- **中位数**。 选择此选项可返回每个关键字段组合的中值。 中位数是对集中趋势的测量，但对于远离中心的值不敏感（这与均值不同，均值容易受到少数极大或极小值的影响）。 也称为第 50 个百分位数或第二个四分位数。
- **计数**。 选择此选项可返回每个关键字段组合的非空值计数。
- **方差**。 选择此选项可返回每个关键字段组合的方差值。 方差是对围绕平均值的离差的测量，该值等于平均偏差的平方和除以观测值数减一。
- **第一个四分位数**。 选择此选项可返回每个关键字段组合的第一个四分位数（第 25 个百分位数）值。
- **第三个四分位数**。 选择此选项可返回每个关键字段组合的第三个四分位数（第 75 个百分位数）值。

注: 在运行包含“汇总”节点的流的情况下，将 SQL 推送回 Oracle 数据库时针对第一个四分位数和第三个四分位数返回的值可能与本机方式下返回的那些值不同。

缺省方式。 指定要用于新添加字段的缺省汇总方式。 如果您频繁使用同一种汇总，请在此处选择一个或多个方式，然后使用右侧的“应用于全部内容”按钮，以便将所选方式应用于上面列出的所有字段。

新字段名称扩展。 选择此选项可添加后缀或前缀（例如，**1** 或 **new**）以复制所汇总的字段。 例如，如果您已选择后缀选项，并指定 **1** 作为扩展名，那么针对字段 **Age** 的最小值汇总结果会生成 `Age_Min_1` 字段名。 请注意，汇总扩展名（例如，`_Min` 或 `Max_`）会自动添加至新字段，以指示所执行的汇总类型。 选择 **后缀** 或 **前缀** 可指明您首选的扩展样式。

在字段中包含记录计数。 缺省情况下，选择此选项可在每条输出记录中包含一个额外的字段 `Record_Count`。 此字段表明汇总了多少输入字段而形成了每个汇总记录。 在编辑字段中键入内容，以便为此字段创建定制名称。

注: 计算汇总时将排除系统空值，但它们会包括在记录计数中。 另一方面，空白值既包括在汇总中也包括在记录计数中。 要排除空白值，您可以使用“填充”节点将空白值替换为空值。 您还可以使用“选择”节点移除空白值。

汇总表达式

表达式类似于根据值、字段名称、运算符和函数创建的公式。 汇总表达式与公式的不同之处在于，公式一次只能操作一条记录，而汇总表达式能够对记录组、记录集或记录集合进行操作。

注: 如果流包含数据库连接（通过“数据库源”节点），那么您只能创建汇总表达式。

将新的表达式创建为派生字段；要创建表达式，请使用表达式构建器中提供的数据库汇总函数。

有关表达式构建器的更多信息，请参阅《IBM SPSS Modeler 用户指南》(`ModelerUsersGuide.pdf`)。

请注意，由于汇总表达式按关键字段进行分组，因此**关键字段**与您创建的任何汇总表达式之间存在关联。

有效的汇总表达式可对汇总结果进行评估；以下是几个有效的汇总表达式以及对它们进行管理的规则的示例：

- 您可以使用标量函数将多个汇总函数组合在一起，以生成单个汇总结果。 例如：

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

- 汇总函数可对多个标量函数的结果执行操作。 例如：

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

汇总优化设置

在“优化”选项卡上，您可以指定以下内容。

键是连续的。 如果您知道输入中具有相同键值的所有记录都分为一组，那么可以选择此选项（例如，如果对关键字段上的输入进行排序）。 这样做有助于提高性能。

允许使用中位数和四分位数的近似值。 在 Analytic Server 中处理数据时，当前不支持顺序统计（中位数、第一个四分位数和第三个四分位数）。 如果您使用的是 Analytic Server，那么可以选中此复选框以便对这些统计量使用近似值，而不是通过对数据进行分级然后根据各个分级之间的分布为统计计算估计值。 缺省情况下，未选中此选项。

分级数。 仅当选中了**允许使用中位数和四分位数的近似值**复选框时，此选项才可用。 选择对统计量进行估计时使用的分级数；分级数会影响**最大误差百分比**。 缺省情况下，分级数为 1000，这对应于最大误差，即范围的 0.1%。

RFM “汇总”节点

通过近因、频率、货币 (RFM) “汇总”节点，您可以利用客户的历史记录事务处理数据，去除所有无用的数据，然后将他们的所有剩余事务处理数据合并到一行并以唯一的客户标识作为关键字，从而列出他们最后一次与您交易的时间（近因），交易的次数（频率）以及这些交易的总值（货币）。

继续执行任一汇总之前，应该花一些时间来清理数据，尤其要关注所有缺失值。

一旦使用“RFM 汇总”节点标识和变换数据之后，您可以使用“RFM 分析”节点执行进一步分析。 有关更多信息，请参阅主题 第 128 页的『RFM 分析节点』。

请注意，如果已通过“RFM 汇总”节点运行数据文件，那么数据文件将不会再具有任何目标值；因此，在利用它作为使用所有建模节点（如 C5.0 或 CHAID）进行进一步预测分析的输入之前，需要将其与其他客户的数据进行合并（例如，通过匹配用户标识）。 有关更多信息，请参阅主题 第 64 页的『合并节点』。

将 IBM SPSS Modeler 中的“RFM 汇总”节点和 RFM 分析节点设置为使用独立分级；即，它们分别按近因、频率、货币值对数据进行排序和分级，而无需考虑它们的值或其他两种标准。

为 RFM “汇总”节点设置选项

“RFM 汇总”节点的“设置”选项卡包含下列字段。

近因计算的相对日期 指定计算交易近因的日期。 该日期可以是您输入的 **固定日期**，也可以是系统设置的 **当前日期**。 **当前日期**由系统缺省输入，并在执行节点时自动更新。

注: 在不同的语言环境中，**固定日期**的显示可能有所不同。 例如，如果值 `2007-8-10` 在流中存储为 `Fri Aug 10 00:00:00 CST 2007`，那么这是时区“UTC+8”中的时间和日期。 但是，在时区“UTC-8”中显示为 `Thu Aug 9 12:00:00 EDT 2007`。

标识是连续的 如果您的数据进行了预先排序，以便所有具有同一标识的记录一起出现在数据流中，那么选择此选项可以加快处理速度。 如果您的数据未经预先排序（或者您不确定），请将此选项保持不选中状态，则该节点将自动对该数据进行排序。

标识 选择该字段以用来识别客户及其交易。 要显示用于选择的字段，请使用右侧的“字段选择器”按钮。

日期 选择将要用来计算近因的日期字段。 要显示用于选择的字段，请使用右侧的“字段选择器”按钮。

请注意，这需要具有适当格式的储存日期或时间戳记的字段以用作输入。 例如，如果您有一个值类似于 *Jan 2007*、*Feb 2007* 等的字符串字段，那么可以使用“过滤器”节点和 `to_date()` 函数将其转换为日期字段。 有关更多信息，请参阅主题 第 120 页的『使用填充节点进行存储类型转换』。

值 选择该字段以用来计算客户交易的总货币值。 要显示用于选择的字段，请使用右侧的“字段选择器”按钮。 注意：该值必须是一个数字值。

新字段名称扩展 选择该字段可将前缀或后缀（如“12_month”）追加到新生成的近因、频率和货币字段。 选择 **后缀** 或 **前缀** 可指明您首选的扩展样式。 例如，这在检查多个时间周期时将可能有用。

废弃具有以下值的记录 如有需要，可在计算 RFM 总计时指定一个最小值，凡低于该值的交易详细信息都不再被使用。 该值单元与所选的 **值** 字段相关。

仅包含最新交易 如果分析的是大型数据库，那么可以指定只使用最近的记录。 无论是在某个特定的日期之后还是在最近的周期内，您都可以选择使用记录的数据：

- **该交易日期之后** 指定交易日期以在分析时包含其之后的记录。
- **以下过去时间段内的交易** 指定从**近因计算的相对日期**开始之前的周期数和周期类型（天、周、月或年），在此日期之后的记录将被包含在您的分析中。

保存第二个最新交易的日期 如果希望了解每个客户第二个最近交易的日期，请选中此框。 此外，您还可以选择 **保存第三个最近交易的日期** 复选框。 例如，这样有助于您识别在很长一段时间之前进行许多交易的客户，但仅限于一个最近交易。

排序节点

您可以使用“排序”节点，根据一个或多个字段的值，按照升序或者降序对记录进行排序。 例如，“排序”节点经常用于查看和选择带有最常见数据值的记录。 通常情况下，您首先要使用“汇总”节点汇总数据，然后使用“排序”节点按照记录计数的降序对汇总后的数据进行排序。 如果在一个表中显示这些结果，您则可以探索这些数据并作出决策，如选择前 10 个最佳客户的记录。

“排序”节点的“设置”选项卡包含下列字段。

排序依据。 在表中显示所有选作排序关键的字段。 如果关键字段为数字字段，那么它最适用于排序。

- 使用右侧的“自动选择器”按钮向此列表**添加字段**。
- 通过单击表中顺序列中的**升序**或**降序**箭头来**选择顺序**。
- 使用红色的删除按钮来**删除字段**。
- **对指令排序**，此操作通过上下方向按钮完成。

缺省排序顺序。 选择**升序**或**降序**用作在上面添加新字段时的缺省排序次序。

注: 如果模型流的下游中存在“区分”节点，那么不会应用“排序”节点。 有关“区分”节点的信息，请参阅第 71 页的『“区分”节点』。

排序优化设置

如果您要对您知道已经按照某些关键字段排序的数据进行操作，那么可以指定哪些字段已经排序，从而使得系统能够更高效地对剩下的数据进行排序。 例如，您要按照 *Age* （降序）和 *Drug* （升序）进行排序，但知道这些数据已经按照 *Age* （降序）进行了排序。

已预先对数据进行排序。 指定数据是否已经按照一个或多个字段进行排序。

指定现有排序顺序。 指定已经排序的字段。 使用“选择字段”对话框，向列表添加字段。 在 顺序 列中，指定每个字段按升序还是降序排序。 如果指定多个字段，请确保按照正确排序顺序列出这些字段。 使用列表右侧的箭头可按照正确顺序排列这些字段。 如果指定现有的正确排序顺序时出错，则当您运行流时会出现一个错误，该错误将显示为一个记录编号，该编号即是出现排序与您所指定的顺序不一致的位置。

注意：启用并行处理有益于提高排序速度。

合并节点

“合并”节点的功能是采用多个输入记录，然后创建一个包含全部或其中部分输入字段的输出记录。 当您要合并来源不同的数据（如内部客户数据和购买的人口统计数据）时，此操作非常有用。 可以通过下列方式合并数据。

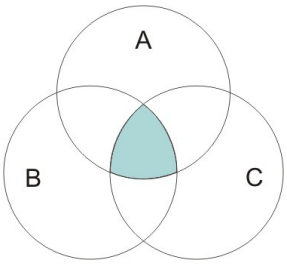
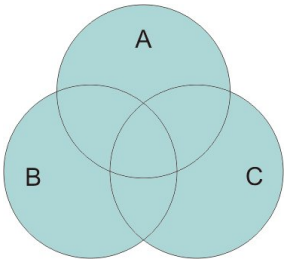
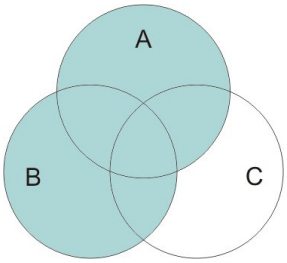
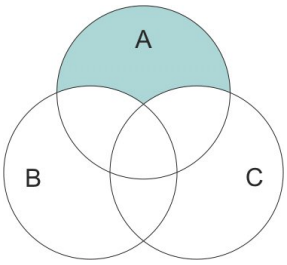
- 按**顺序**合并按输入顺序并置来自所有源的相应记录，直到穷尽最小的数据源为止。 如果使用此选项，务必确保已使用排序节点完成了对数据的排序。
- 使用**键**字段（例如客户标识）来指定如何使来自一个数据源的记录与来自其他数据源的记录匹配，从而进行合并。 连接的类型有许多，其中包括内部连接、完全外部连接、部分外部连接和反连接。 有关更多信息，请参阅主题 第 65 页的『连接类型』。
- 按**条件**合并意味着您可以指定执行合并前所需满足的条件。 可以在节点中直接指定条件，也可以使用“表达式构建器”构建条件。
- 按**排名式条件**合并是一个左侧外连接，在此连接中，您指定执行合并前所需满足的条件，并指定用于实现从低到高排序的排名表达式。 这种合并最常用于合并地理空间数据，您可以直接在节点中指定条件，也可以使用表达式构建器来构建条件。

连接类型

当数据合并使用一个关键字段时，最好先花一些时间来考虑要排除和包括哪些记录。 连接的类型有很多种，详细信息将在下面讨论。

两种基本的连接类型称为内部连接和外部连接。 这些方法经常用于根据关键字段（如 客户标识 ）的公共值，合并来自相关数据集的表。 通过内部连接，可以实现清理合并，以及仅包括完整记录的输出数据集。 外部连接也包括合并数据中的完整记录，但它们还允许包括来自一个或多个输入表的唯一性数据。

以下内容详细介绍了允许的连接类型。

	**内部连接** 只包括其中关键字段的值对于所有输入表都共有的记录。 即，不匹配的记录不会包括在输出数据集中。
	**完全外部连接**包括输入表中的所有记录，既有匹配的记录也有不匹配的记录。 左外部连接和右外部连接称为部分外部连接，将在下面描述。
	**部分外部连接** 包括使用关键字段匹配的所有记录，以及指定的表中的不匹配记录。 （换句话说，包括部分表中的所有记录，以及其他表中的仅匹配记录。） 使用“合并”选项卡上的“选择”按钮，可选择要包括在外部连接中的表（如此处显示的 A 和 B）。 如果只合并两个表，部分连接也称为左外部连接或右外部连接。 因为 IBM SPSS Modeler 允许合并两个以上的表，所以我们称此为部分外部连接。
	**反连接** 仅包括第一个输入表（此处显示的表 A）的不匹配记录。 这种连接类型与内部连接正好相反，在输出数据集中不包括完整记录。

例如，如果您在一个数据集 中包含有关农场的信息，在另一个数据集中包含农场相关的保险索赔信息，则可以使用合并选项将第一个源中的记录与第二个源相匹配。

要确定您农场样本中的客户是否已经提出了保险索赔，请使用内部连接选项返回一个列表，其中显示两个样本中所有标识匹配的记录。

	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...

图 *2:* 内部连接合并的输出示例

使用完全外部连接选项既会返回输入表中的匹配记录也会返回不匹配的记录。 对于任何不完整的值，都将使用系统缺失值 (`$null$`)。

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalu
1	id601	$null$	$null$	$null$	$null$	$null$	$null$	$null$	decomm...	74703.10
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

图 *3:* 完全外部连接合并的输出示例

部分外部连接包括使用关键字段匹配的所有记录，以及指定的表中的不匹配记录。 该表显示了标识字段中所有匹配的记录，以及第一个数据集中匹配的记录。

	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

图 *4:* 部分外部连接合并的输出示例

如果使用反连接选项，该表则只返回第一个输入表的不匹配记录。

	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

图 *5:* 反连接合并的输出示例

指定合并方法和关键字

“合并”节点的“合并”选项卡包含下列字段。

合并方法 请选择用于合并记录的方法。 选择**键**或**条件**将激活此对话框的下半部分。

- **顺序** 按顺序合并记录，以便将每个输入中的第 *n* 个记录合并到一起，从而生成第 *n* 个输出记录。 当任何记录用完匹配输入记录后，将不会再生成任何输出记录。 这意味着，创建的记录数是最小数据集中的记录数。
- **键** 使用键字段（例如交易标识）将键字段中的值相同的记录合并。 此选项等同于数据库的“相等连接”。 如果关键值出现多次，那么返回所有可能的组合。 例如，如果具有相同键字段值 *A* 的记录的其他字段中包含不同的值 *B*、*C* 和 *D*，那么合并后的字段对于 *A* 与值 *B*、*A* 与值 *C* 以及 *A* 与值 *D* 的每个组合都将生成一个单独的记录。

  注意：在按关键字合并的方法中，空值不会被视为相同的值，因此不会连接。

- **条件** 使用此选项可以指定合并条件。 有关更多信息，请参阅 第 67 页的『指定合并的条件』。
- **排名式条件** 使用此选项可以指定是否对主数据集与所有辅助数据集中的每一对行进行合并；使用排名表达式可以将任意的多个匹配项按从低到高顺序排列。 有关更多信息，请参阅 第 67 页的『为“合并”指定排名式条件』。

可能的键 仅列出那些在所有输入数据源中都有完全匹配的字段名称的字段。 从此列表中选择一个字段，并使用箭头按钮将其添加为用于合并记录的关键字段。 可以使用多个键字段。 您可以使用“过滤器”节点或者源节点的“过滤器”选项卡对不匹配的输入字段进行重命名。

用于合并的键 根据键字段的值，列出所有用于对来自所有输入数据源的记录进行合并的字段。 要从列表中移除关键字段，请选择一个关键字段，然后使用箭头按钮将其返回到“可能的关键字”列表中。 如果选择了多个关键字段，那么下面的选项将启用。

组合重复键字段 在上面选择了多个键字段之后，此选项确保只有一个具有该名称的输出字段。 缺省情况下，此选项为启用状态，但已从以前版本的 IBM SPSS Modeler 导入流的情况下除外。 如果禁用了此选项，那么必须使用“合并”节点对话框中的“过滤器”选项卡将重复的键字段重命名或排除。

仅包括匹配的记录（内连接） 选择此项将仅合并完整的记录。

包括匹配的记录和不匹配的记录（全外连接） 选择此项将执行“全外连接”。 这意味着，如果不存在所有输入表中所共有的关键字段值，那么将仍然保留不完整记录。 未定义的值 (`$null$`) 会添加到关键字段，并包括在输出记录中。

包括匹配的记录和选定的不匹配记录（部分外连接） 选择此项表示对子对话框中选择的表执行“部分外连接”。 单击 **选择** 可指定将在合并中为其保留不完整记录的表。

包括第一个数据集中与任何其他记录都不匹配的记录（反连接） 选择此项表示执行某种“反连接”，在此类连接中，只有第一个数据集中的不匹配记录将传递到下游。 您可以使用“输入”选项卡上的箭头指定输入数据集的顺序。 这种连接类型在输出数据集中不包括完整记录。 有关更多信息，请参阅 第 65 页的『连接类型』。

选择用于部分连接的数据

对于部分外部连接，您必须选择要为其保留不完整记录的表。 例如，您可能想保留 Customer 表中的所有记录，同时仅保留 Mortgage Loan 表中匹配的记录。

“外部连接”列。 在外部连接列中，选择要作为整体包括在内的数据集。 对于部分连接，重叠的记录以及此处选中的数据集的不完整记录都将被保留。 有关更多信息，请参阅主题 第 65 页的『连接类型』。

指定合并的条件

通过将合并方法指定为**条件**，可指定要执行合并必须满足的一项或多项条件。

可以直接在“条件”字段中输入条件，也可单击此字段右侧的计算器图标以借助表达式构建器构建条件。

向重复字段名添加标记以避免合并冲突 如果两个或两个以上要合并的数据集包含相同的字段名，那么通过选中此复选框，可以在字段列标题开头添加另一个前缀标记。 例如，如果存在两个名为 *Name* 的字段，那么合并结果将包含 *1_Name* 和 *2_Name*。 如果在数据源中将该标记重命名，那么将使用新名称，而不是使用进行了编号的前缀标记。 如果未选中此复选框，并且数据中存在重复的名称，那么此复选框右侧将显示警告。

为“合并”指定排名式条件

您可以将“已排名的条件”合并视为按条件进行的左侧外连接合并；此合并的左侧是主数据集，其中的每个记录都是一个事件。 例如，对于用来在罪案数据中查找模式的模型，主数据集中的每个记录都是一项罪案及其相关信息（地点和类型等等）。 在此示例中，右侧可能包含相关的地理空间数据集。

此合并同时使用合并条件和排名表达式。 合并条件可以使用 *within* 或 *close_to* 之类的地理空间函数。 在合并期间，右侧数据集中的所有字段都将添加到左侧数据集，但是多项匹配将产生列表字段。 例如：

- 左侧：罪案数据
- 右侧：城镇数据集和道路数据集
- 合并条件：城镇内 (*within*) 且邻近 (*close_to*) 道路的罪案数据，以及所谓“邻近”(*close_to*) 的定义。

在此示例中，如果罪案发生在三条道路（要返回的匹配项数至少设置为 3）的所需邻近 *close_to* 距离内，那么将全部三条道路作为列表项返回。

通过将合并方法设置为**排名式条件**，可以指定执行合并前必须满足的一项或多项条件。

主数据集 请选择用于合并的主数据集；所有其他数据集中的字段都将添加到您选择的数据集中。 您可以将其视为外连接合并的左侧。

您选择主数据集时，所有其他连接到“合并”节点的输入数据集都将自动列示在**合并**表中。

向重复字段名添加标记以避免合并冲突 如果两个或两个以上要合并的数据集包含相同的字段名，那么通过选中此复选框，可以在字段列标题开头添加另一个前缀标记。 例如，如果存在两个名为 *Name* 的字段，那么合并结果将包含 *1_Name* 和 *2_Name*。 如果在数据源中将该标记重命名，那么将使用新名称，而不是使用进行了编号的前缀标记。 如果未选中此复选框，并且数据中存在重复的名称，那么此复选框右侧将显示警告。

合并

数据集
显示作为输入连接到“合并”节点的辅助数据集的名称。 缺省情况下，存在多个辅助数据集时，这些数据集将按它们连接到“合并”节点的顺序列出。

合并条件

请输入用于将表中各个数据集与主数据集合并的唯一条件。 您可以直接在单元格中输入条件，也可以单击此单元格右侧的计算器图标以借助表达式构建器构建条件。 例如，您可以使用地理空间谓词来创建合并条件，用于将一个数据集中的罪案数据放入另一个数据集的城镇数据。 缺省合并条件取决于地理空间测量级别，如以下列表所示。

- “点”、“线串”、“多点”和“多线串”- *close_to* 的缺省条件。
- “多边形”和“多多边形”- *within* 的缺省条件。

有关这些级别的更多信息，请参阅第 105 页的『地理空间测量子级别』。

如果一个数据集包含多个不同类型的地理空间字段，那么使用的缺省条件取决于按以下降序在数据中找到的第一个测量级别。

- 点
- 线串
- 多边形

注: 仅当辅助数据库中存在地理空间数据字段时，缺省值才可用。

排名表达式

请指定一个表达式，用于对数据集的合并进行排名；此表达式将根据排名条件对多个匹配项进行排序。 您可以直接在单元格中输入条件，也可以单击此单元格右侧的计算器图标以借助表达式构建器构建条件。

距离和面积的缺省排名表达式在表达式构建器中提供，这两种表达式都从低到高排名，例如，这表示顶部的距离匹配是最小的值。 下面是按距离进行排名的一个示例：主数据集包含罪案及其相关地点，另外每个数据集都包含具有地点的对象；在这种情况下，罪案与对象之间的距离可以用作排名条件。 缺省排名表达式取决于地理空间测量级别，如以下列表所示。

- “点”、“线串”、“多点”和“多线串”- 缺省表达式为 *distance*。
- “多边形”和“多多边形”- 缺省表达式为 *area*。

注: 仅当辅助数据库中存在地理空间数据字段时，缺省值才可用。

匹配项数
根据条件和排名表达式，指定返回的匹配项数。 缺省匹配项数取决于辅助数据集中的地理空间测量级别，如以下列表所示；但是，您可以在单元格中双击以输入自己的值，最大为 100。

- “点”、“线串”、“多点”和“多线串”- 缺省值为 3。
- “多边形”和“多多边形”- 缺省值为 1。
- 不包含地理空间字段的数据集 - 缺省值为 1。

例如，如果您设置了基于**合并条件** *close_to* 和**排名表达式** *distance* 的合并，那么主数据集中的每个记录在辅助数据集中的前三个（最邻近）匹配项将作为结果列表字段中的值返回。

过滤合并节点中的字段

“合并”节点包括了一种用于过滤或重命名由于合并多个数据源而产生的重复字段的简便方法。 单击对话框中的 **过滤器** 选项卡可选择过滤选项。

此处显示的选项几乎与“过滤”节点的选项完全相同。 但还有其它选项在“过滤”菜单中存在，而此处没有讨论。 有关更多信息，请参阅主题 第 113 页的『过滤或重命名字段』。

字段。 显示当前连接的数据源中的输入字段。

标记。 列出与数据源链接相关联的标记名称（或编号）。 单击 **输入** 选项卡可更改到此合并节点的活动链接。

源节点。 显示要合并其数据的源节点。

已连接的节点。 显示与“合并”节点连接的节点的节点名称。 复杂的数据挖掘经常需要若干可能包括同一个源节点的合并或追加操作。 连接的节点名称提供了一种区分这些内容的方法。

过滤。 显示输入字段和输出字段之间的当前连接。 活动连接会显示一个未断开的箭头。 带有红色 X 的连接表示经过过滤的字段。

字段。 列出合并或追加之后的输出字段。 重复字段显示为红色。 单击上面的过滤字段可禁用重复的字段。

查看当前字段。 选择此选项可查看被选作关键字段的字段信息。

查看未使用的字段设置。 选择此选项可查看当前未使用的字段的相关信息。

设置输入顺序和标记

使用合并节点和追加节点对话框中的“输入”选项卡，可以指定输入数据源的顺序，还可以对每个源的标记名称进行任意更改。

输入数据集的标记和顺序。 选择此选项将只合并或追加完整的记录。

- **标记**。 列出每个输入数据源的当前标记名称。 标记名称（即 **标记** ）是一种唯一标识用于合并或追加操作的数据链接的方法。 例如，这就好像来自不同管道的水在一个点处进行合并，然后流到单个管道中。 IBM SPSS Modeler 中的数据也按照相似的方式流动，合并点通常是不同数据源之间的复杂交互。 标记提供了一种用于管理“合并”节点或“追加”节点的输入（“管道”）的方法，因此，如果保存或断开该节点，这些链接将被保留并可以轻松识别。

将附加数据源与“合并”节点或“追加”节点相连时，将使用编号自动创建缺省标记，以表示您连接这些节点的顺序。 此顺序与字段在输入或输出数据集中的顺序无关。 通过在 标记 列中输入新名称，可以更改缺省标记。

- **源节点**。 显示要合并其数据的源节点。
- **已连接的节点**。 显示与“合并”节点或“追加”节点连接的节点的节点名称。 复杂的数据挖掘经常需要若干可能包括同一个源节点的合并操作。 连接的节点名称提供了一种区分这些内容的方法。
- **字段**。 列出每个数据源中的字段数。

查看当前标记。 选择此选项可查看正在由“合并”节点或“追加”节点使用的活动标记。 换而言之，当前标记标识指向有数据流过的节点的链接。 用管道比喻一下，当前标记就相当于现在有水流过的管道。

查看未使用的标记设置。 选择此选项可查看以前用于连接“合并”节点或“追加”节点、但当前未与数据源连接的标记（或链接）。 这就相当于排水系统中仍然存在的空管道。 您可以选择将这些“管道”与新源连接，也可以选择将其移除。 要从节点删除未使用的标记，请单击 **清除** 。 此操作将马上清除所有未使用的标记。

合并优化设置

系统提供了两个选项，可帮助您在特定的情况下以更高效的方式合并数据。 通过这些选项，您可以在一个输入数据集明显大于其他数据集，或者您的数据已经按照将要用于合并的所有或部分关键字段进行排序的情况下优化合并。

注: 从此选项卡中进行的优化仅适用于 IBM SPSS Modeler 本机节点执行（即，“合并”节点不推送回到 SQL）。 优化设置不影响 SQL 生成。

一个输入数据集相对较大。 选择此选项可表明其中一个输入数据集比其他数据集大很多。 系统会在内存中缓存较小的数据集，然后在不缓存或不对其进行排序的情况下处理较大的数据集来执行合并。 您经常会对使用星形模式或相似方案设计的数据使用这种类型的连接，这种数据中会存在一个较大的共享数据中心表（例如事务处理数据）。 如果选择此选项，请单击 **选择** 指定该较大数据集。 请注意，您只能选择 一个 较大数据集。 下表汇总了哪些连接可以通过此方法进行优化。

表 *16:* 连接优化摘要

连接类型	**是否可针对大型输入数据集进行优化?**
Inner	是
Partial	如果较大数据集中没有不完整记录，是。
全功能	否
反连接	如果较大数据集是第一个输入，是。

所有输入都已按键字段进行排序。 选择此选项可表明输入数据已经按照将要用于合并的一个或多个关键字段进行排序。 请确保 所有 输入数据集均已排序。

指定现有排序顺序。 指定已经排序的字段。 使用“选择字段”对话框，向列表添加字段。 您可以仅从将要用于合并的键段（在“合并”选项卡中指定）中选择。 在 顺序 列中，指定每个字段按升序还是降序排序。 如果指定多个字段，请确保按照正确排序顺序列出这些字段。 使用列表右侧的箭头可按照正确顺序排列这些字段。 如果指定现有的正确排序顺序时出错，则当您运行流时会出现一个错误，该错误将显示为一个记录编号，该编号即是出现排序与您所指定的顺序不一致的位置。

根据数据库使用的排序方法是否区分大小写，当有一个或多个输入由数据库排序时，优化可能不会正常工作。 例如，如果有两个输入分别为区分大小写和不区分大小写，那么排序结果可能有所不同。 合并优化将导致使用记录排序后的顺序来处理记录。 因此，如果输入采用不同的排序方法来排序，则合并节点会报告错误，并显示排序不一致处的记录编号。 如果所有输入均来自相同源，或使用互容的排序方法来排序，则可以成功合并记录。

注意：启用并行处理有益于提高合并速度。

追加节点

您可以使用 Append 节点连接记录集。 “合并”节点将来源不同的记录连接在一起，而“追加”节点与之不同，它读取一个源中的所有记录并将其遍历到下游，直到再也没有更多的记录。 然后会使用与第一个输入（即主输入）相同的数据结构（记录数、字段数等）读取下一个源中的记录。 当主源的字段比另一输入源中的字段多时，对于任何不完整的值都会使用系统空字符串 (`$null$`)。

Append 节点对于合并相似结构的数据集非常有用，但对于结构不同的数据则没什么用处。 例如，您可能将不同时段的事务处理数据存储在了不同文件中，如三月份一个销售数据文件，四月份还有另外一个文件。 假设它们具有相同的结构（字段相同，顺序也相同），那么“追加”节点会将它们连接为一个较大的文件，然后您可以对该文件进行分析。

注意：要追加文件，字段测量级别必须相似。 例如，名义字段无法附加测量级别为连续的字段。

设置追加选项

字段匹配依据。 选择匹配要追加的字段时要使用的方法。

- **位置。** 选择此选项将根据字段在主数据源中的位置追加数据集。 使用此方法时，您的数据应该进行排序，以确保正确的追加。
- **名称。** 选择此选项将根据字段在输入数据集中的位置追加数据集。 同样，选择 **匹配大小写** 可在匹配字段名称时启用大小写的区分。

输出字段。 列出与 Append 节点相邻的源节点。 列表上的第一个节点为主输入源。 您可以通过单击列标题，对显示中的字段进行排序。 此排序并不真正对数据集中的字段进行重新排序。

包含以下来源的字段。 选择**仅主数据集**可根据主数据集中的字段生成输出字段。 主数据集是在“输入”选项卡上指定的第一个输入。 选择 **所有数据集** 可为所有数据集中的所有字段生成输出字段，而不管在所有输入数据集中是否存在匹配字段。

通过在字段中包含源数据集来标记记录。 选择此选项可向输出文件添加一个附加字段，该字段的值将指示每个记录的源数据集。 在文本字段中指定一个名称。 该缺省字段名为 输入 。

“区分”节点

必须移除数据集中的重复记录后才能开始数据挖掘。 例如，在某个市场营销数据库中，个人可能以不同的地址或公司信息多次出现。 您可以使用“区分”节点来查找或移除数据中的重复记录，或者根据一组重复记录创建单个组合记录。

要使用“区分”节点，您必须先定义一组键字段，用于确定何时将两个记录视为重复项。

如果您仅挑选了部分字段用作键字段，那么两个“重复”记录可能并非确实完全相同，这是因为它们的其余字段的值仍可能有所不同。 在这种情况下，您还可以定义在每组重复记录中应用的排序顺序。 此排序顺序使您能够进行微调，以确定要将哪个记录视为组中的第一个记录。 否则，会将所有重复项都视为可交换，并可能选中任意记录。 不会对记录的传入顺序加以考虑，因此，使用上游“排序”节点并无帮助（请参阅下文中的『在“区分”节点中进行记录排序』）。

方式。 指定是否创建组合记录，或者指定包括还是排除（废弃）第一条记录。

- **为每个组创建组合记录**。 提供一种对非数字字段进行汇总的方式。 选中此选项将使“组合”选项卡可用，您可以在该选项卡上指定如何创建组合记录。 有关更多信息，请参阅第 73 页的『区分组合设置』。
- **仅包括每个组中的第一个记录**。 选择每组重复记录中的第一个记录并废弃余下的记录。 第一个记录由下面定义的排序顺序确定，而不是由记录的传入顺序确定。
- **仅废弃每个组中的第一个记录**。 废弃每组重复记录中的第一个记录并选中余下的记录。 第一个记录由下面定义的排序顺序确定，而不是由记录的传入顺序确定。 此选项对于 找出 数据中的重复非常有帮助，因为您随后即可在流中检查这些内容。

关键分组字段。 列出用于确定记录是否相同的一个或多个字段。 您可以:

- 通过使用右侧的字段选择器按钮可向列表添加字段。
- 通过使用红色删除按钮从列表中删除字段。

组内记录的排序依据。 列出用于确定各个记录在每组重复项中的排序方式以及是按升序还是降序排序的字段。 您可以:

- 通过使用右侧的字段选择器按钮可向列表添加字段。
- 通过使用红色删除按钮从列表中删除字段。
- 如果您按多个字段排序，则使用上下按钮移动字段。

如果您已选择包括或排除每个组中的第一个记录，并且将哪个记录视为第一个记录对您而言十分重要，那么必须指定排序顺序。

缺省排序顺序。 指定缺省情况下各个记录是按排序键值的**升序**还是**降序**排列。

在“区分”节点中进行记录排序

如果一组重复项中的记录顺序至关重要，那么您必须使用“区分”节点中的**组内记录的排序依据**选项来指定顺序。 请勿依赖于上游“排序”节点。 请注意，不会对记录的传入顺序加以考虑 - 仅考虑此节点中指定的顺序。

如果未指定任何排序字段（或者指定的排序字段不够充分），那么每组重复项中的记录均为无序（或者不完全地进行了排序），而结果可能不可预测。

例如，假定存在大量与众多机器相关的日志记录。 此日志包含数据，例如:

表 *17:* 机器日志数据

时间戳记	**机器**	**温度**
17:00:22	机器 A	31

表 *17:* 机器日志数据 *(继续)*		
时间戳记	**机器**	**温度**
13:11:30	机器 B	26
16:49:59	机器 A	30
18:06:30	机器 X	32
16:17:33	机器 A	29 日
19:59:04	机器 C	35
19:20:55	机器 Y	34
15:36:14	机器 X	28
12:30:41	机器 Y	25
14:45:49	机器 C	27
19:42:00	机器 B	34
20:51:09	机器 Y	36
19:07:23	机器 X	33

要将记录数减少为每台机器的最新记录，请使用 Machine 作为键字段，并使用 Timestamp 作为排序字段（降序）。 输入顺序不会影响结果，这是因为，排序选择指定了对于给定机器应返回许多行中的哪些行，最终数据输出如下所示。

表 *18:* 排序后的机器日志数据		
时间戳记	**机器**	**温度**
17:00:22	机器 A	31
19:42:00	机器 B	34
19:59:04	机器 C	35
19:07:23	机器 X	33
20:51:09	机器 Y	36

区分优化设置

如果您正在处理的数据只有少量记录，或已排序，您可以优化处理数据的方式，以使 IBM SPSS Modeler 更有效地处理数据。

注意：如果您选择了**输入数据集具有少量区分关键字**，或使用节点的 SQL 生成，那么可能返回区分关键字值内的任何行；要控制区分关键字内返回的行，需要使用“设置”选项卡上**组内记录的排序依据**字段指定排序顺序。 只要您在“设置”选项卡上指定了排序顺序，则优化选项不会影响按照“区分”节点输出的结果。

输入数据集具有少量区分关键字。 如果您具有少量记录和/或少量键字段的唯一性值，请选择此选项。 这样做有助于提高性能。

输入数据集已通过在“设置”选项卡上分组字段并排序字段进行排序。 只有当您的数据已按“设置”选项卡上的**组内记录的排序依据**下面列出的所有字段进行排序，且数据的升序或降序排序方式相同，才能选择此选项。 这样做有助于提高性能。

禁用 SQL 生成。 选择此选项以禁用节点的 SQL 生成。

区分组合设置

如果要处理的数据包含多条记录（例如针对同一人员的多条记录），那么可以通过创建要处理的单个组合（或汇总）记录来优化数据处理方式。

注: 只有在您从“设置”选项卡中选择了**为每个组创建组合记录**的情况下，此选项卡才可用。

设置“组合”选项卡的选项

字段。 此列将以自然排序顺序显示数据模型中除关键字段以外的所有字段。 如果节点未连接，则不显示任何字段。 要按字段名称的字母顺序对行进行排序，请单击列标题。 可以按住 Shift 进行单击或者按住 Ctrl 进行单击以选择多个行。 此外，右键单击某个字段时，将显示一个菜单，您可以从中选择所有行，按字段名称或值的升序或降序对这些行进行排序，按度量或存储类型选择字段，或者选择某个值以自动将同一**值填充依据**条目添加到每个所选行。

值填充依据。 选择要用于**字段**的组合记录的值类型。 可用选项取决于字段类型。

- 对于数字范围字段，您可以从下列选项中进行选择:
  - 组中的第一条记录
  - 组中的最后一条记录
  - 总计
  - 平均值
  - 最小值
  - 最大值
  - 定制
- 对于时间或日期字段，您可以从下列选项中进行选择:
  - 组中的第一条记录
  - 组中的最后一条记录
  - 最早
  - 最新(R)
  - 定制
- 对于字符串或无类型字段，您可以从下列选项中进行选择:
  - 组中的第一条记录
  - 组中的最后一条记录
  - 第一个字母数字
  - 最后一个字母数字
  - 定制

在每种情况下，您都可以使用**定制**选项来更好地控制用于填充组合记录的值。 有关更多信息，请参阅第 73 页的『区分组合 - 定制选项卡』。

在字段中包含记录计数。 选择此选项可在每条输出记录中包括一个额外的字段，缺省情况下该字段名为 `Record_Count`。 此字段表明汇总了多少输入字段而形成了每个汇总记录。 要为此字段创建定制名称，请在编辑字段中输入内容。

区分组合 - 定制选项卡

您可以通过“定制填充”对话框更好地控制使用哪个值完成新的组合记录。 请注意，如果仅在“组合”选项卡上定制了单个字段行，那么必须在对数据进行实例化后使用此选项。

注: 只有您在“组合”选项卡上的**值填充依据**列中选择了定制值的情况下，此对话框才可用。

根据字段类型，您可以选择下列其中一个选项。

- **按频率选择**。 根据在数据记录中的出现频率选择值。

注: 对于类型为“连续”、“无类型”或“日期/时间”的字段，此选项不可用。

- **使用**。 选择“频率最高”或“频率最低”。
- **关系**。 如果有两个或两个以上的记录出现频率相同，请指定如何选择所需记录。 您可以从下列四个选项中进行选择：使用第一条记录、使用最后一条记录、使用频率最低的记录或使用频率最高的记录。

- **包含值 (T/F)**。 选择此选项可将字段转换为标志，该标志用于确定组中是否包含具有指定值的记录。 然后，可以从所选字段的值列表中选择**值**。

  注: 如果在“组合”选项卡上选择了多个字段行，那么此选项不可用

- **列表中的第一个匹配项**。 选择此选项可对要指定给组合记录的值划分优先级。 然后，可以从所选字段的项列表中选择某个**项**。

  注: 如果在“组合”选项卡上选择了多个字段行，那么此选项不可用

- **并置值**。 选择此选项可将组中的所有值连接成一个字符串，从而保留这些值。 必须指定要在每个值之间使用的定界符。

  注: 在您选择了一个或多个类型为“连续”、“无类型”或“日期/时间”的字段行的情况下，这是唯一可用的选项。

- **使用定界符**。 可以选择使用**空格**或**逗号**作为合并字符串中的定界符值。 或者，您可以在**其他**字段中输入自己的定界符值字符。

  注: 仅在选择了**连接值**选项的情况下可用。

“流式时间序列”节点

“流式时间序列”节点用于在一个步骤中对时间序列模型同时进行构建和评分。 针对每个目标字段构建了一个单独的时间序列模型，但是不会将模型块添加到已生成的模型选用板中，并且无法浏览模型信息。

时间序列数据建模法需要在每个测量之间有一致的区间，并由空行表示所有缺失值。 如果数据尚未满足此需求，那么将必须根据需要对值进行变换。

有关时间序列节点的其他注意事项有：

- 字段必须是数字。
- 不得将日期字段用作输入。
- 忽略分区。

“流式时间序列”节点对时间序列的指数平滑法模型、单变量自回归整合移动平均值 (ARIMA) 模型和多变量 ARIMA（或转换函数）模型进行估计，并根据时间序列数据产生预测。 另外，还提供了专家建模器，此建模器尝试针对一个或多个目标字段自动确定并估计最佳拟合 ARIMA 模型或指数平滑法模型。

有关时间序列建模的更多信息，请参阅《SPSS Modeler 建模节点指南》的“时间序列模型”部分。

通过 IBM SPSS Modeler Solution Publisher，使用 IBM SPSS 协作和部署服务 评分服务，支持在流式部署环境中使用“流式时间序列”节点。

“流式时间序列”节点 - 字段选项

使用预定义角色：此选项使用上游类型节点（或上游源节点的“类型”选项卡）的角色设置（目标、预测变量等等）。

使用定制字段分配。 要手动分配目标、预测变量和其他角色，请选中此选项。

注: 如果已对数据分区，那么选择**使用预定义角色**时将考虑分区，但选择**使用定制字段分配**时则不考虑。

字段。 使用箭头按钮可以从列表中将项目手动分配到屏幕右侧的各类角色字段。 图标表示每个角色字段的有效测量级别。

要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

目标 选择一个或多个字段作为预测目标。

候选输入 选择一个或多个字段作为预测输入。

事件和干预 使用此区域将特定输入字段指定为事件字段或干预字段。 此指定会将字段标识为包含可受事件（可预测的重现情况，例如促销）或干预（一次性事件，例如停电或员工罢工）所影响的时间序列数据。

“流式时间序列”节点 - 数据规范选项

通过“数据规范”选项卡，您可以设置用于将数据包含在模型中的所有选项。 只要同时指定**日期/时间字段**和**时间间隔**，便可以单击**运行**按钮来构建包含所有缺省选项的模型，但通常您会想要根据自己的用途定制构建。

该选项卡包含多个不同的窗口，您可以在其中设置特定于自己的模型的定制。

“流式时间序列”节点 - 观测值

使用此窗格中的设置可以指定用于定义观测值的字段。

由日期/时间字段指定的观测值

您可以指定观测值由日期、时间或时间戳记字段定义。 除了用于定义观测值的字段以外，请选择用于描述观测值的适当时间间隔。 根据指定的时间间隔不同，您还可以指定其他设置，例如观测值之间的间隔（增量）或者每周的天数。 对于时间间隔，注意事项如下所示：

- 如果各个观测值之间的时间间距不定期（例如处理销售订单的时间），请使用**不定期**值。 如果选择了**不定期**，那么必须从“数据指定项”选项卡上的**时间间隔**设置中指定用于分析的时间间隔。
- 如果观测值表示日期和时间，并且时间间隔为小时、分钟或秒，请使用**每天的小时数**、**每天的分钟数**或**每天的秒数**。 如果观测值表示时间（持续时间）并且未引用日期，而时间间隔为小时、分钟或秒，请使用**小时数（非周期性）**、**分钟数（非周期性）**或**秒数（非周期性）**。
- 根据选择的时间间隔，此过程可以检测缺失的观测值。 由于此过程假定所有观测值之间的时间间距相等，并假定未缺失观测值，因此有必要检测缺失的观测值。 例如，如果时间间隔为“天”，并且日期 2015-10-27 后跟 2015-10-29，那么表示缺失 2015-10-28 的观测值。 对于任何缺失观测值，将插补值；使用“数据规范”选项卡的**缺失值处理**区域可以指定用于处理缺失值的设置。
- 指定的时间间隔使此过程能够检测到同一时间间隔内的多个需要汇总到一起的观测值并使各个观测值在时间间隔边界（例如每个月的第一天）处对齐，以确保各个观测值之间的间距相等。 例如，如果时间间隔为“月”，那么同一个月内的多个日期将汇总到一起。 此类汇总称为分组。 缺省情况下，进行分组时，将计算观测值的总和。 通过“数据指定项”选项卡上的**汇总和分布**设置，您可以指定另一种分组方法，例如计算各个观测值的平均值。
- 对于某些时间间隔，附加设置可以定义正常等间距时间间隔中的中断。 例如，如果时间间隔为“天”，但只有工作日有效，那么您可以指定每周有 5 天，并且每周从星期一开始。

定义为周期或循环周期的观测值

观测值可以由一个或多个表示周期或周期反复循环（直至达到任意数目的循环级别为止）的整数字段定义。 借助此结构，您可以描述任何标准时间间隔都无法支持的观测值序列。 例如，可以使用表示年份的循环字段和表示月份的周期字段（一个循环的长度为 10）来描述仅包含 10 个月的财年。

用于指定循环周期的字段定义了周期性级别的层次结构，在此层次结构中，最低级别由**周期**字段定义。 次高级别由级别为 1 的循环字段指定，接着由级别为 2 的循环字段指定，依此类推。 除最高级别以外，每个级别的字段值对于次高级别都必须具有周期性。 最高级别的值不得具有周期性。 例如，对于由 10 个月组成的财年，月在年中具有周期性，而年不具有周期性。

- 在特定级别，循环长度是次低级别的周期长度。 在财年示例中，只有一个循环级别，并且循环长度为 10，这是因为次低级别表示月，而指定的财年包含 10 个月。
- 指定不从 1 开始的任何周期字段的起始值。 此设置是检测缺失值所必需的。 例如，如果周期性字段起始于 2，但起始值指定为 1，那么此过程将假定该字段的每个循环中的第一个周期都有一个缺失值。

“流式时间序列”节点 - 分析时间间隔

用于分析的时间间隔可以与观测值的时间间隔不同。 例如，观测值的时间间隔为“天”时，您可以选择“月”用作进行分析的时间间隔。 系统先将每日数据汇总为每月数据，然后再构建模型。 您还可以选择将时间间隔

较长的数据分布到较短的时间间隔内。 例如，如果观测值是按季度的，那么您可以将数据从季度分发到月度数据。

使用此窗格中的设置指定用于分析的时间间隔。 汇总或分布数据的方法是在“数据指定项”选项卡上的**汇总和分布**设置中指定的。

执行分析所采用的时间间隔的可用选项取决于定义观测值的方式以及这些观测值的时间间隔。 特别是，如果观测值由循环周期定义，那么仅支持汇总。 在此情况下，分析的时间间隔必须大于或等于观测值的时间间隔。

“流式时间序列”节点 - 汇总和分布选项

使用此窗格中的设置可以指定用于对观测值的时间间隔的相关输入数据进行汇总或分布的设置。

汇总函数
如果用于分析的时间间隔比观测值的时间间隔长，那么将对输入数据进行汇总。 例如，当观测值的时间间隔为“天”并且分析时间间隔为“月”时，将执行汇总。 可用的汇总函数如下所示：mean、sum、mode、min 或 max。

分布函数
如果用于分析的时间间隔比观测值的时间间隔短，那么将对输入数据进行分布。 例如，当观测值的时间间隔为“季度”并且分析时间间隔为“月”时，将执行分布。 可用的分布函数如下所示：mean 或 sum。

分组函数
当观测值由日期/时间定义，并且同一个时间间隔内存在多个观测值时，将进行分组。 例如，如果观测值的时间间隔为“月”，那么同一个月的多个日期将分组到一起，并与它们所在的月份相关联。 以下是可用的分组函数：mean、sum、mode、min 或 max。当观测值由日期/时间定义，并且观测值的时间间隔指定为“不定期”时，将始终执行分组。

注: 尽管分组是一种汇总形式，但在对缺失值进行任何处理之前执行，而正式的汇总是在对所有缺失值进行处理之后执行。 如果观测值的时间间隔指定为“不定期”，那么将仅使用分组函数来执行汇总。

将跨天观测值汇总到前一天
指定是否将时间跨天边界的观测值汇总到前一天的值。 例如，对于在 20:00 开始的 8 小时一天的每小时观测值，此设置指定是否将介于 00:00 与 04:00 之间的观测值包含在前一天的汇总结果中。 仅当观测值的时间间隔为“每天的小时数”、“每天的分钟数”或“每天的秒数”，并且分析时间间隔为“天”时，此设置才适用。

所指定字段的定制设置
您可以对每个字段指定汇总函数、分布函数和分组函数。 这些设置将覆盖汇总函数、分布函数和分组函数的缺省设置。

“流式时间序列”节点 - 缺失值选项

使用此窗格中的设置可以指定输入数据中要替换为插补值的缺失值数。 提供了下列替换方法：

线性插值
使用线性插值替换缺失值。 缺失值之前的最后一个有效值以及之后的第一个有效值用于插值。 如果序列中的第一个或最后一个观测值具有缺失值，那么将使用序列开头或末尾的两个最近邻非缺失值。

使用线性插值替换缺失值。

- 对于非季节数据，缺失值之前的最后一个有效值以及之后的第一个有效值用于插值。 如果缺失值位于时间序列的开始或结束位置，那么基于两个最近的有效值使用线性推断方法。
- 对于季节数据，使用缺失值之前相同时间段的最后一个有效值和缺失值之后相同时间段的第一个有效值，线性内插缺失值。 如果无法针对缺失值找到相同时间段的两个值之一，那么数据将被视为非季节数据，并且使用非季节数据的线性插值来插补缺失值。

序列平均值
将缺失值替换为整个序列的平均值。

邻近点的平均值
使用有效周围值的均值替换缺失值。 邻近点的跨度为缺失值前后用于计算平均值的有效值数目。

邻近点的中位数
使用有效周围值的中位数替换缺失值。 邻近点的跨度为缺失值前后用于计算中位数的有效值数目。

线性趋势
此选项使用序列中的所有非缺失观测值来拟合简单线性回归模型，该模型随后用于插补缺失值。

其他设置：

最低数据质量评分 (%)
针对时间变量以及对应于每个时间序列的输入数据计算数据质量度量。 如果数据质量评分低于此阈值，那么将丢弃对应的时间序列。

“流式时间序列”节点 - 估计期

在“估计期”窗格中，您可以指定要在模型估计中使用的记录的范围。 缺省情况下，估计期从所有序列中的最早观测值的时间开始，并以最晚观测值的时间结束。

按开始时间和结束时间
您可以同时指定估计期的开始时间和结束时间，也可以仅指定开始时间或者仅指定结束时间。 如果省略了估计期的开始时间或结束时间，那么将使用缺省值。

- 如果观测值由日期/时间字段定义，请以用于该日期/时间字段的格式输入开始时间值和结束时间值。
- 对于由循环周期定义的观测值，请对每个循环周期字段指定值。 每个字段都将显示在单独的列中。

按最晚或最早时间间隔

将估计期定义为指定数目的时间间隔，这些时间间隔从数据中的最早时间间隔开始或以最晚时间间隔结束，并具有可选偏移量。 在此上下文中，时间间隔是指分析时间间隔。 例如，假定观测值按月获取，但分析时间间隔为季度。 指定**最晚**并对**时间间隔数目**指定值 24 表示最晚的 24 个季度。

您可以选择性地排除指定数目的时间间隔。 例如，指定最晚的 24 个时间间隔并对排除数目指定 1 表示估计期由最后一个时间间隔之前的 24 个时间间隔组成。

“流式时间序列”节点 - 构建选项

通过“数据规范”选项卡，您可以设置用于构建模型的所有选项。 当然，您只需单击**运行**按钮，即可采用所有缺省选项来构建模型；不过，通常您需要根据具体用途定制构建选项。

此选项卡包含两种不同的窗格，您可以在这些窗格中设置特定于模型的定制内容。

“流式时间序列”节点 - 常规构建选项

此窗格中的可用选项取决于您从**方法**列表中选择以下三项设置中的哪一项：

- **专家建模器**。 选择此选项以使用“专家建模器”，此组件将自动为每个相依序列查找拟合度最高的模型。
- **指数平滑法**。 使用此选项可指定定制的指数平滑法模型。
- **ARIMA**。 使用此选项可指定定制的 ARIMA 模型。

专家建模器

在**模型类型**下，选择您要构建的模型的类型：

- **所有模型**。 专家建模器同时考虑 ARIMA 模型和指数平滑法模型。
- **仅指数平滑法模型**。 “专家建模器”仅考虑指数平滑法模型。
- **仅 ARIMA 模型**。 “专家建模器”仅考虑 ARIMA 模型。

专家建模器考虑季节性模型。 只有在为活动数据集定义了周期性时才启用此选项。 选中此选项时，Expert Modeler 将同时考虑季节模型和非季节模型。 如果未选择此选项，那么专家建模器将仅考虑非季节性模型。

专家建模器考虑复杂的指数平滑法模型。 选择了此选项时，“专家建模器”将搜索所有 13 个指数平滑法模型（其中 7 个存在于原始时间序列节点中，而剩下 6 个是 V18.1 中新增的节点）。 如果 未选择此选项，那么“专家建模器”将搜索原始的 7 个指数平滑法模型。

在**离群值**下，从以下选项中进行选择

自动检测离群值。 缺省情况下，不自动检测离群值。 选中此选项以执行离群值自动检测，然后选择所需的离群值类型。

输入字段必须具有标志、名义或有序测量级别，并且必须是数字（例如，对于标志字段，必须为 1/0，而非 True/False），才能包含在此列表中。

对于在**字段**选项卡上标识为事件字段或干预字段的输入，Expert Modeler 仅考虑简单回归而不是任意变换函数。

指数平滑法

模型类型。 指数平滑法模型分类为季节性模型或非季节性模型。[1] 仅当使用“数据规范”选项卡上的“时间间隔”窗格定义的周期性为季节性时，季节性模型才可用。 季节性周期如下：循环周期、年、季度、月、每周的天数、每天的小时数、每天的分钟数以及每天的秒数。 提供了以下模型类型：

- **简式**。 此模型适合于没有趋势或季节性的序列。 其唯一的相关平滑参数是水平。 简单的指数平滑法非常类似于自回归阶数为 0、差分阶数为 1、移动平均值阶数为 1 且没有常量的 ARIMA 模型。
- **Holt 线性趋势**。 此模型适合于其中有线性趋势但没有季节性的序列。 其相关的平滑参数是水平和趋势，并且在此模型中，这些参数的值不会彼此限制。 Holt's 模型比 Brown's 模型更加常用，但在计算大型序列的估计值时会花费更多的时间。 霍特指数平滑非常类似于自回归阶数为 0、差分阶数为 2 而且移动平均值阶数为 2 的 ARIMA 模型。
- **阻尼趋势**。 此模型适合于具有逐渐消失的线性趋势但没有季节性的序列。 其相关的平滑参数是水平、趋势和阻尼趋势。 阻尼指数平滑模型非常类似于自回归阶数为一、差分阶数为一且移动平均值阶数为二的 ARIMA 模型。
- **乘法趋势**。 该模型适合于具有一种随序列量级而变的趋势且没有季节性的序列。 其相关的平滑参数是水平和趋势。 乘性趋势指数平滑与任何 ARIMA 模型都不相似。
- **Brown 线性趋势**。 此模型适合于其中有线性趋势但没有季节性的序列。 其相关的平滑参数是水平和趋势，但在此模型中，这些参数的值假设相等。 因此，Brown 模型是 Holt 模型的特例。 布朗指数平滑法非常类似于自回归阶数为 0、差分阶数为 2 且移动平均值阶数为 2 的 ARIMA 模型，其第二阶移动平均值的系数等于第一阶系数一半的平方。
- **简单季节性**。 此模型适合于没有趋势且季节效应不随时间变化的序列。 其相关的平滑参数是水平和季节。 季节指数平滑模型非常类似于自回归阶数为零、差分阶数为一、季节差分阶数为一且移动平均值阶数为 1、*p* 和 *p*+1 的 ARIMA 模型，其中 *p* 是一个季节区间中的周期数。 对于以月为时间单位的数据， *p* = 12。
- **Winters 加法**。 此模型适合于具有线性趋势且季节效应不随时间变化的序列。 其相关的平滑参数是水平、趋势和季节。 Winters 加法指数平滑模型非常类似于自回归阶数为零、差分阶数为一、季节差分阶数为一且移动平均值阶数为 *p*+1 的 ARIMA 模型，其中 *p* 是一个季节区间中的周期数。 对于以月为时间单位的数据， *p* = 12。
- **具有加性季节性的阻尼趋势**。 此模型适合于具有逐渐消退的线性趋势且季节效应不随时间变化的序列。 其相关的平滑参数是水平、趋势、阻尼趋势和季节。 阻尼趋势和加性季节性指数平滑与任何 ARIMA 模型都不相似。
- **具有加性季节性的乘法趋势**。 该模型适合于具有随序列量级而变的趋势且季节效应不随时间变化的序列。 其相关的平滑参数是水平、趋势和季节。 乘性趋势和加性季节性指数平滑与任何 ARIMA 模型都不相似。
- **乘法季节性**。 此模型适合于不具有趋势且季节效应随序列量级而变的序列。 其相关的平滑参数是水平和季节。 乘性季节性指数平滑与任何 ARIMA 模型都不相似。
- **Winters 乘法**。 此模型适合于具有线性趋势且季节效应随序列的大小而变化的序列。 其相关的平滑参数是水平、趋势和季节。 Winters 的可乘指数平滑法与任何 ARIMA 模型都不相似。
- **具有乘法季节性的阻尼趋势**。 此模型适合于具有逐渐消退的线性趋势且季节效应随序列的大小而变化的序列。 其相关的平滑参数是水平、趋势、阻尼趋势和季节。 阻尼趋势和乘性季节性指数平滑与任何 ARIMA 模型都不相似。
- **具有乘法季节性的乘法趋势**。 此模型适合于具有随序列的量级发生变化的趋势和季节效应的序列。 其相关的平滑参数是水平、趋势和季节。 乘性趋势和乘性季节性指数平滑与任何 ARIMA 模型都不相似。

目标转换。 您可以指定在对每个因变量建模前对其执行的变换。

[1] Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

- **无**。 未执行变换。
- **平方根**。 将执行平方根变换。
- **自然对数**。 执行自然对数变换。

ARIMA

指定定制 ARIMA 模型的结构。

ARIMA 阶数。 在网格的相应单元格中，输入模型的各个 ARIMA 成分的值。 所有的值都必须是非负整数。 对于自回归和移动平均值组件来说，该值表示最大阶数。 所有较低的正阶数都将包括在模型中。 例如，如果指定 2，那么模型包含顺序 2 和 1 。 只有在为活动数据集定义了周期性的情况下，才会启用季节列中的单元格。

- **自回归的 (p)**。 模型中的自回归阶数。 自回归阶数指定序列中哪些以前的值用于预测当前值。 例如，自回归阶数 2 指定序列中过去两个时间段的值用于预测当前值。
- **差分 (d)**。 指定在估计模型之前应用于序列的差分的阶。 当趋势出现时（具有趋势的序列通常是不稳定的，而 ARIMA 建模时假定是稳定的），差分是必需的并可用于去除这些趋势的影响。 差分阶数对应于序列趋势的程度；第一阶差分表示线性趋势，第二阶差分表示二次趋势，依此类推。
- **移动平均值 (q)**。 模型中移动平均值阶数的值。 移动平均值阶数指定如何使用与序列以前值均值之间的偏差来预测当前值。 例如，移动平均值阶数 1 和 2 指定在预测序列的当前值时，可考虑与序列（来自过去两个时限中的每一个）均值之间的偏差。

季节性。 季节性自回归成分、移动平均值成分和差分成分与其非季节性对应成分起着相同的作用。 但是，对于季节阶数，当前的序列值会受到由一个或多个季节周期分隔的序列值的影响。 例如，对于月度数据（季节性周期为 12），季节性阶 1 表示当前序列值将受到当前周期之前 12 个周期的序列值的影响。 因此，对于以月为时间单位数据，将季节阶数指定为 1 相当于将非季节阶数指定为 12。

自动检测离群值。 选中此选项可以对离群值执行自动检测，并选择可用的一个或多个离群值类型。

要检测的离群值的类型。 选择要检测的离群值类型。 支持的类型有：

- 加性（缺省值）
- 水平变动（缺省值）
- 创新的
- 瞬时的
- 季节加性
- 局部趋势
- 可加的修补

转换函数顺序和变换。 要指定变换并为 ARIMA 模型中的任何或所有输入字段定义转换函数，请单击**设置**；这将显示另一个对话框，您可以在其中输入转换和变换详细信息。

在模型中包含常量。 除非您确定整体平均序列值为 0，否则包含常量是标准。 如果应用差分，那么建议排除常量。

转换函数和变换函数

使用“转换函数顺序和变换”对话框可以指定变换以及为 ARIMA 模型中的任何或所有输入字段定义转换函数。

目标变换。 在此窗格中，您可以指定对每个目标变量进行建模之前要对其执行的变换。

- **无**。 未执行变换。
- **平方根**。 将执行平方根变换。
- **自然对数**。 执行自然对数变换。

候选输入转换函数和变换。 通过使用转换函数，您可指定以何种方式使用输入字段的过去值来预测目标序列的未来值。 左侧窗格的列表中显示了所有的输入字段。 此窗格中的其余信息特定于您选择的输入字段。

转换函数的阶数。 将转换函数的各个组件的值输入到**结构** 网格的相应单元格中。 所有的值都必须是非负整数。 对于分子和分母组件来说，该值表示最大阶数。 所有较低的正阶数都将包括在模型中。 此外，对于分子组件通常会包括阶数 0。 例如，如果为分子指定 2，那么模型包含阶数 2、1 和 0。 如果为分母指定 3，那么模型包含阶数 3、2 和 1。 只有在为活动数据集定义了周期性的情况下，才会启用季节列中的单元格。

分子。 转换函数的分子阶数指定选定的独立（预测变量）序列中哪些先前的值用于预测相依序列的当前值。 例如，分子阶数 1 指定使用过去某个时间段的独立序列的值（以及独立序列的当前值）来预测每个相依序列的当前值。

分母。 转换函数的分母阶数指定如何使用与选定独立（预测变量）序列的先前值均值之间的偏差来预测相依序列的当前值。 例如，分母阶数 1 指定在预测每个相依序列的当前值时，需要考虑与过去一个时间周期的独立序列的均值偏差。

差分。 指定在估计模型之前应用于所选独立（预测变量）序列的差分的阶数。 当趋势出现时，差分是必需的并可用于去除这些趋势的影响。

季节性。 季节性分子、分母和差分成分与其非季节性对应成分起着相同的作用。 但是，对于季节阶数，当前的序列值会受到由一个或多个季节周期分隔的序列值的影响。 例如，对于以月为时间单位的数据（季节周期为 12），季节阶数 1 表示当前序列值会受到当前序列之前的 12 个周期内的序列值的影响。 因此，对于以月为时间单位数据，将季节阶数指定为 1 相当于将非季节阶数指定为 12。

延迟。 设置延迟会将输入字段的影响延迟，延迟的时间为指定的时间间隔数。 例如，如果延迟设置为 5，那么输入字段在时间 *t* 不会产生影响，直到此后五个时限后 (*t* + 5) 才会对预测产生影响。

变换。 为一组自变量指定的转换函数还包括要对这些变量执行的可选变换。

- **无**。 未执行变换。
- **平方根**。 将执行平方根变换。
- **自然对数**。 执行自然对数变换。

“流式时间序列”节点 - 模型选项

置信限制宽度 (%)。 为模型预测和残差自相关计算置信区间。 您可以指定任何小于 100 的正数值。 缺省情况下，将使用 95% 置信区间。

将记录扩展至将来选项用于设置时间间隔数目，以预测估计期结束之后的情况。 在这种情况下，时间间隔为您在“数据指定项”选项卡上指定的分析时间间隔。 请求进行预测时，将为所有并非同时作为目标的输入序列自动构建自回归模型。 然后，使用这些模型生成这些输入序列在预测期内的值。 此设置没有最大限制。

要在预测中使用的未来值

- **计算输入的未来值** 如果选择此选项，那么会自动计算预测变量、噪声预测、差异估算和未来时间值的预测值。 请求进行预测时，将为所有并非同时作为目标的输入序列自动构建自回归模型。 然后，使用这些模型生成这些输入序列在预测期内的值。
- **选择要将其值添加到数据中的字段**。 对于要预测的每条记录（不包括保留值），如果您使用的是预测变量字段（角色设置为 `Input`），那么可以为每个预测变量指定预测周期的估计值。 您可以手动指定值，也可以从列表中选择值。
  - **字段**。 单击“字段选择器”按钮并选择可用作预测变量的任何字段。 请注意，在此选择的字段可能用于建模，也可能不用于建模；要将某个字段实际用作预测变量，必须在某个下游建模节点中选择该字段。 此对话框为您简单提供了指定未来值的便捷环境，这样可以在多个下游建模节点之间共享这些值，而无需在每个节点中单独进行指定。 另请注意，可用字段的列表可能会受到“构建选项”选项卡中选项的约束。

    请注意，如果为流中不再可用（因为已将其删除或由于“构建选项”选项卡中的选项更新）的字段指定了未来值，那么此字段将显示为红色。
  - **值**。 对于每个字段，可以在函数列表中进行选择，也可以单击**指定**手动输入值或从预定义值列表中选择值。 如果预测变量字段与您所控制的项目或其他预先可知的项目相关，则应手动输入值。 例如，如果要根据房间预订数目预测饭店下个月的收入，可以指定在该期间实际具有的预订数目。 相反，如果预测变量字段与您无法控制的某些因素（如股票价格）相关，那么可以使用函数，如“最近值”或“最近点的均数”。

可用的函数取决于字段的测量级别。

表 *19:* 可用于测量级别的函数

测量级别	**函数**
连续或名义字段。	空 最近点的均数 最近的值 指定
标志字段	空 最近的值 True False 指定

最近点的均数根据最后三个数据点的均数计算未来值。

最近值将未来值设置为最近数据点的值。

真/假将标志字段的未来值设置为指定的真值或假值。

指定打开一个对话框，用于手动指定未来值或从预定义列表中选择未来值。

使其可用于评分

您可以在此设置模型块的对话框中显示的评分选项的缺省值。

- **计算置信度的上限和下限**。 如果选择了此选项，那么对于每个目标字段，将为置信区间上限和下限创建新字段（带有缺省前缀 $TSLCI- 和 $TSUCI-）。
- **计算噪声残差**。 如果选中了此选项，那么对于每个目标字段，此选项将为模型残差创建新字段（带有缺省前缀 $TSResidual-），并同时创建这些值的总计。

模型设置

要在输出中显示的最大模型数。 指定您要包含在输出中的最大模型数。 请注意，如果构建的模型数超过了此阈值，那么模型不会显示在输出中，但它们仍可用于评分。 缺省值为 10。 显示大量模型可能会导致性能不佳或不稳定。

SMOTE 节点

合成少数类过采样技术 (Synthetic Minority Over-sampling Technique, SMOTE) 节点提供了用于处理不平衡数据集的过采样算法。 它提供了用于均衡数据的高级方法。 SMOTE 过程节点使用 Python 进行实现并且需要 `imbalanced-learn`© Python 库。 有关 imbalanced-learn 库的详细信息，请参阅 https://imbalanced-learn.org/stable/[1]。

节点选用板上的 Python 选项卡包含 SMOTE 节点和其他 Python 节点。

[1]Lemaître, Nogueira, Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research*，卷 18，第 17 号，2017 年，第 1-5 页。 (http://jmlr.org/papers/v18/16-365.html)

SMOTE 节点设置

在 SMOTE 节点的**设置**选项卡上定义下列设置。

目标设置

目标字段。 选择目标字段。 支持所有“标志”、“名义”、“有序”和“独立”测量类型。 如果在“分区”部分中选择了**使用分区数据**选项，那么将对训练数据进行过采样。

过采样比率

选择**自动**以自动选择过采样比率，或者选择**设置比率（少数对多数）**以设置定制比率值。 此比率是少数类中的样本数与多数类中的样本数之比。 此值必须大于 **0** 并小于或等于 **1**。

随机种子

设置随机种子值。 选择此信息并单击**生成**可以生成由随机数字生成器使用的种子。

方法

算法种类。 选择您要使用的 SMOTE 算法的类型。

样本规则

K 近邻。 指定要用于构建合成样本的最近邻居的数量

M 近邻。 指定要用于确定是否少数样本处于危险状态的最近邻居的数量。 仅当选择 **Borderline1** 或 **Borderline2** SMOTE 算法类型时，才会使用此选项。

分区

使用分区数据。 如果您仅希望对训练数据进行过采样，请选择此选项。

此 SMOTE 节点需要 `imbalanced-learn`© Python 库。 下表显示 SPSS Modeler SMOTE 节点对话框中的设置和 Python 算法之间的关系。

表 *20:* 映射到 *Python* 库参数的节点属性

SPSS Modeler 设置	脚本名称（属性名称）	Python API 的参数名称
过采样比率（数字输入控制）	`sample_ratio_value`	`ratio`
随机种子(D)	`random_seed`	`random_state`
K 邻居	`k_neighbours`	`k`
M 邻居	`m_neighbours`	`m`
算法种类	`algorithm_kind`	`kind`

“扩展变换”节点

借助“扩展变换”节点，您可以使用 R 脚本编写或 Python for Spark 脚本编写从 IBM SPSS Modeler 流中获取数据并对该数据应用变换。 完成修改后，数据将返回到流中以进行进一步处理、模型构建和模型评分。 “扩展变换”节点支持使用以 R 或 Python for Spark 编写的算法来变换数据，并使用户能够开发针对特定问题进行定制的数据变换方法。

要将此节点与 R 配合使用，您必须安装 IBM SPSS Modeler - Essentials for R。 请参阅 *IBM SPSS Modeler - Essentials for R: Installation Instructions* 以获取安装指示信息和兼容性信息。 另外，还必须在计算机上安装 R 的兼容版本。

“扩展变换”节点 -“语法”选项卡

选择语法类型 - **R** 或 **Python for Spark**。 请参阅以下部分以获取更多信息。 语法就绪时，您可以单击**运行**来执行“扩展变换”节点。

R 语法

R 语法。 您可以在此字段中输入或粘贴用于数据分析的定制 R 脚本语法。

转换标志字段。 指定标志字段的处理方式。 共有两个选项：**将字符串转换为因子，将整数和实数转换为双精度数**和**逻辑值（True 和 False）**。 如果选择**逻辑值（True 和 False）**，那么标志字段的原始值将丢失。 例如，如果某个字段的值为 Male 和 Female，那么这些值将更改为 True 和 False。

将缺失值转换为 R "不可用" 值 (NA)。 选中时，任何缺失值都将转换为 R NA 值。 R 使用值 NA 来标识缺失值。 您使用的某些 R 函数可能有一个参数，可用于控制当数据包含 NA 时函数的行为方式。 例如，该函数可能会允许您选择自动排除包含 NA 的记录。 如果未选择此选项，那么所有缺失值都将按原样传递到 R，并可能导致执行 R 脚本时发生错误。

将日期/时间字段转换为特殊时区控制的 R 类。 如果选择此选项，那么会将带有日期或日期时间格式的变量转换为 R 日期/时间对象。 必须选择下列选项之一：

- 将具有日期或日期时间格式的 **RPOSIXct.** 变量将转换为 R `POSIXct` 对象。
- **R POSIXlt （列表）**。 将具有日期或日期时间格式的变量转换为 R `POSIXlt` 对象。

注: POSIX 格式是高级选项。 仅当您的 R 脚本指定以需要这些格式的方式处理日期时间字段时才使用这些选项。 POSIX 格式不适用于具有时间格式的变量。

Python 语法

Python 语法。 您可以针对数据分析向此字段中输入或粘贴定制的 Python 脚本语法。 有关 Python for Spark 的更多信息，请参阅 Python for Spark 和 Python for Spark 的脚本编制。

“扩展变换”节点 -“扩展输出”选项卡

控制台输出选项卡包含当“语法”选项卡上的 R 脚本或 Python for Spark 脚本运行时接收到的任何输出（例如，如果使用 R 脚本，当执行**语法**选项卡上的 **R 语法**字段中的 R 脚本时，它显示从 R 控制台接收到的输出）。 此输出可能包括执行 R 或 Python 脚本时生成的 R 或 Python 错误消息或警告。 输出可主要用于调试脚本。 **控制台输出**选项卡还包含 **R 语法**或 **Python 语法**字段中的脚本。

每次执行“扩展变换”脚本时，都会使用从 R 控制台或 Python for Spark 接收到的输出来覆盖**控制台输出**选项卡的内容。 输出不能编辑。

“空间时间限制”节点

空间时间限制 (STB) 是进行了 Geohash 计算的空间位置的扩展。 更具体地说，STB 是一个字母数字字符串，它表示形状规则的空间和时间区域。

例如，STB **dr5ru7|2013-01-01 00:00:00|2013-01-01 00:15:00** 由以下三个部分组成：

- 地理散列 **dr5ru7**
- 开始时间戳记 **2013-01-01 00:00:00**
- 结束时间戳记 **2013-01-01 00:15:00**

例如，您可以使用空间和时间信息来提高两个实体是同一实体的置信度，因为这两个实体几乎在同一时间位于同一位置。 另外，通过显示两个实体由于其空间和时间接近而相关，您可以提高关系确定的准确性。

您可以根据需要选择**各个记录**或**逗留**方式。 这两种方式需要相同的基本详细信息，如下所示：

纬度字段。 请选择用于标识 WGS84 坐标系中的纬度的字段。

经度字段。 请选择用于标识 WGS84 坐标系中的经度的字段。

时间戳记字段。 选择用于标识时间或日期的字段。

“各个记录”选项

使用此方式可向记录添加一个附加字段，用于标识该记录在指定时间的位置。

派生。 选择一个或多个空间和时间密度，将根据所选内容派生新字段。 有关更多信息，请参阅第 85 页的『定义空间时间限制密度』。

字段名称扩展。 输入要向新字段名称添加的扩展名。 您可以选择将此扩展添加为**后缀**，也可以将其添加为**前缀**。

“逗留”选项

可以将逗留视作实体在其中持续或重复出现的位置和/或时间。 例如，可以使用它来标识进行定期运输的车辆，并确定与标准值的任何偏差。

逗留检测器监视实体的移动，并标记观察到实体在某个区域内“逗留”的情况。 逗留检测器将标记的每次逗留自动分配给一个或多个 STB，并使用内存内的实体和事件跟踪功能以最优效率检测逗留。

STB 密度。 选择空间和时间密度，将根据所选内容派生新字段。 例如，值 **STB_GH4_10MINS** 对应于大小约为 20 公里乘 20 公里且时间窗口为 10 分钟的四字符地理散列限制空间。 有关更多信息，请参阅第 85 页的『定义空间时间限制密度』。

实体标识字段。 选择要用作逗留标识的实体。 此标识字段用于标识事件。

最小事件数。 事件是数据中的行。 选择将实体视为逗留实体时，该实体的某一事件发生的最少次数。 逗留还必须满足以下**最短停留时间**字段所指定的条件。

最短停留时间。 指定实体必须在同一位置停留的最短持续时间。 这有助于排除操作，例如可以将等待红绿灯的车辆从逗留车辆中排除的操作。 逗留还必须满足以上**最小事件数**字段所指定的条件。

以下是更多有关逗留所必须满足的条件的详细信息：

让 $e_1, \ldots, e_n$ 表示在持续时间 (t_1, t_n) 期间从给定实体标识接收到的所有按时间排序的事件。在以下情况下，这些事件符合逗留条件：

- n >= 最小事件数
- $t_n - t_1$ >= 最短停留时间
- 所有事件 $e_1, \ldots, e_n$ 都在同一个 STB 中发生

允许逗留跨 STB 边界。 选择此选项后，逗留的定义将不再那么严格并且可以包括一些实体，例如在多个空间时间限制中逗留的实体。 例如，如果 STB 定义为整小时，那么选中此选项会将逗留一小时的实体识别为有效，即使这一小时由午夜前的 30 分钟和午夜后的 30 分钟组成也是如此。 如果未选中此选项，那么 100% 的逗留时间都必须在单个空间时间限制之内。

符合条件的时间段内事件的最小比例 (%)。 仅当选择了**允许跨 STB 边界逗留**时才可用。 使用此选项可控制一个 STB 中报告的逗留事实上可能与另一逗留重叠的程度。 选择必须在单一 STB 中发生的事件的最小比例可以标识逗留。 如果设置为 25%，并且事件比例为 26%，那么符合逗留条件。

例如，假定将逗留检测器配置为至少需要两个事件（最小事件数为 = 2），并且连续徘徊时间是在 4 字节地理散列限制空间和 10 分钟限制时间 (STB_NAME = STB_GH4_10MINS) 内至少徘徊 2 分钟。 检测到逗留时，假定下午 4:57 与 5:07 之间的 10 分钟时间范围内，在下午 4:58、5:01 和 5:03 发生三个合格事件时，即可认为实体在同一个 4 字节地理散列限制空间内停留。 合格时间百分比值指定允许用于逗留的 STB，如下所示：

- **100%**。 在下午 5:00 - 5:10 的限定时间内报告逗留，而在下午 4:50 - 5:00 的限定时间内不报告逗留（下午 5:01 和 5:03 发生的事件满足合格逗留所需的所有条件，并且其中 100% 的事件在 5:00 - 5:10 限定时间内发生）。
- **50%**。 两个限定时间内的逗留都将进行报告（下午 5:01 和 5:03 发生的事件满足合格逗留所需的所有条件，在这些事件中，至少 50% 发生在 4:50 - 5:00 限定时间内，并且至少 50% 发生在 5:00 - 5:10 限定时间内）。
- **0%**。 两个限定时间内的逗留都将进行报告。

如果指定了 0%，那么逗留报告将包括那些表示合格持续时间所触及的每个限定时间的 STB。 合格持续时间必须小于或等于 STB 中的限定时间的相应持续时间。 换而言之，如果配置的合格持续时间为 20 分钟，那么不得将 STB 配置为 10 分钟。

一旦满足合格条件就会报告逗留，并且对于每个 STB 不会多次进行报告。 假定有三个事件符合逗留条件，并且合格持续时间内发生的全部 10 个事件在同一个 STB 内。 在这种情况下，将在发生第三个合格事件时报告逗留。 另外 7 个事件都不会触发逗留报告。

注:

- 逗留检测器的内存内事件数据在进程之间不共享。 因此，特定实体与特定逗留检测器节点之间存在亲缘关系。 即，一个实体的传入移动数据必须始终一致地传递到跟踪该实体的逗留检测器节点，在整个运行过程中，这通常是同一个节点。
- 逗留检测器的内存内事件数据并非持久存储。 每当逗留检测器退出和重新启动时，所有工作中的逗留都将丢失。 这意味着停止并重新启动进程可能会导致系统错过报告实际逗留。 潜在的补救措施涉及重放部分历史移动数据（例如，后退 48 小时并重放适用于任何已重新启动的节点的移动记录）。
- 必须以时间顺序向逗留检测器输送数据。

定义空间时间限制密度

通过指定要包括在每个空间时间限制 (STB) 中的物理区域和耗用时间，可以选择空间时间限制的大小（密度）。

地理密度。 选择要包括在每个 STB 中的区域的大小。

时间间隔。 选择要包括在每个 STB 中的小时数。

字段名称。 以 STB 作为前缀，将根据前两个字段中的选择自动补全此名称。

流式 TCM 节点

“流式 TCM”节点可用于在一个步骤中对时间因果模型进行构建和评分。

有关时间因果建模的更多信息，请参阅《SPSS Modeler 建模节点》指南的“时间序列模型”部分中的“时间因果模型”主题。

流式 TCM 节点 -“时间序列”选项

在“字段”选项卡上，请使用**时间序列**设置来指定要包括在模型系统中的序列。

请选择应用于数据的数据结构选项。 对于多维数据，请单击**选择维度**以指定维度字段。 指定维度字段的顺序定义了这些字段在所有后续对话框和输出中的显示顺序。 请使用“选择维度”子对话框上的向上和向下箭头按钮对维度字段进行重新排序。

对于基于列的数据，术语序列的含义与术语字段相同。 对于多维数据，包含时间序列的字段称为度量字段。 对于多维数据，时间序列由度量字段以及每个维度字段的值定义。 对于基于列的数据和多维数据，注意事项如下所示。

- 将对指定为候选输入或同时作为目标和输入的序列加以考虑，以便将其包括在每个目标的模型中。 每个目标的模型都始终包含该目标自身的延迟值。
- 指定为强制输入的序列将始终包括在每个目标的模型中。
- 必须将至少一个序列指定为目标或者同时指定为目标和输入。
- 如果选择了**使用预定义角色**，那么角色为“输入”的字段将设置为候选输入。 没有任何预定义角色映射到强制输入。

多维数据

对于多维数据，请在网格中指定度量字段及相关角色，网格中的每一行都指定单个度量和角色。 缺省情况下，模型系统包含此网格中每一行的所有维度字段组合的序列。 例如，如果存在 *region* 维度和 *brand* 维度，那么在缺省情况下，指定度量 *sales* 作为目标意味着对于 *region* 与 *brand* 的每个组合，都存在单独的 sales 目标序列。

对于网格中的每一行，您可以通过单击维度的省略号按钮来定制任何维度字段的值集合。 此操作将打开“选择维度值”子对话框。 另外，您还可以添加、删除或复制网格行。

序列计数列显示当前对相关度量指定的维度值集合的数目。 显示的值可能大于序列的实际数目（每个集合各有一个对应的序列）。 当指定的某些维度值组合未与相关度量所包含的序列相对应时，将发生这种情况。

流式 TCM 节点 - 选择维度值

对于多维数据，您可以通过指定哪些维度值将应用于具有特定角色的特定度量字段来定制分析。 例如，如果 *sales* 是一个度量字段，*channel* 是值为“retail”和“web”的维度，那么您可以指定“web”销售是输入，“retail”销售是目标。 另外，还可以指定将应用于分析中使用的所有度量字段的维度子集。 例如，如果 *region* 是指示地理区域的维度字段，那么您可以将分析限制在特定区域。

所有值
指定包括当前维度字段的所有值。 这是缺省选项。

选择要包括或排除的值
使用此选项可以指定当前维度字段的值集。 如果对**方式**选择了**包括**，那么将仅包括**选择的值**列表中指定的值。 如果对**方式**选择了**排除**，那么将包括除**选择的值**列表中指定的值之外的所有值。

您可以对要从中进行选择的值集进行过滤。 满足过滤条件的值将显示在**匹配**选项卡中，不满足过滤条件的值将显示在**未选择的值**列表的**不匹配**选项卡中。 **全部**选项卡列示所有未选择的值，而无论指定了什么过滤条件。

- 指定过滤器时，可以使用星号 (*) 来表示通配符。
- 要清除当前过滤器，请在“过滤显示的值”对话框中对搜索项指定空值。

流式 TCM 节点 -“观测值”选项

在“字段”选项卡上，使用**观测**设置来指定用于定义观测的字段。

由日期/时间定义的观测值

您可以指定观测值由日期、时间或时间戳记字段定义。 除了用于定义观测值的字段以外，请选择用于描述观测值的适当时间间隔。 根据指定的时间间隔，您还可以指定其他设置，例如两次观测之间的时间间隔（增量）或每周的天数。 以下注意事项适用于时间间隔：

- 如果各个观测值之间的时间间距不定期（例如处理销售订单的时间），请使用**不定期**值。 选择**不规则**时，必须在“数据规范”选项卡上的**时间间隔**设置中指定用于分析的时间间隔。
- 如果观测值表示日期和时间，并且时间间隔为小时、分钟或秒，请使用**每天的小时数**、**每天的分钟数**或**每天的秒数**。 如果观测值表示时间（持续时间）并且未引用日期，而时间间隔为小时、分钟或秒，请使用**小时数（非周期性）**、**分钟数（非周期性）**或**秒数（非周期性）**。
- 根据选择的时间间隔，此过程可以检测缺失的观测值。 由于此过程假定所有观测值之间的时间间距相等，并假定未缺失观测值，因此有必要检测缺失的观测值。 例如，如果时间间隔为“天”，并且日期 2014-10-27 后面跟着 2014-10-29，那么表明缺失 2014-10-28 的观测值。 对于任何缺失的观测值，将插补值。 您可以在“数据规范”选项卡上指定用于处理缺失值的设置。
- 指定的时间间隔使此过程能够检测到同一时间间隔内的多个需要汇总到一起的观测值，并使各个观测值使用统一的时间间隔边界（例如每个月的第一天），以确保各个观测值之间的间距相等。 例如，如果时间间隔为“月”，那么同一个月内的多个日期将聚集到一起。 此类汇总称为分组。 缺省情况下，分组时将计算观测值的总和。 通过“数据指定项”选项卡上的**汇总和分布**设置，您可以指定另一种分组方法，例如计算各个观测值的平均值。
- 对于某些时间间隔，附加设置可以定义正常等间距时间间隔中的中断。 例如，如果时间间隔为“天”，但仅工作日有效，那么可以指定一周有五天，每周第一天为星期一。

周期或循环周期定义的观测

观测可以由一个或多个表示周期或周期反复循环（直至达到任意数目的循环级别为止）的整数字段定义。 借助此结构，您可以描述任何标准时间间隔都无法支持的观测值序列。 例如，要描述只有 10 个月的财年，可以使用表示年的循环字段和表示月的周期字段，并且一个循环的长度为 10。

指定循环周期的字段定义周期性级别层次结构，最低级别由**周期**字段定义。 次高级别由级别为 1 的循环字段指定，接着由级别为 2 的循环字段指定，依此类推。 除最高级别以外，每个级别的字段值对于次高级别都必须具有周期性。 最高级别的值不得具有周期性。 例如，如果是 10 个月的财年，年中的月份是周期性值，而年不是周期性的。

- 特定级别的循环长度是下一个最低级别的周期。 在财年示例中，只有一个循环级别，并且循环长度为 10，这是因为次低级别表示月，而指定的财年包含 10 个月。
- 指定不从 1 开始的任何周期字段的起始值。 此设置是检测缺失值的必备步骤。 例如，如果周期性字段起始于 2，但起始值指定为 1，那么此过程将假定该字段的每个循环中的第一个周期都有一个缺失值。

流式 TCM 节点 -“时间间隔”选项

用于分析的时间间隔可以与观测的时间间隔不同。 例如，如果观测时间间隔为“天”，可以为分析时间间隔选择“月”。 然后，系统在构建模型之前将数据从每日数据汇总为每月数据。 您还可以选择将时间间隔较长的数据拆分到较短的时间间隔内。 例如，如果观测值是每季度数据，那么您可以将每季度数据分布为每月数据。

执行分析的时间间隔的可用选项取决于观测的定义方式以及这些观测的时间间隔。 特别是，如果观测值由循环周期定义，那么仅支持汇总。 在这种情况下，分析时间间隔必须大于或等于观测值的时间间隔。

分析时间间隔在“数据指定项”选项卡上的**时间间隔**设置中指定。 汇总或分布数据的方法在“数据指定项”选项卡上的**汇总和分布**设置中指定。

流式 TCM 节点 -“汇总和分布”选项

聚集函数
如果用于分析的时间间隔比观测值的时间间隔长，那么将对输入数据进行聚集。 例如，当观测值的时间间隔为“天”并且分析时间间隔为“月”时，将执行聚集。 可用的聚集函数如下所示：mean、sum、mode、min 或 max。

分布函数
如果用于分析的时间间隔比观测值的时间间隔短，那么将对输入数据进行分布。 例如，当观测值的时间间隔为“季度”并且分析时间间隔为“月”时，将执行分布。 可用的分布函数如下所示：mean 或 sum。

分组函数
当观测值由日期/时间定义，并且同一个时间间隔内存在多个观测值时，将进行分组。 例如，如果观测值的时间间隔为“月”，那么同一个月的多个日期将分组到一起，并与它们所在的月份相关联。 可用函数有：mean、sum、mode、min 或 max。当观测值由日期/时间定义，并且观测值的时间间隔指定为“不定期”时，将始终执行分组。

注: 尽管分组是一种聚集形式，但在对缺失值进行任何处理之前执行，而正式的聚集是在对所有缺失值进行处理之后执行。 如果观测值的时间间隔指定为“不规则”，那么将仅使用分组函数来执行聚集。

将跨天观测值聚集到前一天
指定是否将时间跨天边界的观测值汇总到前一天的值。 例如，如果每一天的时间范围是从 20:00 开始的 8 小时，那么对于每小时观测值，此设置指定是否将介于 00:00 与 04:00 之间的观测值包括在前一天的聚集结果中。 仅当观测值的时间间隔为“每天的小时数”、“每天的分钟数”或“每天的秒数”，并且分析时间间隔为“天”时，此设置才适用。

所指定字段的定制设置
您可以对每个字段指定聚集函数、分布函数和分组函数。 这些设置将覆盖聚集函数、分布函数和分组函数的缺省设置。

流式 TCM 节点 -“缺失值”选项

输入数据中的缺失值将替换为插补值。 可用的替换方法如下所示：

线性插值
使用线性插值法替换缺失值。 缺失值之前的最后一个有效值以及之后的第一个有效值用于插值。 如果序列中的第一个或最后一个观测值具有缺失值，那么将使用序列开头或结尾的两个最近的非缺失值。

序列平均值
将缺失值替换为整个序列的平均值。

临近点的平均值
使用有效周围值的平均值替换缺失值。 邻近点的跨度为缺失值前后用于计算平均值的有效值数目。

邻近点的中值
使用有效周围值的中值替换缺失值。 邻近点的跨度为缺失值前后用于计算中位数的有效值数目。

线性趋势
此选项使用序列中的所有非缺失观测值来拟合简单线性回归模型，该模型随后用于插补缺失值。

其他设置：

缺失值的最大百分比 (%)
指定针对任何序列允许的缺失值的最大百分比。 将从分析中排除缺失值数量超过指定最大值的序列。

流式 TCM 节点 -“常规数据”选项

每个维度字段的不同值的最大数目
此设置适用于多维数据，它指定任何一个维度字段所允许的不同值的最大数量。 缺省情况下，此限制设置为 10000，但可以增大到任意大的数字。

流式 TCM 节点 -“常规构建”选项

置信区间宽度 (%)
此设置控制预测及模型参数的置信区间。 您可以指定任何小于 100 的正数值。 缺省情况下，将使用 95% 置信区间。

每个目标的最大输入数目
此设置指定每个目标的模型中允许的最大输入数目。 您可以指定 1 到 20 范围内的整数。 每个目标的模型都始终包含自身的延迟值，因此将此值设置为 1 表示唯一的输入是目标自身。

模型容差
此设置控制用于确定每个目标的最佳输入集合的迭代式过程。 您可以指定任何大于零的值。 缺省值为 0.001。 模型容差是预测变量选择的停止条件。 它可以影响最终模型中包含的预测变量数。 但是，如果目标自身预测良好，那么最终模型中可能不包含其他预测变量。 可能需要一些试验和错误（例如，如果将此设置位置为较高的值，那么可尝试将其设置为较小的值以查看是否可包含其他预测变量）。

离群值阈值 (%)
如果根据模型计算而得的可能性指出观测值为超出此阈值的离群值，那么会将该观测值标记为离群值。 您可以指定 50 到 100 范围内的值。

每个输入的延迟项数
此设置指定每个目标的模型中每个输入的延迟项数。 缺省情况下，延迟项数根据用于分析的时间间隔自动确定。 例如，如果时间间隔为“月”（增量为 1 个月），那么延迟项数为 12。 您可以选择性地明确指定延迟项数。 指定的值必须是 1 - 20 范围内的整数。

使用现有模型继续估算
如果已生成时间因果模型，那么选择此选项将复用对该模型指定的条件设置，而不构建新模型。 这样，就可以基于先前模型设置但使用较新的数据来重新估算并生成新预测，从而节省时间。

流式 TCM 节点 -“估计期”选项

缺省情况下，估计期从所有序列中的最早观测值的时间开始，并以最晚观测值的时间结束。

按开始时间和结束时间
您可以同时指定估计期的开始时间和结束时间，也可以仅指定开始时间或者仅指定结束时间。 如果省略了估计期的开始时间或结束时间，那么将使用缺省值。

- 如果观测值由日期/时间字段定义，请以用于该日期/时间字段的格式输入开始时间值和结束时间值。
- 对于由循环周期定义的观测值，请对每个循环周期字段指定值。 每个字段都将显示在单独的列中。

按最晚或最早时间间隔

将估计期定义为指定数目的时间间隔，这些时间间隔从数据中的最早时间间隔开始或者以最晚时间间隔结束，并可以使用可选的偏移量。 在此上下文中，时间间隔是指分析时间间隔。 例如，假定观测值按月获取，但分析时间间隔为季度。 指定**最晚**并对**时间间隔数目**指定值 24 表示最晚的 24 个季度。

您可以选择性地排除指定数目的时间间隔。 例如，指定最晚的 24 个时间间隔并对排除数目指定 1 表示估计期由最后一个时间间隔之前的 24 个时间间隔组成。

流式 TCM 节点 -“模型”选项

模型名称

您可以对模型指定定制名称，也可以接受自动生成的名称，即 *TCM*。

预测

将记录扩展至将来选项用于设置时间间隔数目，以预测估计期结束之后的情况。 在这种情况下，时间间隔为“数据指定项”选项卡上指定的分析时间间隔。 请求进行预测时，将为所有并非同时作为目标的输入序列自动构建自回归模型。 然后，使用这些模型生成这些输入序列在预测期内的值。 此设置没有最大限制。

“CPLEX 优化”节点

借助 CPLEX Optimization 节点，可以通过优化编程语言 (OPL) 模型文件来使用基于优化的复杂数学。 此功能可在不再受支持的 IBM Analytical Decision Management 产品中使用，但在 SPSS Modeler 中，现在还可以使用 CPLEX 节点，而不需要 IBM Analytical Decision Management。

有关 CPLEX 优化和 OPL 的更多信息，请参阅 IBM ILOG CPLEX Optimization Studio 文档。

CPLEX Optimization 节点支持多个数据源或多维入局数据。 可以将多个节点连接到 CPLEX Optimization 节点，并且每个先验节点可用于向 OPL 模型计算提供数据 - 通过个别字段映射来设置为个别元组集合。

输出由 CPLEX Optimization 节点生成的数据时，数据源中的原始数据可以作为单一索引或作为结果的多维索引一起输出。

CPLEX Optimization 节点的决策变量不支持复杂数组。

注:

- 在 **IBM SPSS Modeler Server** 上运行包含 CPLEX Optimization 节点的流时，缺省情况下将使用嵌入式 Community 版本 CPLEX 库。 其限制为 1000 个变量和 1000 约束。 如果安装完整版本的 IBM ILOG CPLEX 并且想要改为使用完整版本的 CPLEX 引擎（无此类限制），那么请针对您的平台完成以下步骤。
  - 在 Windows 上，编辑 `options.cfg` 并添加 OPL 库路径。 例如：

    ```
    cplex_opl_lib_path="<CPLEX_path>\opl\bin\<Platform_dir>"
    ```

    其中，`<CPLEX_path>` 是 CPLEX 安装目录，例如，`C:\Program Files\IBM\ILOG\CPLEX_Studio127`，而 `<Platform_dir>` 是特定于平台的目录，例如，`x64_win64`。
  - 在 Linux 上，编辑 `modelersrv.sh` 并添加 OPL 库路径。 例如：

    ```
    CPLEX_OPL_LIB_PATH=<CPLEX_path>/opl/bin/<Platform_dir>
    ```

    其中，`<CPLEX_path>` 是 CPLEX 安装目录，例如，`/root/Libs_127_FullEdition/Linux_x86_64`，而 `<Platform_dir>` 是特定于平台的目录，例如，`x86-64_linux`。

注:

- 在 **SPSS Modeler Solution Publisher** 中运行包含 CPLEX Optimization 节点的流时，缺省情况下，将使用嵌入式 Community 版本 CPLEX 库。 其限制为 1000 个变量和 1000 约束。 如果安装完整版本的 IBM ILOG CPLEX 并且想要改为使用完整版本的 CPLEX 引擎（无此类限制），那么请针对您的平台完成以下步骤。
  - 在 Windows 上，添加 OPL 库路径作为 `modelerrun.exe` 的命令行自变量。 例如：

    ```
    -o cplex_opl_lib_path="<CPLEX_path>\opl\bin\<Platform_dir>"
    ```

    其中，`<CPLEX_path>` 是 CPLEX 安装目录，例如，`C:\Program Files\IBM\ILOG\CPLEX_Studio127`，而 `<Platform_dir>` 是特定于平台的目录，例如，`x64_win64`。
  - 在 Linux 上，编辑 `modelerrun` 并添加 OPL 库路径。 例如：

    ```
    CPLEX_OPL_LIB_PATH=<CPLEX_path>/opl/bin/<Platform_dir>
    ```

其中，<CPLEX_path> 是 CPLEX 安装目录，例如，/root/Libs_127_FullEdition/Linux_x86_64，而 <Platform_dir> 是特定于平台的目录，例如，x86-64_linux。

- CPLEX 在 MacOS 上不受支持。 您可以使用该节点（将其添加到流中，编辑其属性等），但无法运行该节点。

设置 CPLEX Optimization 节点的选项

CPLEX Optimization 节点的选项卡“选项”包含下列字段。

OPL 模型文件。 选择优化编程语言 (OPL) 模型文件。

OPL 模型。 选择 OPL 模型后，内容将显示在此处。

输入数据

在“输入数据”选项卡上，**数据源**下拉列表将列出连接到 CPLEX Optimization 节点的所有数据源（先验节点）。 从下拉列表中选择数据源将刷新下方的**输入映射**部分。 单击**应用所有字段**可自动生成所选数据源的所有字段映射。 **输入映射**表将自动进行填充。

在 OPL 模型中输入与入局数据对应的元组集合名称。 然后，如果需要，请根据元组字段在元组定义中的顺序来验证映射到数据输入字段的所有元组字段。

设置数据源的输入映射后，可以从下拉列表中选择其他数据源并重复该过程。 先前的数据源映射将自动进行保存。 完成时，单击**应用**或**确定**。

其他数据

在“其他数据”选项卡上，如果您需要指定任何其他数据进行优化，请使用 **OPL 数据**部分。

输出

当输出是决策变量时，它必须采用先验数据源（入局数据）作为索引，并且索引必须在“输入数据”选项卡上的**输入映射**部分中进行预定义。 当前不支持任何其他类型的决策变量。 决策变量可以具有单个索引或多个索引。 SPSS Modeler 会将 CPLEX 结果与全部或部分原始入局数据一起输出，这与其他 SPSS Modeler 节点一致。 必须在下述的**输出元组**字段中指定所引用的对应索引。

在“输出”选项卡上，选择输出模式（**原始输出**或**决策变量**）并指定其他选项（如果适用）。 “原始输出”选项将直接输出目标函数，而与名称无关。

OPL 中的目标函数值变量名称。 如果您已选择**决策变量**输出模式，则将启用此字段。 输入 OPL 模型中目标函数值变量的名称。

输出的目标函数值字段名称。 输入要用在输出中的字段名称。 缺省值为 _OBJECTIVE。

输出元组。 输入入局数据中的预定义元组的名称。 此属性充当决策变量的索引，并且预期通过“变量输出”进行输出。 “输出元组”必须与 OPL 中的决策变量定义一致。 如果有多个索引，那么元组名称必须以逗号 (,) 连接。

变量输出。 添加一个或多个要包含在输出中的变量。

第 4 章 字段操作节点

字段操作概述

经过初始数据研究之后，您可能需要在准备分析的过程中选择、清除或构造数据。“字段操作”选项板包含许多适用于这种变换和准备的节点。

例如，使用“派生”节点，可以创建当前数据中并未呈现的属性。 或者，使用“分级”节点可以自动针对目标分析进行字段值的重新编码。 您可能会发现自己使用“类型”节点的频率很高，该节点可用于为数据集中每个字段分配测量级别、值和建模角色。 其操作对于处理缺失值和下游建模十分实用。

“字段操作”选项板包含下列节点：

“自动数据准备 (ADP)”节点可分析您的数据并标识修正，筛选出存在问题或可能无用的字段，并在适当的情况下派生新的属性，通过智能筛选和抽样技术改进性能。 您可以采用完全自动化方式使用此节点，从而允许此节点选择并应用修订，另外也可以在应用修订前预览更改并根据需要接受、拒绝或进行修改。

“类型”节点指定字段元数据和属性。 例如，您可以指定每个字段的测量级别（连续、名义、有序或标志）、设置用于处理缺失值和系统 Null 值的选项、设置用于建模的字段的角色、指定字段标签和值标签以及为字段指定值。

“过滤”节点用于过滤（废弃）字段、对字段进行重命名以及将字段从一个源节点映射到另一个节点。

“派生”节点修改数据值或者根据一个或多个现有字段创建新字段。 它创建类型为公式、标志、名义、状态、计数和条件的字段。

“整体”节点将两个或多个模型块组合在一起，以获得比可从任何一个模型实现的预测更准确的预测。

“填充器”节点用于替换字段值并更改存储。 您可以选择基于 CLEM 条件（例如 `@BLANK(@FIELD)`）的替换值。 或者，也可以选择将所有空白值或空值替换为特定值。 “填充”节点通常与“类型”节点一起使用以替换缺失值。

“匿名化”节点用于转换字段名和字段值在下游的表示方式，从而掩饰原始数据。 如果要允许其他用户构建含有敏感数据（例如客户名称或其他详细信息）的模型，那么这种节点十分有用。

“重新分类”节点将一组分类值转换为另一组值。 对于折叠类别或者进行数据重新分组以执行分析而言，重新分类非常有用。

“分箱”节点根据一个或多个现有连续（数字范围）字段的值自动创建新的名义（集合）字段。 例如，您可以将连续收入字段转换为一个包含各组收入（作为与均值之间的偏差）的新分类字段。 一旦创建新字段分级后，即可根据割点创建“衍生”节点。

利用“近因、频数和货币 (RFM) 分析”节点，您能够通过检查客户最近一次从您那里购买的时间（近因）、购买的频率（频数）以及他们在所有交易中花费的金额（货币），以定量方式确定哪些客户可能是最佳客户。

分区节点可生成分区字段，该字段可将数据分割为单独的子集以便在模型构建的训练、测试和验证阶段使用。

“设为标志”节点根据针对一个或多个名义字段定义的分类值派生多个标志字段。

“重构”节点将名义字段或标志字段转换为一组字段，这组字段可以使用另一字段的值进行填充。 例如，如果给定名为支付类型的字段，且值为信用、现金和借记，那么将创建三个新字段（信用、现金和 借记），其中每个字段都可能包含实际支付的值。

“转置”节点会交换行和列中的数据，以便记录成为字段，字段成为记录。

使用“时间间隔”节点可以指定时间间隔并派生用于估算或预测的新时间字段。 支持全部范围的时间间隔，从秒到年。

“历史记录”节点创建新字段，这些字段包含先前记录中的字段的数据。 “历史记录”节点最常用于顺序数据，例如时间序列数据。 在使用“历史记录”节点之前，您可能希望使用“排序”节点对数据进行排序。

“字段重新排序器”节点定义用于显示下游字段的自然顺序。 此顺序将影响字段在各种位置（例如表、列表和字段选择器）的显示方式。 在使用宽数据集以提高感兴趣字段的可见性时，此操作很有用。

在 SPSS Modeler 中，表达式构建器空间函数、“空间-时间预测”(STP) 节点和“地图可视化”节点之类的项使用投影坐标系。 使用“重新投影”节点可以更改所导入的任何使用了地理坐标系的数据的坐标系。

其中某些节点可以通过“数据审核”节点所创建的审核报告直接生成。 有关更多信息，请参阅主题 第 243 页的『生成其他用于数据准备的节点』。

自动数据准备

准备数据以进行分析是任何项目中最重要的步骤之一，而从传统上说也是最耗时的步骤之一。 “自动数据准备 (ADP)”为您处理任务，分析您的数据并识别修正，筛选出存在问题或可能无用的字段，并在适当的情况下派生新的属性，通过智能筛选技术改进性能。 您可以通过完全**自动**的方式使用算法，这种方式可以允许选择

并应用修正；或者也可以通过**交互式**方式使用算法，这种方式可以在做出更改前对其进行预览，并根据需要进行接受或拒绝。

通过使用 ADP，您可以轻松、快速地准备好用于模型构建的数据，而无需事先具备所涉及的统计概念的知识。 您可以更快速地构建模型并进行评分。此外，使用 ADP 还能提高自动化建模过程（例如，模型刷新和 Champion-Challenger 分析）的稳健程度。

注: 当 ADP 准备字段进行分析时，它将创建包含调整或转换的新字段，而不是替换旧字段的现有值和属性。旧字段未用于进一步分析；其角色设置为“无”。

示例。 在调查业主保险理赔方面拥有有限资源的保险公司希望构建一个模型来标记具有潜在欺骗性的可疑理赔。 构建模型前，他们将使用自动数据准备来准备数据进行建模。 由于他们希望能够在应用转换前查看建议的转换，他们将在交互模式下使用自动数据准备。

某汽车集团希望跟踪各类私人汽车的销售情况。 为了能够标识表现良好和表现不好的型号，他们希望建立汽车销售和汽车特性之间的关系。 他们将使用自动数据准备来准备数据进行分析，同时使用准备“之前”和“之后”的数据构建模型以查看结果的差别。

您的目标是什么? 自动数据准备可以推荐能够加快其他算法的建模速度、并增强这些模型的预测能力的数据准备步骤。 可包括变换、构建和选择功能。 也可对目标进行变换。 您可以指定数据准备过程应遵循的建模优先级次序。

- **平衡速度和准确度**。 此选项可以准备数据，以使建模算法处理数据的速度和预测的精确度具有同等优先级。
- **为速度而优化**。 此选项可以准备数据，以使建模算法处理数据的速度具有较高优先级。 当您处理超大数据集，或要求快速得到结果时，请选中此选项。
- **为准确度而优化**。 此选项可以准备数据，以使建模算法生成的预测结果的准确性具有较高优先级。
- **定制分析**。 如果您希望手动修改“设置”选项卡上的算法，请选择此选项。 注意，如果您随后在“设置”选项卡上更改了与其他目标之一不一致的选项，则会自动选择该设置。

训练节点

ADP 节点以过程节点实现，其工作方式与类型节点相似。**训练** ADP 节点相当于类型节点实例化。 一旦执行分析后，只要上游数据模型无变化，就可对数据应用指定的转换，而无需进一步分析。 与类型和过滤节点类似，在 ADP 节点断开连接后，它会记住数据模型和转换，这样当它重新连接时，就不需要再次训练。这允许您在典型数据子集上训练该节点，然后进行复制或部署，以便在实时数据上多次使用。

使用工具栏

工具栏允许您运行和更新数据分析显示，并生成可与原始数据结合使用的节点。

- **生成** 通过此菜单，您可以生成过滤节点或派生节点。 请注意，仅当在“分析”选项卡上显示有分析时，该菜单才可用。

  过滤节点删除转换后的输入字段。 如果您将 ADP 节点配置为保留数据集中的原始输入字段，那么这将恢复原始输入集，从而允许您根据输入解释评分字段。 例如，如果要针对不同输入生成评分字段图表，这可能非常有用。

  “派生”节点可以恢复原始数据集和目标单位。 只有当 ADP 节点包含对范围目标重新标度（即，在“准备输入和目标”面板上选择了 Box-Cox 重定比）的分析时，才能生成“派生”节点。 如果目标不是一个范围，或未选中 Box-Cox 重定比，则不能生成派生节点。 有关更多信息，请参阅主题 第 102 页的『生成“派生”节点』。
- **视图** 包含可以控制“分析”选项卡上所显示内容的选项。 其中包括图形编辑控件，以及主面板和链接视图的显示选择。
- **预览**显示将在输入数据上应用的转换样例。
- **分析数据** 使用当前设置启动分析，并在“分析”选项卡上显示结果。
- **清除分析**删除现有分析（仅当存在当前分析时可用）。

节点状态

ADP 节点在 IBM SPSS Modeler 工作区上的状态通过图标上的箭头或勾号进行指示，即是否已运行过分析。

有关“自动数据准备”节点执行的计算的更多信息，请参阅《IBM SPSS Modeler 算法指南》的『自动数据准备算法』一节。 该指南以 PDF 格式提供，位于安装磁盘的 \Documentation 目录中，可能作为产品下载的一部分提供，也可能位于 Web 上。

“字段”选项卡

在构建模型之前，需要指定要将哪些字段用作目标和输入。 某些特殊情况下，所有建模节点将采用上游的类型节点的字段信息。 如果正在使用类型节点选择输入和目标字段，则不必在此选项卡上做任何更改。

使用类型节点设置。 该选项通知节点使用来自上游类型节点的字段信息。 这是缺省选项。

使用定制设置。 该选项通知节点使用在此处指定的字段信息，而不是在任何上游类型节点中给出的字段信息。 选中此选项后，请根据需要指定下面的字段。

目标。 对于需要一个或多个目标字段的模型，请选择目标字段或字段。 此操作与在“类型”节点中将字段的角色设置为目标类似。

输入。 选择输入字段或字段。 此操作与在“类型”节点中将字段的角色设置为输入类似。

“设置”选项卡

“设置”选项卡包含若干组不同的设置，您可以对其进行修改以对算法处理数据的方式进行微调。 如果您对与其他目标不一致的缺省设置进行了更改，则“目标”选项卡会自动更新为选择**自定义分析**选项。

字段设置

使用频率字段。 此选项允许您选择一个字段作为频率权重。 如果您的培训数据中的每个记录代表多个单位（例如，如果您正使用汇总数据），请使用此选项。 字段值应是每个记录代表的单位的数量。

使用权重字段。 此选项允许您选择一个字段作为观测值权重。 观测值权重将作为对输出字段各个水平上方差的差异的一种考量。

如何处理从建模中排除的字段。 指定如何处理排除的字段；您可以选择将它们从数据中过滤掉或仅把它们的角色设置为**无**。

注: 此操作还将应用于已变换的目标。 例如，如果目标的新派生版本用作**目标**字段，原始目标将被过滤或设置为**无**。

如果传入的字段与现有分析不匹配。 指定在您执行经过训练的 ADP 节点时，如果接收数据集中缺失一个或多个所需输入字段会怎样。

- **停止执行并保留现有分析**。 这将停止执行过程，保留当前分析信息并显示错误。
- **清除现有分析，并分析新数据**。 这将清除现有分析、分析接收数据并对该数据应用建议的转换。

准备日期和时间

许多建模算法无法直接处理日期和时间细节；这些设置允许您从现有数据中的日期和时间派生新的持续时间数据，以用作模型输入。 必须采用日期或时间存储类型预定义包含日期和时间的字段。 不建议在自动数据准备后将原始日期和时间字段用作模型输入。

为建模准备日期和时间。 取消选择该选项将在保持选择的同时禁用所有其他“准备日期&时间”控件。

计算到参考日期的耗用时间。 这将为包含日期的每个变量生成自参考日期后的年/月/日数。

- **参考日期**。 指定以该日期为参考，根据输入数据中的日期信息计算持续时间的日期。 如果选择**当前日期**，那么在执行 ADP 时将始终使用当前系统日期。 要使用特定日期，选择**固定日期**，并输入所需日期。 首次创建节点时，将自动在**固定日期**字段中输入当前日期。
- **持续日期单位**。 指定 ADP 是自动确定持续日期的单位，还是从**固定单位**（年、月或日）中选择。

计算到参考时间的耗用时间。 这将为包含时间的每个变量生成自参考日期后的小时/分钟/秒数。

- **参考时间**。 指定以该时间为参考，根据输入数据中的时间信息计算持续的时间。 如果选择**当前时间**，则 ADP 执行时始终使用当前系统时间。 要使用特定时间，选择**固定时间**，并输入所需具体时间。 首次创建节点时，将自动在**固定时间**字段中输入当前时间。
- **持续时间单位**。 指定 ADP 是自动确定持续时间的单位，还是从**固定单位**（小时、分或秒）中选择。

抽取周期时间元素。 使用这些设置将单个日期或时间字段分割成一个或多个字段。 例如，如果您选择了全部三个日期复选框，那么输入日期字段“1954-05-23”将分割成三个字段：1954、5 和 23，这三部分均使用**字段名称**面板上定义的后缀，并且将忽略原始日期字段。

- **从日期中抽取**。 对于任何日期输入，请指定是否要提取年、月、日或任意组合。
- **从时间中抽取**。 对于任何时间输入，请指定是否要如果要提取小时、分、秒或任意组合。

排除字段

质量较差的数据会影响到预测的准确性，因此需要为输入特征指定可接受的质量级别。 所有为常量或缺失值达 100% 的字段自动被排除。

排除低质量输入字段。 取消选择该选项将在保持选择的同时禁用所有其他“排除字段”控件。

排除缺失值过多的字段。 删除缺失值超过指定百分比的字段，而不会用于进一步分析。 指定大于或等于 0 的值等同于取消选择此选项，同时，指定小于或等于 100 的值将自动排除具有所有缺失值的字段。 缺省值为 50。

排除具有过多唯一类别的名义字段。 移除类别超过个数的字段，而不会用于进一步分析。 请指定正整数。 缺省值为 100。 这对于自动从建模中删除包含记录特有信息（如标识、地址或名称）的字段非常有用。

排除单个类别中具有过多值的分类字段。 移除在单个类别中包含超过指定百分比的记录的有序和名义字段，而不会用于进一步分析。 指定大于或等于 0 的值等同于取消选择此选项，同时，指定小于或等于 100 的值将自动排除常数字段。 缺省值为 95。

准备输入和目标

由于没有数据处于适合处理的完美状态，您可能希望在运行分析之前调整一些设置。 例如，这可能包括删除离群值，指定如何处理缺失值或调整类型。

注: 如果您在此面板上更改值，那么**目标**选项卡将自动更新为选择**定制分析**选项。

准备用于建模的输入和目标字段。 将面板上的所有字段切换为打开或关闭。

调整类型并提高数据质量。 对于输入和目标，您可以分别指定几个数据转换；这是因为您可能不希望更改目标值。 例如，以美元为单位的收入预测变量比以对数（美元）度量的预测变量更有意义。 此外，如果目标有缺失值，将没有预测增益来填充缺失值，而在输入中填充缺失值可启用一些算法来处理可能会丢失的信息。

这些转换的其他设置（如离群值分界值）对于目标和输入都通用。

您可以为输入或目标、或两者选择以下设置：

- **调整数字字段的类型**。 选择此选项以确定有序测量级别的数字字段是否可以转换为连续，反之亦然。 您可以指定最小和最大阈值以控制转换。
- **对名义字段重新排序**。 选择此选项以按从小到大的类别顺序排序名义（集合）字段。
- **替换连续字段中的离群值**。 指定是否替换离群值；将其与以下的**替换离群值的方法**结合使用。
- **连续字段：将缺失值替换为平均值**。 选择本选项以替换连续（范围）特征的缺失值。
- **名义字段：将缺失值替换为众数**。 选择本选项以替换名义（集合）特征的缺失值。
- **有序字段：将缺失值替换为中位数**。 选择本选项以替换有序（有序集合）特征的缺失值。

有序字段的最大值数目。 指定重新定义有序（有序集合）字段为连续（范围）的阈值。 缺省值为 10；因此，如果一个有序字段有超过 10 个类别，它将被重新定义为连续（范围）。

连续字段的最小值数目。 指定重新定义尺度或连续（范围）字段为有序（有序集合）的阈值。 缺省值为 5；因此，如果连续字段有少于 5 个值，它将被重新定义为有序（有序集合）。

离群值截断值。 指定离群值截断标准（采用标准差测量），缺省值为 3。

用于替换离群值的方法。 选择是否通过修整（强制）分界值、将其删除或设置为缺失值来替换离群值。 在任何离群值被设置为缺失值后，将按照以上所选的缺失值处理设置进行处理。

将所有连续输入字段放在一个共同的尺度上。 要标准化连续输入字段，选中本复选框并选择正态化方法。 缺省值为 **z-score 变换**，您可以在其中指定**最终均值**（缺省值为 0）和**最终标准偏差**（缺省值为 1）。 或者，您可以选择使用 **Min/max 变换**并指定最小值和最大值，缺省值分别为 0 和 100。

在“构造和选择功能部件”面板上选择**执行功能部件构造**时，此字段特别有用。

使用 Box-Cox 变换重新调整连续目标。 要标准化连续（尺度或范围）目标字段，请选中此复选框。 Box-Cox 转换的**最终均值**缺省值为 0，同时**最终标准差**缺省值为 1。

注: 如果您选择将目标标准化，那么将会变换目标的维度。 这时，您可能需要生成“派生”节点以应用逆转换，将转换后的单位转回可识别的格式，以供进一步处理。 有关更多信息，请参阅主题 第 102 页的『生成“派生”节点』。

构建和特征选择

为提高数据预测能力，您可以根据现有字段转换输入字段或构建新的字段。

注: 如果您在此面板上更改值，那么**目标**选项卡将自动更新为选择**定制分析**选项。

转换、构造和选择输入字段以提高预测能力。 将面板上的所有字段切换为打开或关闭。

合并稀疏类别以最大化与目标的关联。 选中此选项可以减少与目标关联的需处理的变量数，得到更简约的模型。 如果需要，更改 0.05 的缺省概率值。

注意，如果所有类别合并为一个类别，字段的原始和派生版本将被排除，因为它们没有作为预测变量的值。

当没有目标时，根据计数合并稀疏类别。 如果您处理的是没有目标的数据，那么可以选择合并有序（有序集合）和/或名义（集合）特征的松散类别。 指定标识要合并类别的数据中观测值或记录的最小百分比，缺省值为 10。

使用以下规则合并类别:

- 合并不能在二元字段上执行。
- 如果在合并过程中只有两个类别，合并将停止。
- 如果没有原始类别，或者合并期间所创建类别的观测值百分比均不少于指定最小观测值百分比，合并将停止。

在保留预测能力的同时对连续字段进行分级。 如果您拥有的数据包含分类目标，那么可以采用强关联对连续输入分级，以改进处理性能。 如果需要，更改 0.05 的缺省齐次子集概率值。

如果特定字段的离散化结果为单个块，则会排除字段的原始和分级版本，因为它们没有值作为预测变量。

注: ADP 中的分级与 IBM SPSS Modeler 其他部分中使用的最佳分级不同。 最佳离散化使用熵信息将连续变量转换为分类变量；这需要在内存中对全部数据进行排序和存储。 ADP 使用齐次子集来离散化连续变量，这意味着 ADP 离散化不需要在内存中对全部数据进行排序和存储。 通过使用齐次子集方法离散化连续变量，离散化后的类别数总是小于或等于目标类别数。

执行特征选择。 选择此选项将移除相关系数低的特征。 如果需要，更改 0.05 的缺省概率值。

该选项仅适用于目标为连续的连续输入特征，以及类别输入特征。

执行特征构造。 选中此选项，以从包含多个现有特征的组合中派生新特征，现有特征随后将从建模过程中丢弃。

该选项仅适用于目标为连续或不存在目标的连续输入特征。

字段名称

为方便识别新的和转换后的特征，ADP 可以创建并应用基本新名称、前缀或后缀。 您可以更改这些名称，以使其与您的要求和数据更加相关。 如果要指定其他标签，您将需要在下游“类型”节点中进行更改。

已转换和已构造的字段。 指定要应用到转换目标和输入字段的名称扩展。

注意，在 ADP 节点中，将字符串字段设置为空可能会引起错误，这具体取决于您选择用来处理未使用字段的方法。 如果在“设置”选项卡的“字段设置”面板上将**如何处理从建模中排除的字段**设置为**过滤掉未使用字段**，则输入和目标的名称扩展将被设置为空。 原始字段将被过滤掉，并替换为转换后的字段。在这种情况下，转换后的新字段与原始字段的名称相同。

不过，如果您选择了**将未使用字段的方向设置为“无”**，那么目标和输入的名称扩展将为空，并会引起错误，因为您试图创建重复的字段名称。

此外，还需要指定要应用到通过“选择和构建”设置所构建的任何特征的前缀名称。 新名称将通过向此前缀根名称添加数字后缀生成。 数字格式取决于生成的特征数目，例如：

- 第 1-9 个构建的特征将命名为：feature1 到 feature9。
- 第 10-99 个构建的特征将命名为：feature01 到 feature99。
- 第 100-999 个构建的特征将命名为：feature001 到 feature999，依此类推。

这可以确保不论有多少个特征，都将按有意义的顺序排列。

根据日期和时间计算的持续时间。 指定要应用到从日期和时间计算的持续时间的名称扩展。

从日期和时间提取的循环元素。 指定要应用到从日期和时间提取的循环元素的名称扩展。

“分析”选项卡

1. 在完成对 ADP 的设置（包括对“目标”、“字段”和“设置”选项卡所作的任何更改）后，单击**分析数据**。算法将设置应用到数据输入，并在“分析”选项卡上显示结果。

“分析”选项卡包含表格和图形输出，其中显示数据处理概要，并显示有关如何修改或改进数据以提高评分的建议。 您可以审核这些建议，并加以接受或拒绝。

“分析”选项卡包含两个面板，主视图位于左侧，链接或辅助视图位于右侧。 有三个主视图：

- 字段处理摘要（缺省视图）。 有关更多信息，请参阅主题 第 97 页的『字段处理概要』。
- 字段。 有关更多信息，请参阅主题 第 98 页的『字段』。
- 操作摘要。 有关更多信息，请参阅主题 第 99 页的『操作摘要』。

有四个链接/辅助视图：

- 预测能力（缺省视图）。 有关更多信息，请参阅主题 第 99 页的『预测能力』。
- 字段表。 有关更多信息，请参阅主题 第 99 页的『字段表』。
- 字段详细信息。 有关更多信息，请参阅主题 第 99 页的『字段详细信息』。
- 操作详细信息。 有关更多信息，请参阅主题 第 100 页的『操作详细信息』。

视图间链接

在主视图内，表格中的下划线文本控制链接视图中的显示。 单击文本将显示有关特定字段、字段集合或处理步骤的详细信息。 您最近一次选择的链接显示为深色，这可帮助您识别两个视图面板内容间的联系。

重置视图

要重新显示原始分析建议，并放弃对分析视图的任何更改，请单击主视图面板底部的**重置**。

字段处理概要

“字段处理摘要”表格提供了有关字段处理的预计总体影响的快照，包括对特征状态的更改和构建的特征数目。

请注意，这里不会实际构建模型，因此并不存在总体预测能力在数据准备前后的变化测量或图表，您只能显示单个建议预测变量的预测能力图表。

该表格显示以下信息：

- 目标字段数。

- 原始（输入）预测变量数。
- 建议用于分析和建模的预测变量。 其中包括建议字段总数；原始、未转换以及建议的字段数；建议的已转换字段数（排除任何字段的中间版本、从日期/时间预测变量派生的字段以及构建的预测变量）；从日期/时间字段派生的建议字段数，以及建议的构建字段数。
- 不建议在任何格式下使用（无论是在其原始格式下用作派生字段，还是用作构造预测变量的输入）的输入预测变量数。

如果任何**字段**信息带有下划线，单击可在链接视图中显示更多信息。 在“字段表链接视图”中显示**目标**、**输入特征**和**未使用输入特征**的详细信息。 有关更多信息，请参阅主题 第 99 页的『字段表』。 **建议在分析中使用的特征**将显示在“预测能力”链接视图中。 有关更多信息，请参阅主题 第 99 页的『预测能力』。

字段

“字段”主视图显示处理过的字段，以及 ADP 是否建议在下游模型中使用这些字段。 您可以覆盖任何字段建议；例如，排除构建的特征或包含 ADP 建议排除的特征。 如果字段已转换，您可以决定是接受建议转换，还是使用原始版本。

“字段”视图包含两个表，一个用于目标，另一个用于已处理或已创建的预测变量。

目标表

仅当数据中定义有目标时，才会显示**目标**表。

该表包含两列：

- **名称**。 此为目标字段的名称或标签；不论字段是否已转换，都始终使用原始名称。
- **测量级别**。 此列显示代表测量级别的图标；将鼠标悬停在图标上可以显示数据描述标签（连续、有序、名义等）。

  如果目标已转换，则**测量级别**列将反映最终转换版本。 注意：您不能关闭目标转换。

预测变量表格

预测变量表格总是显示。 表格的每一行代表一个字段。 缺省情况下，按预测能力的降序来排列行。

对于普通特征，原始名称始终用作行名称。 日期/时间字段的原始版本和派生版本都会显示在该表中（以单独的行显示）；该表还包括已构造的预测变量。

注意，在表格中显示的字段转换后版本始终代表最终版本。

缺省情况下，“预测变量”表中仅显示建议的字段。 要显示其余字段，选中表格上方的**在表中包括非推荐字段**复选框，这些字段随即显示在表格底部。

该表包含以下列：

- **要使用的版本**。 此列将显示一个下拉列表，以控制字段是否将在下游使用，以及是否使用建议的转换。 缺省情况下，下拉列表将反映建议。

  对于已变换的普通预测变量，下拉列表提供了以下三个选项：**已变换**、**原始**和**不使用**。

  对于未变换的普通预测变量，提供的选项为**原始**和**不使用**。

  对于派生的日期/时间字段和已构建的预测变量，提供的选项为**已变换**和**不使用**。

  对于原始日期字段，下拉列表被禁用，并设置为**不使用**。

  注: 对于同时具有原始版本和变换后版本的预测变量，如果切换**原始**和**已变换**版本，那么将自动更新这些特征的**测量级别**和**预测能力**设置。
- **名称**。 每个字段的名称均为链接。 单击名称可以在链接视图中显示有关该字段的更多信息。 有关更多信息，请参阅主题 第 99 页的『字段详细信息』。
- **测量级别**。 此列显示代表数据类型的图标；将鼠标悬停在图标上可以显示数据描述标签（连续、有序、名义等）。

- **预测能力**。 只会对 ADP 建议的字段显示预测能力。 如果未定义目标，那么不会显示此列。 预测能力范围从 0 到 1，其中较大的值表示“更好的”预测变量。 通常，预测能力对于比较一个 ADP 分析内的预测变量有用，但不应跨分析比较预测能力值。

操作摘要

对于自动数据准备所执行的各个操作，对输入预测变量进行了变换和/或过滤；执行某个操作后保留的字段将用于下一个操作。 在最后步骤中保留的字段将被建议用于建模，而转换和构造预测变量的输入将被过滤掉。

“操作摘要”是一个样本表，该表列出了 ADP 所执行的处理操作。 如果任何**操作**带有下划线，单击可在链接视图中显示有关所执行操作的更多信息。 有关更多信息，请参阅主题 第 100 页的『操作详细信息』。

注: 只会显示每个字段的原始版本和最终变换版本，而不会显示在分析过程中使用的任何中间版本。

预测能力

在首次运行分析时缺省显示，或者在“字段处理概要”主视图中选择了**建议在分析中使用的预测变量**时显示，该图表显示建议预测变量的预测能力。 字段按其预测能力排序，预测能力值最高的字段显示在顶端。

对于变换版本的普通预测变量，字段名称反映了您在“设置”选项卡的“字段名称”面板中选择的后缀；例如：*_transformed*。

测量级别图标显示在各个字段名称之后。

每个建议预测变量的预测能力通过线性回归或朴素贝叶斯模型模型进行计算，具体取决于目标是连续还是分类。

字段表

在“字段处理摘要”主视图中单击**目标**、**预测变量**或**未使用预测变量**时显示，“字段表”视图显示一个简单表，其中列出了相关特征。

该表包含两列：

- **名称**。 预测变量名。

  对于目标，这是所使用的字段的原始名称或标签，即使此目标已进行变换也是如此。

  对于变换版本的普通预测变量，名称反映了您在“设置”选项卡的“字段名称”面板中选择的后缀；例如：*_transformed*。

  对于派生自日期和时间的字段，将使用最终变换版本的名称；例如：*bdate_years*。

  对于构造的预测变量，将使用构造预测变量的名称；例如：*Predictor1*。
- **测量级别**。 此列显示代表数据类型的图标。

  对于目标，**测量级别**始终反映转换后的版本（如果目标已转换）。例如，从有序（有序集合）转换为连续（范围、尺度），反之亦然。

字段详细信息

在“字段”主视图中单击任何**名称**时显示，“字段详细信息”视图包括选定字段的分布、缺失值和预测能力图表（如果适用）。 此外，字段的处理历史记录和变换后的字段名称也将显示（如果适用）。

对于每个图表集，两个版本将并排显示，以比较字段在应用转换前后的情况；如果字段的转换后版本不存在，那么将只显示原始版本的图表。 对于派生日期或时间字段和构造的预测变量，将仅对新的预测变量显示图表。

注: 如果字段因为类别太多而遭排除，那么将仅显示处理历史记录。

分布图

连续字段分布显示为直方图，并叠放一条正态分布曲线，还有一条均值垂直参考线。类别字段显示为条形图。

直方图带有标签以显示标准差和偏度。不过，如果值个数等于或低于 2，或原始字段的方差低于 10-20，则不会显示偏度。

将鼠标悬停在图表的上方，可以显示直方图的均值，或条形图中类别计数与占记录总数的百分比。

缺失值图表

该图表显示为饼图，以比较在应用转换前后的缺失值百分比。图表标签显示百分比。

如果 ADP 执行了缺失值处理，则转换后的饼图还应包含替换值作为标签，即用于替换缺失值的值。

将鼠标悬停在图表的上方，可以显示缺失值计数和占记录总数的百分比。

预测能力图表

对于建议的字段，以条形图形式显示转换前后的预测能力。 如果目标已经过转换，则计算的预测能力对应于转换后的目标。

注: 如果未定义目标，或者在“主视图”面板中单击了目标，那么不会显示预测能力图表。

将鼠标悬停在图表的上方，可以显示预测能力值。

处理历史记录表

该表格显示字段的转换后版本是如何派生的。 ADP 采取的操作按照其执行顺序列出。不过，对于某些步骤，可能对特定字段执行了多个操作。

注: 对于未变换的字段，不会显示此表格。

此表中的信息分为两列或三列：

- **操作**。 操作的名称。 例如，连续预测变量。 有关更多信息，请参阅主题 第 100 页的『操作详细信息』。
- **详细信息**。 所执行处理的列表。 例如，转换成标准单位。
- **函数**。 仅针对构建的预测变量显示，显示输入字段的线性组合，例如，0.06*age + 1.21*height。

操作详细信息

在“操作摘要”主视图中选择任何带有下划线的**操作**时显示，“操作详细信息”链接视图显示所执行的每个处理步骤的操作相关与通用信息。首先显示操作相关的详细信息。

对于每个操作，描述用作标题位于链接视图的顶部。 特定于操作的详细信息在标题下方显示，并且可能包括有关以下各项数目的详细信息：派生的预测变量、已重新强制转换的字段、目标变换、已合并或重新排序的类别，以及已构造或排除的预测变量。

处理各项操作时，处理过程中使用的预测变量数可能会发生更改，例如，由于排除或合并预测变量而发生的更改。

注: 如果某项操作已关闭，或者未指定任何目标，那么在“操作摘要”主视图中单击该操作时，将会显示错误消息而非操作详细信息。

有 9 个可能的操作，不过对于每个分析而言，这些操作并非都有必要使用。

“文本字段”表

该表显示下列项的数目：

- 被去除的尾部空白值。
- 已从分析中排除的预测变量。

“日期和时间预测变量”表

该表显示下列项的数目：

- 派生自日期和时间预测变量的持续时间。

- 日期和时间元素。
- 派生的日期和时间预测变量（总计）。

如果已计算了任何日期持续时间，则参考日期或时间将显示为脚注。

“预测变量筛选”表

该表显示从处理中排除的下列预测变量的数目：

- 常量。
- 具有过多缺失值的预测变量。
- 单一类别中具有过多观测值的预测变量。
- 具有过多类别的名义字段（集合）。
- 已筛选掉的预测变量（总计）。

“检查测量级别”表

该表显示重新设计、分解成以下项的字段数目：

- 已重新强制转换为连续字段的有序字段（有序集）。
- 已重新强制转换为有序字段的连续字段。
- 重新设计总数。

如果没有连续或有序的输入字段（目标或预测变量），那么此部分将显示为脚注。

“离群值”表

该表显示离群值处理方式的计数。

- 发现并修整其离群值的连续字段数，或发现离群值并将其设为缺失值的连续字段数，具体取决于您在“设置”选项卡的“准备输入和目标”面板上的设置。
- 由于在离群值处理后为常量，而被排除的连续字段数。

一个脚注显示离群值分界值；而另一脚注在没有连续输入字段（目标或预测变量）时显示。

“缺失值”表

该表显示已替换缺失值、分解为以下项目的字段数：

- 目标。 如果未指定目标，那么将不显示此行。
- 预测变量。 它将进一步分解为名义（集合）、有序（有序集合）和连续特征数。
- 被替换的缺失值总数。

目标表

该表显示目标是否被转换，显示为：

- 到正态的 Box-Cox 转换。 这将进一步分解为显示指定标准（均值和标准差）和 Lambda 的列。
- 对其重新排序以提高稳定性的目标类别。

“分类预测变量”表

该表显示分类预测变量数：

- 其类别已按从低到高顺序进行重新排序以提高稳定性的预测变量。
- 合并其类别以最大化目标关联。
- 合并其类别以处理松散类别。
- 由于与目标关联程度过低而被排除。

- 由于在合并后为常量而被排除。

如果没有分类预测变量，那么将显示一个脚注。

“连续预测变量”表

有两个表。 第一个表格显示以下转换数之一：

- 已变换为标准单位的预测变量值。 此外，还显示已变换的预测变量数、指定的均值和标准偏差。
- 已映射至公共范围的预测变量值。 此外，还显示使用最值法变换进行变换的预测变量数，以及指定的最小值和最大值。
- 已分级的预测变量值和已分级的预测变量数。

第二个表显示预测变量空间构造详细信息，此信息将显示为预测变量数：

- 已构造。
- 由于与目标关联程度过低而被排除。
- 由于在离散化后为常量而被排除。
- 由于在构建后为常量而被排除。

如果没有任何连续预测变量作为输入，那么将显示一个脚注。

生成“派生”节点

当您生成派生节点时，它会将目标逆转换应用到评分字段。 缺省情况下，节点输入由自动建模节点（如“自动分类器”或“自动数值”）或“整体”节点生成的评分字段名称。 如果已转换刻度（范围）目标，那么评分字段将以转换后的单位显示；例如，log($) 而不是 $。为了解释和使用结果，您必须将预测值转换回原始刻度。

注: 只有当 ADP 节点包含对范围目标重新标度（即，在“准备输入和目标”面板上选择了 Box-Cox 重定比）的分析时，才能生成“派生”节点。 如果目标不是一个范围，或未选中 Box-Cox 重定比，则不能生成派生节点。

在“多个”模式下创建派生节点，并在表达式中使用 @FIELD，以便可以在需要时添加转换后的目标。 例如，使用以下详细信息：

- 目标字段名称：response
- 变换后的目标字段名称：response_transformed
- 评分字段名称：$XR-response_transformed

“派生”节点将创建一个新字段：$XR-response_transformed_inverse。

注: 如果您未使用自动建模节点或“整体”节点，那么需要编辑“派生”节点，以便为模型转换正确的评分字段。

标准化连续目标

缺省情况下，如果在“准备输入和目标”面板上选中**使用 Box-Cox 变换重新调整连续目标**复选框，那么将转换目标，并创建一个新字段，该字段将作为模型构建的目标。 例如，如果原始目标为 *response*，则新目标将为 *response_transformed*。ADP 节点的模型下游将自动选取该新目标。

但这可能会引发问题，具体取决于原始目标。 例如，如果目标为 *Age*，则新目标的值将不是 *Years*，而是 *Years* 的转换版本。 这意味着，由于它们的单位不可识别，因此您无法看到评分和解释。 这时，您可以应用逆转换，将转换后的单位转回到它们原来的含义。 为此：

1. 单击**分析数据**以运行 ADP 分析，然后从生成菜单中选择派生节点。
2. 在模型工作区上，将派生节点放置在模型块后面。

派生节点会将评分字段恢复为原始维度，这样预测值的单位将为 *Years*。

缺省情况下，“派生”节点会转换由自动建模节点或整体模型生成的评分字段。 如果构建单独的模型，那么需要编辑“派生”节点，以便从实际评分字段中派生。 如果要对您的模型进行评估，则应当将转换后的目标添加

到派生节点的**导出自**字段中。 这也会将相同的逆转换应用到目标，并且任何下游评估或分析节点都能正确地使用已转换数据，只要您将其切换到使用字段名，而不是元数据。

如果还想恢复原始名称，那么可以使用“过滤”节点移除原始目标字段（如果仍然存在），然后重新命名目标和评分字段。

类型节点

字段属性可在源节点中指定也可在单独的“类型”节点中指定。 两种节点的功能相似。 可用的属性如下：

- **字段** 双击任何字段名均可指定 IBM SPSS Modeler 中的数据的值标签和字段标签。 例如，从 IBM SPSS Statistics 导入的字段元数据可在此处查看或修改。 与之相似，您也可以为字段及其值创建新的标签。 您在此处指定的标签将根据您在“流属性”对话框中的选项显示在整个 IBM SPSS Modeler 中。
- **测量** 这是测量级别，用于描述给定字段中的数据的特征。 如果已经了解某个字段的所有详细信息，则称为**已完全实例化**。 有关更多信息，请参阅 第 104 页的『测量级别』。

  注: 字段的测量级别与字段的存储类型不同，后者指示数据是作为字符串、整数、实数、日期、时间、时间戳记还是列表进行存储。
- **值** 此列使您能够指定用于从数据集中读取数据值的选项，或者使用**指定**选项在单独的对话框中指定测量级别和值。 您还可以选择遍历字段，而不读取它们的值。 有关更多信息，请参阅 第 107 页的『数据值』。

  注: 如果相应的**字段**条目包含列表，那么您无法对此列中的单元格进行修改。
- **缺失** 用于指定如何处理此字段的缺失值。 有关更多信息，请参阅 第 110 页的『定义缺失值』。

  注: 如果相应的**字段**条目包含列表，那么您无法对此列中的单元格进行修改。
- **检查** 在此列中，您可以设置选项，以确保字段值符合指定的值或范围。 有关更多信息，请参阅 第 110 页的『检查类型值』。

  注: 如果相应的**字段**条目包含列表，那么您无法对此列中的单元格进行修改。
- **角色** 用于向建模节点指示字段将成为用于机器学习过程的**输入**（预测变量字段）还是**目标**（预测的字段）。 **两者**、**无**以及**分区**也是可用角色，最后一个可用角色表明字段用于将记录分区到不同的样本中，以用于进行训练、检验和验证。 值 **分割** 指定将为字段的每个可能值构建单独的模型。 有关更多信息，请参阅 第 111 页的『设置字段角色』。

使用类型节点窗口可以指定另外一些选项：

- 使用“工具”菜单按钮，可以选择在某个类型节点已实例化（通过规范设置、读取值或运行流）时**忽略唯一性字段**。 忽略唯一性字段将自动忽略仅有一个值的字段。
- 使用“工具”菜单按钮，可以选择在某个类型节点已实例化时**忽略大型集合**。 忽略大型集合将自动忽略具有大量成员的集合。
- 使用工具菜单按钮，您可在实例化类型节点后选择**转换连续整数为有序**。 有关更多信息，请参阅主题 第 106 页的『转换连续数据』。
- 使用“工具”菜单按钮，可以生成过滤节点以丢弃选定的字段。
- 使用墨镜切换按钮，可以将所有字段的缺省值设置为“读取”或“遍历”。 缺省情况下，源节点中的“类型”选项卡将传递字段，而类型节点本身则会读取值。
- 使用 **清除值** 按钮，可以清除在该节点中对字段值所做的更改（非继承值），并重新读取上游操作的值。此选项对于重置在上游对特定字段所进行的更改十分有用。
- 使用 **清除所有值** 按钮，可以重置读入该节点的 **所有** 字段的值。 此选项可以有效地针对所有字段将 **值** 列设置为 **读取** 。 此选项对于重置所有字段值以及重新读取上游操作的值和类型十分有用。
- 使用上下文菜单，可以选择将属性从一个字段 **复制** 到另一个字段。 有关更多信息，请参阅主题 第 111 页的『复制类型属性』。
- 使用 **查看未使用的字段设置** 选项，可以查看不再存在于数据中或曾经连接到该类型节点的字段的类型设置。 此选项在对已更改的数据集重新使用某个类型节点时十分有用。

测量级别

测量级别（以前称为“数据类型”或“用途类型”）用于描述数据字段在 IBM SPSS Modeler 中的用法。 测量级别可以在源节点或“类型”节点的“类型”选项卡中指定。 例如，您可能希望将值为 1 和 0 的某个整数字段的测量级别设置为标志。 这通常表明 1 = 真 ， 0 = 假 。

存储与测量。 请注意，字段的测量级别不同于字段的存储类型，后者是指数据的存储形式是字符串、整数、实数、日期、时间还是时间戳记。 数据类型可以使用类型节点在流中的任意位置进行修改，而存储类型必须在将数据读入 IBM SPSS Modeler 时在源中确定（当然，之后也可以使用转换函数对其进行更改）。 有关更多信息，请参阅主题 第 6 页的『设置字段存储类型和格式』。

某些建模节点通过其“字段”选项卡上的图标指示其输入字段和目标字段所允许的测量级别类型。

测量级别图标

表 *21:* 测量级别图标

图标	测量级别
	缺省
	连续
	分类
	标志
	名义
	有序
	无类型
[ ]	收集
	地理空间

可以使用以下测量级别：

- **缺省值** 其存储类型和值未知（例如，因为尚未读取）的数据显示为 **<Default>**。
- **连续** 用于描述数字值，例如范围 0-100 或 0.75-1.25。 连续值可以是整数、实数或日期/时间。
- **分类** 在不同值的准确数目未知时用于字符串值。 这是一种 **非实例化** 数据类型，表示有关数据存储类型和用法的所有可用信息均未知。 读取数据后，测量级别将为标志、名义或无类型，具体取决于在“流属性”对话框中指定的名义字段的最大成员数。
- **标志** 用于具有两个不同值的数据，这两个值用于指示特性存在与否（例如 `true` 与 `false`、`Yes` 与 `No` 或者 `0` 与 `1`）。 使用的值可以有所不同，但是必须始终将一个值指定为“true”值，将另一个值指定为“false”值。 数据可表示为文本、整数、实数、日期、时间或时间戳记。
- **名义** 用于描述具有多个不同值的数据，其中的每个值都被视为集合的一个成员，例如 `small/medium/large`。 名义数据可具有任何存储数值、字符串或日期/时间。 请注意，将测量级别设置为名义不会自动将值更改为字符串存储。
- **有序** 用于描述具有多个顺序固定的不同值的数据。 例如，工资类别或满意度排秩可以归类为有序数据。 顺序由数据元素的自然排列顺序定义。 例如，`1, 3, 5` 是一组整数的缺省排序顺序，而 `HIGH, LOW, NORMAL`（按字母顺序升序）是一组字符串的顺序。 使用有序测量级别可以将一组分类数据定义为有序数据，以进行可视化处理、模型构建以及导出到将有序数据识别为不同类型的其他应用程序（如 IBM SPSS Statistics）。 您可以在任何能够使用名义字段的位置使用有序字段。 此外，可以将任何存储类型（实数、整数、字符串、日期、时间等等）的字段定义为有序。
- **无类型** 用于不属于上述任何类型的数据、具有单个值的字段或者集合成员数超过定义的最大数目的名义数据。 当测量级别为包含许多成员（如帐号）的集合时，这种类型也将十分有用。 When you select 字段的

无类型，角色将自动设置为**无**，并且**记录标识**将作为唯一的替代项。 集合的最大缺省容量为 250 个唯一值。 可在“流属性”对话框（通过“工具”菜单访问）的“选项”选项卡中调整或禁用该数字。

- **集合** 用于标识列表中记录的非地理空间数据。 集合实际上是深度为零的列表字段，该列表中的元素具有另外某种测量级别。

  有关列表的更多信息，请参阅《SPSS Modeler 的“源”节点、“过程”节点和“输出”节点》指南的『“源”节点』部分中的『列表存储以及相关联的测量级别』主题。

- **地理空间** 与“列表”存储类型配合使用以标识地理空间数据。 列表可以是列表深度介于 0 与 2（含首尾值）之间的“整数列表”或“实数列表”字段。

  有关更多信息，请参阅《SPSS Modeler 的“源”节点、“过程”节点和“输出”节点》指南的『“类型”节点』部分中的『地理空间测量子级别』主题。

可以手动指定测量级别，也可以由软件读取数据并根据所读取的值确定其测量级别。

此外，如果有多个连续数据字段需视为类别数据，可以选择一个用于对这些字段进行转换的选项。 有关更多信息，请参阅主题 第 106 页的『转换连续数据』。

要使用自动输入

1. 在“类型”节点或者源节点的“类型”选项卡中，将**值**列设置为 **<Read>** 以获取期望的字段。 此操作将使元数据可用于所有下游节点。 您可以使用对话框上的太阳镜按钮快速将所有字段设置为 **<Read>** 或 **<Pass>**。
2. 单击 **读取值** 可立即读取数据源中的值。

要为字段手动设置测量级别

1. 在表中选择一个字段。
2. 在**测量**列的下拉列表中为该字段选择测量级别。
3. 或者，可以先采用 Ctrl+A 或按住 Ctrl 并单击的方式选择多个字段，再使用下拉列表选择测量级别。

地理空间测量子级别

与“列表”存储类型配合使用的“地理空间”测量级别具有 6 个子级别，这些子级别用于标识不同类型的地理空间数据。

- **点** - 指示特定位置；例如城市中心。
- **多边形** - 点序列，用于标识区域及其位置（例如城镇）的单个边界。
- **线串** - 也称为“折线”或简称“线条”，“线串”是点的序列，用于标识线条路径。 例如，线串可能是固定项，例如道路、河流或铁路；或者是移动的对象的轨迹，例如飞机的飞行路径或轮船的航线。
- **多点** - 在数据中的每一行都包含每个区域的多个点时使用。 例如，如果每一行都表示城市街道，那么可以使用每条街道的多个点来标识每个街灯。
- **多多边形** - 在数据中的每一行都包含多个多边形时使用。 例如，如果每一行都表示国家或地区的轮廓，那么可以将美国记录为多个多边形以标识不同的地区，例如本土、阿拉斯加和夏威夷。
- **多线串** - 在数据中的每一行都包含多条线条时使用。 由于线条无法分支，因此您可以使用“多线串”来标识一组线条。 例如，每个国家或地区的通航水道或铁路网络之类的数据。

这些测量子级别与“列表”存储类型配合使用。 有关更多信息，请参阅 第 8 页的『列表存储以及相关联的测量级别』。

限制

在使用地理空间数据时，您必须了解一些限制。

- 坐标系可能会影响数据格式。 例如，投影坐标系使用坐标值 x 和 y 并在需要时使用 z，而地理坐标系使用坐标值经度和纬度并在需要时使用高度值或深度值。

  有关坐标系的更多信息，请参阅《SPSS Modeler 用户指南》的『使用流』部分中的『设置流的地理空间选项』主题。

- 线串不得与自身交叉。
- 多边形不会自行闭合；对于每个多边形，您必须确保将第一个点与最后一个点定义为相同。
- 多多边形中的数据方向至关重要；顺时针方向表示实心形状，而逆时针方向表示空心形状。 例如，如果您记录国家或地区中存在湖泊的地区，那么可以按顺时针方向记录大陆地区的边界，并按逆时针方向记录每个湖泊的形状。
- 多边形不得与自身相交。 这种相交的一个示例是，您尝试将多边形边界绘制成图 8 所示的连续线条形状。
- 多多边形不得相互重叠。
- 对于地理空间字段，唯一的相关存储类型为**实数**和**整数**（缺省设置为**实数**）。

地理空间测量子级别图标

表 *22:* 地理空间测量子级别图标

图标	**测量级别**
	点
	多边形
	线串
	多点
	多多边形
	多线串

转换连续数据

将分类数据视为连续可能对模型的质量产生严重影响，特别是它作为目标字段的情况下，例如，生成回归模型而不是二元模型。 为避免这种情况，可以将整数范围转换成类别类型，例如有序或标志。

1. 在“操作和生成”菜单按钮（带有工具符号）中，选择**转换连续整数为有序**。 此时，将显示转换值对话框。
2. 指定将自动转换的范围大小，这会应用到小于和等于输入大小的任何范围。
3. 单击**确定**。 受影响的范围转换为标志或有序，并显示在类型节点的“类型”选项卡上。

转换结果

- 如果某个以整数形式存储的连续字段转换为有序，则上限和下限值将扩展以包括从下限值到上限值之间的所有整数值。 例如，如果范围为 1, 5，则值集合为 1, 2, 3, 4, 5。
- 如果连续字段转换为标志时，下限值和上限值成为标志字段的真值和假值。

什么是实例化?

实例化是读取或指定信息（如数据字段的存储类型和值）的过程。 为优化系统资源，实例化是一种用户导向过程 - 您通过在源节点的“类型”选项卡中指定选项，或通过“类型”节点运行数据，指导软件读取值。

- 类型未知的数据也称为非实例化数据。 存储类型和值未知的数据在“类型”选项卡的测量列中显示为**<Default>**。
- 如果已知字段存储类型的某些相关信息（如字符串或数字），那么这种数据称为部分实例化。 **分类**或**连续**都是部分实例化测量级别。 例如，**分类**指定字段为符号，但无法得知其测量级别是名义、有序还是标志。
- 当某个类型的所有相关详细信息（包括值）均已知时，将在此列中显示一种完全实例化测量级别 - 名义、有序、标志或连续。 请注意，连续类型可用于部分实例化和完全实例化的数据字段。 连续数据可以是整数，也可以是实数。

在通过“类型”节点执行数据流的过程中，非实例化类型将立即根据初始数据值变为部分实例化类型。 通过节点传递所有数据后，除非将值设置为 **<Pass>**，否则所有数据都将完全实例化。如果执行中断，那么数据将保持部分实例化状态。 “类型”选项卡实现实例化后，字段的值在流的这一点上是静态的。 这意味着，任何上游更改都不会影响某个特定字段的值，即使重新运行流也是如此。 要根据新数据或添加的操作来更改或更新值，就需要在“类型”选项卡本身中编辑这些内容，或者将字段的值设置为 **<Read>** 或 **<Read +>**。

何时进行实例化

通常，如果数据集不是非常大，并且不打算稍后在流中添加字段，那么在源节点上进行实例化是最方便的方法。 但对于下列情况，在单独的类型节点中进行实例化更为实用：

- 数据集较大，且流在类型节点之前过滤子集。
- 数据已在流中完成过滤。
- 数据已在流中完成合并或追加。
- 在处理过程中有新的数据字段被导出。

注: 如果要在数据库导出节点中导出数据，那么该数据必须完全实例化。

数据值

使用“类型”选项卡的**值**列，可以自动读取数据的值，也可以在单独的对话框中指定测量级别和值。

“值”下拉列表中的选项提供了有关自动输入的指示信息，如下表中所示。

表 *23:* 自动输入的指示信息

选项	**函数**
<Read>	执行节点时将读取数据。
<Read+>	将读取数据并附加到当前数据（如果存在）。
<Pass>	未读取任何数据。
<Current>	保留当前数据值。
指定...	将打开另一个对话框，以便您指定值和测量级别选项。

执行“类型”节点或单击**读取值**将根据您的选择进行自动归类并从数据源中读取值。 此外，也可以使用“指定”选项或通过在**字段**列中双击单元格来手动指定这些值。

在“类型”节点中对字段进行更改后，您可以使用对话框工具栏中的下列按钮重置值信息：

- 通过使用 **清除所有值** 按钮，可以清除对此节点中的字段值 (非继承值) 所作的更改，并从上游操作重新读取值。 此选项对于重置在上游对特定字段所进行的更改十分有用。
- 通过使用 **清除值** 按钮，您可以重置读入节点的 **所有** 字段的值。 此选项可以有效地针对所有字段将 值 列设置为 **读取** 。 此选项对于重置所有字段值以及重新读取上游操作的值和测量级别十分有用。

值列中的灰色文本

在“类型”节点或“源”节点中，如果**值**列中的数据以黑色文本显示，那么它表示该字段的值已读取并存储在该节点中。 如果此字段中不存在黑色文本，那么该字段的值未读取并且是确定的进一步上游。

在某些情况下，您可以看到数据显示为灰色文本。 当 SPSS Modeler 可以确定或推断出某个字段的有效值而无需实际读取和存储数据时，会发生此情况。 如果您使用下列其中一个节点，那么可能发生此情况：

- 用户输入节点。 由于数据是在节点中定义的，因此某个字段的值范围始终是已知的，即使这些值尚未存储在该节点中也是如此。
- Statistics 文件源节点。 如果存在对应于数据类型的元数据，那么它使 SPSS Modeler 可以推断出值的可能范围，而无需读取或存储数据。

在上述任何一种节点中，值将以灰色文本显示，直到您单击**读取值**为止。

警告: 如果您未将流中的数据实例化，并且数据值以灰色显示，那么不会对您在**检查**列中设置的类型值进行任何检查。

使用值对话框

在“类型”选项卡中单击**值**或**缺失**列显示预定义值的下拉列表。 选择此列表上的**指定...** 选项将打开一个单独的对话框，您可以在其中设置用于读取、指定、标注和处理所选字段的值的选项。

很多控件是所有数据类型通用的。 下面介绍这些通用控件。

测量 显示当前选择的测量级别。 您可以更改设置以反应希望使用数据的方式。 例如，如果名为 `day_of_week` 的字段包含代表各天的数字，您可能希望将此更改为名义数据，以创建用于分别检查每个类别的分布节点。

存储 显示存储类型（如果已知）。 存储类型不受您选择的测量级别的影响。 要改变存储类型，可以使用“固定文件和可变文件”源节点中的“数据”选项卡或使用“过滤”节点中的转换函数。

模型字段 对于对模型块进行评分时生成的字段，还可以查看模型字段的详细信息。 这些详细信息包括目标字段的名称以及建模时此字段的角色（预测值、概率和倾向等等）。

值 选择用于确定所选字段的值的方法。 您在此做出的选择将覆盖之前在类型节点对话框的 **值** 列中进行的任何选择。 用于读取值的选项包括：

- **从数据中读取** 选择此项表示在执行节点时读取值。 此选项与**<Read>**相同。
- **经过** 选择此项表示不读取当前字段的数据。 此选项与**<Pass>**相同。
- **指定值和标签** 这里的选项用于指定所选字段的值和标签。 将此选项与值检查功能配合使用，可以根据您对当前字段的了解指定值。 此选项可针对不同字段类型激活该类型所特有的控件。 后续主题将分别介绍用于值和标签的选项。

  注: 对于测量级别为“无类型”或“<缺省>”的字段，无法指定值或标签。
- **使用数据扩展值** 选择此项可以对当前数据追加此处输入的值。 例如，如果 `field_1` 的范围为 (0,10)，您输入 (8,16) 中的一系列值，那么将通过添加 16 扩展范围，而不除去原始最小值。 新的范围将是 (0,16)。选择此选项会自动将自动输入选项设置为 **<Read+>**。

最大列表长度 仅适用于测量级别为“地理空间”或“集合”的数据。 通过指定列表可以包含的元素数目来设置列表的最大长度。

最大字符串长度 仅适用于无类型数据；在生成 SQL 以创建表时使用该字段。 输入数据中最大字符串的值；这样会在表中生成一个足够容纳该字符串的列。 如果字符串长度值不可用，将使用可能不适用于该数据的缺省字符串大小（例如，如果值太小，向表中写入数据时可能会发生错误；如果值太大，可能会对性能产生不利影响。）

检查值 请选择强制转换值以使其符合指定的连续、标志或名义值的方法。 此选项与类型节点对话框中的 **检查** 列对应，在此进行的设置将覆盖该对话框中的设置。 通过将值检查功能与**指定值和标签**选项配合使用，可以使数据中的值与期望的值一致。 例如，如果指定值为 1、0，然后使用 **丢弃** 选项，则可以丢弃所有值不是 1 或 0 的记录。

定义空白值 选择此项可以激活下列控件，这些控件可用于声明数据中的缺失值或空白值。

- **缺失值** 使用此表可以将特定的值（例如 99 或 0）定义为空白值。 该值应适用于字段的存储类型。
- **范围** 用于指定缺失值的范围，例如，年龄 1-17 或大于 65。 如果将某个界限值保留为空，那么范围将不受限；例如，如果仅指定下限为 100 而未指定上限，那么会将所有大于或等于 100 的值定义为缺失。 界限值包括在内；例如，下限为 5 且上限为 10 的范围定义将包括 5 和 10。 可以为任何存储类型定义缺失值范围，这些类型包括日期/时间和字符串（在这种情况下，将采用字母排列顺序来确定某个值是否在范围内）。
- **Null/空格** 您还可以将系统 null 值（在数据中显示为 `$null$`）和空白（不带可见字符的字符串值）指定为空白值。

  注: 为了执行分析，“类型”节点还会将空字符串视为空白，尽管它们在内部以不同方式进行存储，并且在某些情况下以不同方式进行处理。

注: 要将空白值编码为未定义值 `$null$`，请使用“填充”节点。

描述 使用此文本框可以指定字段标签。 这些标签将根据您在“**流属性**”对话框中选择的选项出现在各种位置，例如出现在图形、表、输出和模型浏览器中。

指定连续数据的值和标签

连续测量级别仅用于数值字段。 连续数据的存储类型有以下三种：

- 实数
- 整数
- 日期/时间

所有连续字段都将通过同一个对话框进行编辑，显示的存储类型仅供参考。

指定值

以下控件是连续字段所独有的，用于指定值的范围：

下限。 指定值范围的下限。

上限。 指定值范围的上限。

指定标签

可以为范围字段的任意值指定标签。 单击 **标签**按钮可打开一个新的对话框，用于指定值标签。

值和标签子对话框

在范围字段的“值”对话框中单击**标签**将打开一个新的对话框，您可在其中指定该范围内任意值的标签。

可以使用此表中的 值 和 标签 列定义值和标签对。 当前已定义的对将在此显示。 通过单击空单元格并输入值及其对应标签，可以添加新的标签对。 注意：向此表添加值/值-标签对不会向字段添加任何新值。 该操作只是创建字段值的元数据而已。

您在“类型”节点中指定的标将按按照您在“流属性”对话框中选择的选项显示在多个位置（显示为工具提示、输出标签等）。

指定名义和有序数据的值和标签

名义（集合）和有序（有序集合）测量级别表明数据值将分别用作集合的一个成员。 集合的存储类型可以是字符串、整数、实数或日期/时间。

以下控件是名义和有序字段独有的，用于指定值和标签：

值。 您可以使用 **值** 列根据您对当前字段的了解来指定值。 通过使用此表，您可以输入字段的期望值，并使用 **检查值** 列表来检查数据集是否符合这些值。 通过使用箭头和删除图标，您可以修改现有值以及重新排序或删除值。

标签。 可以使用 **标签** 列为集合中的每个值指定标签。 这些标签显示在各种位置，例如图形，表，输出和模型浏览器中，具体取决于您在 "流属性" 对话框中所作的选择。

指定标志的值

标志字段用于显示具有两个不同值的数据。 标记的存储类型可以是字符串、整数、实数或日期/时间。

True。 指定条件成立时字段的标志值。

False。 指定条件不成立时字段的标志值。

标签。 为标志字段中的每个值指定标签。 这些标签将按照您在“流属性”对话框中选择的选项出现在多个位置，如图形、表格、输出和模型浏览器中。

指定集合数据的值

集合字段用于显示列表中的非地理空间数据。

唯一可以为“集合”**测量**级别设置的项是**列表测量**。 缺省情况下，此测量设置为“无类型”，但您可以选择另一个值，以设置列表中的元素的测量级别。 您可以选择下列其中一个选项：

- 无类型
- 连续
- 名义
- 有序
- 标志

指定地理空间数据的值

地理空间字段用于显示列表中的地理空间数据。

对于地理空间**测量**级别，您可以设置下列选项，以设置该列表中的元素的测量级别：

类型 请选择地理空间字段的测量子级别。 可用的子级别由列表字段的深度确定；缺省值为："点"（深度为零）、"线串"（深度为 1）和"多边形"（深度为 1）。

有关子级别的更多信息，请参阅第 105 页的『地理空间测量子级别』。

有关列表深度的更多信息，请参阅第 8 页的『列表存储以及相关联的测量级别』。

坐标系 仅当您将测量级别由非地理空间级别更改为地理空间级别时，此选项才可用。 要对您的地理空间数据应用坐标系，请选中此复选框。 缺省情况下，将显示**工具 > 流属性 > 选项 > 地理空间**窗格中设置的坐标系。 要使用另一个坐标系，请单击**更改**按钮以显示"**选择坐标系**"对话框，并选择所需的坐标系。

有关坐标系的更多信息，请参阅《SPSS Modeler 用户指南》的『使用流』部分中的『设置流的地理空间选项』主题。

定义缺失值

"类型"选项卡的**缺失**列指示是否已为字段定义缺失值处理。 可能的设置为：

开启 (*)。指示为该字段定义了缺失值处理。 可使用下游填充节点，或使用"指定"选项（见下文）通过明确规范来进行此操作。

关闭。 字段没有定义缺失值处理。

指定。 选择此选项将显示对话框，您可在其中声明将明确值视为此字段的缺失值。

检查类型值

打开每个字段的"检查"选项将检查该字段中的所有值，以确定它们是否符合当前类型设置或已在"指定值"对话框中指定的值。 使用这种方法，单项操作即可实现对数据集的整理以及数据集规模的缩减。

类型节点对话框中 检查 列的设置将决定在发现超出类型限制的值时的操作。 要更改字段的"检查"设置，请使用检查列中该字段的下拉列表。 要设置所有字段的检查设置，请在字段列中单击并按 Ctrl-A。 然后，使用检查列中任何字段的下拉列表。

可用的"检查"设置如下：

无。 将遍历值而不进行检查。 这是缺省设置。

使无效。 将超出限制的值更改为系统空值 (`$null$`)。

强制。 将在测量级别已完全实例化的字段中查找超出指定范围的值。 未指定的值将被转换为该测量级别的合法值，应用的规则如下：

- 对于标志，将真值和假值以外的所有值都转换为假值。
- 对于集合（名义或有序），将所有未知值转换为集合值的第一个成员。
- 大于范围上限的数值将替换为上限。
- 小于范围下限的数值将替换为下限。
- 范围内的空值将获得该范围的中点值。

废弃。 在发现非法值时，丢弃整个记录。

警告。 读取所有数据后，会在“流属性”对话框中计算并报告非法项数。

中止。 遇到第一个非法值时，将终止流的运行。 错误将在“流属性”对话框中报告。

设置字段角色

字段的角色用于指定其在模型构建过程中的用法 - 例如，字段是输入还是目标（预测的对象）。

注: “分区”、“频率”和“记录标识”角色只能分别应用到单个字段。

可用的角色如下：

输入。 字段将用作机器学习的输入（预测变量字段）。

目标。 字段将用作机器学习的输出或目标（模型将尝试预测的字段之一）。

两者。 字段将被 Apriori 节点同时用作输入和输出。 所有其他建模节点都将忽略该字段。

无。 机器学习将忽略该字段。 测量级别已设置为**无类型**的字段将在**角色**列中自动设置为**无**。

分区。 指明字段用于将数据分区为单独的样本（用于训练、测试，也可用于验证）。 该字段必须属于实例化集合类型，具有两个或三个可能值（在“字段值”对话框中定义）。 第一个值表示训练样本，第二个值表示测试样本，第三个值（如果存在）表示验证样本。 所有其他值都将被忽略，且不能使用标志字段。 请注意，要在分析中使用分区，必须在相应的模型构建或分析节点的“模型选项”选项卡中启用分区。 启用分区时，会将对于分区字段具有空值的记录从分析中排除。 如果已在流中定义多个分区字段，那么必须在每个相应建模节点的“字段”选项卡中指定单一分区字段。 如果数据中不存在适合的字段，您可以使用“分区”节点或“派生”节点进行创建。 有关更多信息，请参阅主题 第 131 页的『分区节点』。

分割。 （仅名义、有序和标志字段）指定为字段的每个可能值构建一个模型。

频率。 （仅数字字段）设置此角色允许将字段值用作记录的频率加权因子。 仅 C&R 树、CHAID、QUEST 和线性模型支持此功能；所有其他节点将忽略此角色。 在支持此功能的建模节点的“字段”选项卡上，选择**使用频率权重**以启用频率加权。

记录标识。 此字段将用作唯一记录标识。 大多数节点都会忽略此特征；但它受线性模型支持，并且是 IBM Netezza 数据库内挖掘节点所必需的。

复制类型属性

可以轻松地将某种类型的属性（如值、检查选项和缺失值）从一个字段复制到另一个字段：

1. 右键单击要复制其属性的字段。
2. 在上下文菜单中选择 **复制** 。
3. 右键单击要更改其属性的字段。
4. 在上下文菜单中选择**选择性粘贴**。 注：您可以使用 Ctrl+单击方法或通过使用上下文菜单中的**选择字段**选项来选择多个字段。

此时将打开一个新的对话框，您可在其中选择要粘贴的特定属性。 如果要粘贴至多个字段，在此选择的选项将应用于所有目标字段。

粘贴以下属性。 在下面的列表中进行选择，将属性从一个字段粘贴至另一个字段。

- **类型**。 选择此选项可粘贴测量级别。
- **值**。 选择此选项可粘贴字段值。
- **缺失**。 选择此选项可粘贴缺失值设置。
- **选中这一项**。 选择此选项可粘贴值检查选项。
- **角色**。 选择此选项可粘贴字段的角色。

字段格式设置选项卡

“表”节点和“类型”节点的“格式”选项卡将显示当前字段或未用字段的列表，以及每个字段的格式设置选项。 下面是字段格式设置表中每个列的说明：

字段。 此列显示所选字段的名称。

格式。 通过双击此列中的单元格，可以使用打开的对话框指定各个字段的格式设置。 有关更多信息，请参阅主题 第 112 页的『设置字段格式选项』。 在此指定的格式设置将覆盖总体流属性中指定的格式设置。

注意：统计信息 导出节点和 统计信息 输出节点导出的 *.sav* 文件的元数据中包括根据字段选择格式设置。 如果指定了 IBM SPSS Statistics *.sav* 文件格式所不支持的根据字段格式，那么节点将采用 IBM SPSS Statistics 缺省格式。

对齐。 使用此列可指定表列中的值的对齐方式。 缺省设置为**自动**，该设置将对符号值进行左对齐，对数字值进行右对齐。 您可通过选择 **左** 、 **右** 或 **中心** 来覆盖缺省设置。

列宽。 缺省情况下，将根据字段值自动计算列宽。 要覆盖自动宽度计算，请单击表单元格，然后使用下拉列表选择新的宽度。 要输入此处未列出的定制宽度，请通过双击“字段”或“格式”列中的表单元格打开“字段格式”子对话框。 或者，也可以右键单击某个单元格，然后选择 **设置格式** 。

查看当前字段。 缺省情况下，对话框将显示当前处于活动状态的字段的列表。 要查看未使用字段的列表，请选择 **查看未使用的字段设置** 。

上下文菜单。 此选项卡的上下文菜单提供了多种选择和设置更新选项。 在列中单击鼠标右键可显示此菜单。

- **全选**。 选中所有字段。
- **选择“无”**。 清除选择内容。
- **选择字段**。 依据类型或存储特征选择字段。 选项包括**选择分类**、**选择连续**（数值）、**选择无类型**、**选择字符串**、**选择数值**或**选择日期/时间**。 有关更多信息，请参阅主题 第 104 页的『测量级别』。
- **设置格式**。 打开用于指定每个字段的日期、时间和小数选项的子对话框。
- **设置对齐方式**。 设置所选字段的对齐方式。 选项包括 **自动** 、 **中心** 、 **左** 或 **右** 。
- **设置列宽**。 设置所选字段的字段宽度。 指定 **自动** 将从数据中读取宽度。 您也可以将字段宽度设为 5、10、20、30、50、100 或 200。

设置字段格式选项

字段格式设置在一个子对话框中进行指定，该对话框在“类型”节点和“表格”节点的“格式”选项卡中提供。 如果在打开此对话框之前已选择多个字段，那么选中的第一个字段的设置将用于所有选中字段。 在此进行指定后单击**确定**会将这些设置应用于“格式”选项卡中选定的所有字段。

以下选项针对每个字段提供。 其中很多设置也可以在“流属性”对话框中指定。 任何在字段级别进行的设置都将覆盖流指定的缺省设置。

日期格式。 选择日期存储字段要使用的日期格式或 CLEM 日期函数将字符串解析为日期时使用的日期格式。

时间格式。 选择时间存储字段要使用的时间格式或 CLEM 时间函数将字符串解析为时间时使用的时间格式。

数字显示格式。 可以在“标准”(**####.###**)、“科学表示法”(**#.###E+##**) 和“货币”(**$###.##**) 显示格式中选择。

十进制符号。 选择逗号 (,) 或句号 (.) 作为小数分隔符。

分组符号。 针对数字显示格式，选择用于对值进行分组的符号（例如，*3,000.00* 中的逗号）。 选项包括“无”、“句号”、“逗号”、“空格”和“定义的语言环境”（在该情况下将采用当前语言环境的缺省设置）。

小数位（标准、科学表示法、货币和导出）。 针对数字显示格式，指定要在显示实数时使用的小数位数。 此选项将分别为每种显示格式指定一个值。 请注意，**导出小数位**设置仅适用于平面文件导出，将覆盖流属性。 平面文件导出的流缺省值是为流属性中的**标准小数位**设置指定的任何值。 “XML 导出”节点导出的小数位数始终为 6。

对齐。 指定列中的值应采用的对齐方式。 缺省设置为**自动**，该设置将对符号值进行左对齐，对数字值进行右对齐。 您可通过选择“左”、“右”或“中心”来覆盖缺省设置。

列宽。 缺省情况下，将根据字段值自动计算列宽。 您可使用列表框右边的箭头指定以五为间隔的定制宽度。

过滤或重命名字段

您可以在流中的任意时间点上重命名或排除字段。 例如，作为医学研究人员，您可能不关心患者（记录级别数据）的钾水平（字段级别数据）；因此，您可以过滤掉 K（钾）字段。 也可以使用单独的“过滤”节点，或者源节点或输出节点上的“过滤”选项卡实现此操作。 无论使用哪种节点，结果都是一样的。

- 可以在将数据从源节点（如变量文件、固定文件、统计信息 文件、XML 或扩展导入）读入 IBM SPSS Modeler 时对字段进行重命名或过滤。
- 使用过滤节点，可以在流的任何位置对字段进行重命名或过滤。
- 通过 统计信息 导出、统计信息 变换、统计信息 模型和 统计信息 输出节点，可以对字段进行过滤或重命名，使之符合 IBM SPSS Statistics 命名标准。
- 您可以使用以上任何节点中的“过滤器”选项卡来定义或编辑多响应集。
- 最后，可以使用“过滤”节点将一个源节点中的字段映射至另一个源节点。

设置过滤选项

“过滤”选项卡中使用的表格可显示每个字段进入和离开节点时的名称。 可以使用此表中的选项对重复的或下游操作不需要的字段进行重命名或过滤。

- **字段**。 显示当前连接的数据源中的输入字段。
- **过滤**。 显示所有输入字段的过滤状态。 过滤后的字段在此列中带有红色 X，指示该字段不会向下游传递。单击选定字段的 过滤 列可打开和关闭过滤功能。 此外，也可以采用按住 Shift 并单击的选择方法同时选择多个字段的选项。
- **字段**。 在字段离开“过滤”节点时显示这些字段。 重复名称显示为红色。 可以通过单击此列并输入新名称来编辑字段名。 也可以通过单击 过滤 列删除字段，以禁用重复字段。

通过单击列标题，可以对表中所有列进行排序。

查看当前字段。 选择此选项可查看当前连接到“过滤”节点的数据集的字段。 此选项缺省处于选中状态，是最常用的过滤节点使用方法。

查看未使用的字段设置。 选择此选项可查看曾连接到“过滤”节点但已断开连接的数据集的字段。 将过滤节点从一个流复制到另一个流时，或保存并重新加载过滤节点时，此选项将十分有用。

过滤按钮菜单

单击对话框左上角的“过滤”按钮可访问含有多个快捷键和其他选项的菜单。

可以选择执行下列操作：

- 删除所有字段。
- 包括所有字段。
- 切换所有字段。
- 删除副本。 请注意，选择此选项将移除具有该重复名称的所有字段，包括第一次出现该名称的字段。
- 重命名字段和多重响应集以满足其他应用程序的要求。 有关更多信息，请参阅主题 第 278 页的『重命名或过滤 IBM SPSS Statistics 的字段』。
- 截断字段名。
- 匿名化字段和多重响应集名称。
- 使用输入字段名。
- 编辑多重响应集。 有关更多信息，请参阅主题 第 114 页的『编辑多重响应集』。
- 设置缺省过滤状态。

您还可以使用对话框顶部的箭头切换按钮指定要在缺省情况下包括还是丢弃字段。 对于要在下游纳入的字段很少的大型数据集，此方法十分有用。 例如，可以仅选择要保留的字段，并指定所有其他字段都应废弃（而不是逐一选择所有要废弃的字段）。

截断字段名

通过过滤按钮菜单（“过滤”选项卡的左上角），您可以选择截断字段名。

最大长度。 指定字段名称长度的限制字符数。

位数。 如果截断后的字段名称不再唯一，那么将对其进行进一步的截断，并通过向名称添加数字位来进行区分。 您可指定使用的数字位数。 使用箭头按钮可调整该位数。

例如，下表说明了如何使用缺省设置（最大长度=8，数字位数=2）对某个医学数据集中的字段名称进行截断。

表 *24:* 字段名称截断

字段名	**截断后的字段名**
Patient Input 1	Patien01
Patient Input 2	Patien02
Heart Rate	HeartRat
BP	BP

匿名化字段名称

通过单击左上角的过滤按钮菜单并选择**对字段名称进行匿名化**，您可以对包含“过滤”选项卡的任何节点中的字段名称进行匿名化。 已匿名化的字段名称由一个字符串前缀及一个基于数字的唯一值组成。

将名称匿名化。 选择**仅所选字段**将仅对“过滤”选项卡中的所选字段的名称进行匿名化。 缺省设置为**所有字段**，这表示将对所有字段名称进行匿名化。

字段名称前缀。 已匿名化的字段名称的缺省前缀为 **anon_**；如果要使用其他前缀，请选择**自定义**并键入您自己的前缀。

匿名化多重响应集。 按照对字段进行匿名化的方法对多重响应集名称进行匿名化。 有关更多信息，请参阅主题 第 114 页的『编辑多重响应集』。

要复原原始字段名称，请在过滤按钮菜单中选择**使用输入字段名称**。

编辑多重响应集

通过单击左上角的过滤按钮菜单并选择**编辑多重响应集**，可添加或编辑包含“过滤”选项卡的任意节点中的多重响应集。

多重响应集可用于记录对每个问题都具有多个值的数据，例如，询问被调查者参观过哪些博物馆或阅读过哪些杂志。 可以使用 数据收集 源节点或 Statistics 文件源节点将多重响应集导入 IBM SPSS Modeler 中，也可使用过滤节点在 IBM SPSS Modeler 中进行定义。

单击 **新建** 可创建新的多重响应集，或者单击 **编辑** 可修改现有的多重响应集。

名称和标签。 指定多重响应集的名称和描述。

类型。 可以使用下列两种方法之一处理多响应问题：

- **多二分集**。 为每个可能的响应创建一个单独的标志字段；例如，如果有 10 本杂志，那么将创建 10 个标志字段，其中每个字段都可以拥有值，如 0 或 1 分别表示真或假。 计数值可以指定哪个值为真。 通过该方法，有助于响应者选择适用的所有选项。
- **多类别集**。 为每个响应创建的一个名义字段，该响应中特定响应者提供的答案数量可以达到最多。 每个名义字段的值均表示可能的答案，如 *1* 表示《时代周刊》、*2* 表示《新闻周刊》以及 *3* 表示《个人计算机周刊》。 该方法在限制答案数量时非常有用，如让响应者选择最常看的三本杂志。

集合中的字段。 使用右侧的图标可添加或移除字段。

注释

- 多重响应集中的所有字段都必须具有同一存储类型。

- 这些集合不同于它们所含的字段。 例如，删除某个集合并不会删除它所包含的字段，只会删除这些字段之间的链接。 从删除点上游仍可以查看该集合，但在删除点下游将看不到该集合。
- 如果使用“过滤”节点重命名字段（直接通过选项卡，或者通过选择“过滤”菜单上的为 IBM SPSS Statistics 重命名、**截断**或**匿名化**选项），那么还将更新对多重响应集中使用的这些字段的所有引用。 但是，不会从多重响应集中移除通过“过滤”节点删除的任何字段。 尽管无法在流中再查看这些字段，但多重响应集仍可以引用它们；在导出等操作中，需要格外注意此问题。

“派生”节点

IBM SPSS Modeler 中最强大的功能之一是可以修改数据值并从现有数据中派生新字段。 在漫长的数据挖掘工程中，执行若干派生操作是很常见的，如从 Web 日志数据的字符串中抽取客户标识，或根据事务和人口统计数据创建客户生命周期值。 所有这些变换均可使用各种字段操作节点完成。

若干节点可提供导出新字段的功能：

“派生”节点修改数据值或者根据一个或多个现有字段创建新字段。 它创建类型为公式、标志、名义、状态、计数和条件的字段。

“重新分类”节点将一组分类值转换为另一组值。 对于折叠类别或者进行数据重新分组以执行分析而言，重新分类非常有用。

“分箱”节点根据一个或多个现有连续（数字范围）字段的值自动创建新的名义（集合）字段。 例如，您可以将连续收入字段转换为一个包含各组收入（作为与均值之间的偏差）的新分类字段。 一旦创建新字段分级后，即可根据割点创建“衍生”节点。

“设为标志”节点根据针对一个或多个名义字段定义的分类值派生多个标志字段。

“重构”节点将名义字段或标志字段转换为一组字段，这组字段可以使用另一字段的值进行填充。 例如，如果给定名为支付类型的字段，且值为信用、现金和借记，那么将创建三个新字段（信用、现金和 借记），其中每个字段都可能包含实际支付的值。

“历史记录”节点创建新字段，这些字段包含先前记录中的字段的数据。 “历史记录”节点最常用于顺序数据，例如时间序列数据。 在使用“历史记录”节点之前，您可能希望使用“排序”节点对数据进行排序。

使用派生节点

使用导出节点，可以根据一个或多个现有字段创建六种类型的新字段：

- **公式**。 新字段是任意 CLEM 表达式的结果。
- **标志**。 新字段是代表指定条件的标志。
- **名义**。 新字段是名义的，表示其成员是一组指定值。
- **状态**。 新字段是两种状态之一。 通过指定条件触发这两种状态之间的切换。
- **计数**。 新字段以满足某个条件的次数为基准。
- **条件**。 新字段是两个表达式的其中一个的值，具体取决于条件的值。

其中每个节点在 Derive 节点对话框中都包含一组特殊选项。 这些选项将在后续主题中进行论述。

请注意，使用以下各项可能会更改行顺序：

- 通过 SQL 回送在数据库中执行
- 通过远程 IBM SPSS Analytic Server 执行
- 使用在嵌入式 IBM SPSS Analytic Server 中运行的函数
- 派生列表（例如，请参阅第 118 页的『派生列表或地理空间字段』）
- 调用任何空间函数

为导出节点设置基本选项

在导出节点的对话框顶部，有用于选择所需导出节点类型的若干选项。

方式。 根据是否要派生多个字段，选择**单个**或**多个**。 选择**多个**时，对话框将变为显示用于多个派生字段的选项。

派生字段。 对于简单的 Derive 节点，请指定要派生并添加到每条记录的字段的名称。 缺省名称为 Derive*N*，其中 *N* 是截止到目前由当前会话所创建的派生节点数。

导出为。 从下拉列表中选择 Derive 节点的类型，例如“公式”或“名义”。 对于每个类型，都会根据您在该类型特定的对话框中指定的条件创建新字段。

从下拉列表中选择某个选项会将一组新的控件添加到主对话框，具体取决于每种 Derive 节点类型的属性。

字段类型。 为新派生的节点选择测量级别，如“连续”、“分类”或“标志”。 此选项对于导出节点的所有形式是通用的。

注意：导出新字段通常需要使用特殊的函数或数学表达式。 所有类型的派生节点的对话框中都提供了表达式构建器，用于帮助您创建这些表达式，并提供了规则检查和 CLEM 表达式的完整列表。

导出多个字段

在派生节点内将模式设置为**多个**可以依据同一节点内的同一条件派生多个字段。 如果要对数据集中的多个字段进行相同变换，使用此功能可节省时间。 例如，如果要构建一个回归模型，用于根据起始工资和原有经验预测当前工资，那么对所有三个非对称变量应用对数变换可能很有帮助。 您可以对所有字段同时应用同一函数，而不是针对每种转换添加一个新的“派生”节点。 只需选择要从中派生新字段的所有字段，然后使用 `@FIELD` 函数在字段括号内键入派生表达式。

注：`@FIELD` 函数是一个重要工具，用于同时派生多个字段。 它使您可以在无需指定确切字段名的情况下引用当前字段的内容。 例如，用于将对数变换应用于多个字段的 CLEM 表达式为 `log(@FIELD)`。

当您选择 **多个** 模式时，会将下列选项添加到对话框中：

派生自。 使用“字段选择器”选择要从其中派生新字段的字段。 对于每个选定字段，将生成一个输出字段。 注意：所选字段不需要具有同一存储类型；但是，如果条件并非对于所有字段均成立，那么派生操作将会失败。

字段名称扩展。 输入要向新字段名称添加的扩展名。 例如，对于包含 *Current Salary* 的新字段，可以为字段名添加扩展部分 *log_* ，从而产生 *log_Current Salary* 。 使用单选按钮选择要将扩展名添加为字段名的前缀（开头）还是后缀（末尾）。 缺省名称为 Derive*N*，其中 *N* 是截止到目前由当前会话所创建的派生节点数。

在单个模式的“派生”节点中，此时需要创建用于派生新字段的表达式。 根据所选派生操作的类型，可有多种创建条件的选项。 这些选项将在后续主题中进行论述。 要创建表达式，可以简单地输入公式字段，也可以通过单击计算器按钮使用表达式构建器。 在引用对多个字段的操作时，请记住使用 `@FIELD` 函数。

选择多个字段

对于对多个输入字段执行操作的所有节点（如导出节点（多个模式）、“汇总”节点、排序节点、多重散点图节点和时间散点图节点），您可以使用“选择字段”对话框轻松选择多个字段。

排序依据。 可以通过选择下列其中一个选项可用于查看的字段进行排序：

- **自然**。 数据流向下遍历数据时，当前节点接收字段的顺序即为字段的查看顺序。
- **名称**。 采用字母顺序对字段进行排序以便于查看。

- **类型**。 查看字段时按其测量级别排序。 此选项在选择具有特定测量级别的字段时非常有用。

一次从列表中选择一个字段，或采用按住 Shift 并单击和按住 Ctrl 并单击的方法选择多个字段。 此外，也可以使用列表下面的按钮根据测量级别选择多组字段，或选择或取消选择表中所有字段。

设置导出公式选项

“派生公式”节点根据 CLEM 表达式的结果为数据集中的每条记录创建一个新字段。 此表达式不能是条件表达式。 要派生基于条件表达式的值，请使用标志或条件类型的“派生”节点。

公式 使用 CLEM 语言指定用于为新字段派生值的公式。

注: 由于 SPSS Modeler 无法知道要用于派生列表字段的子测量级别，因此对于“集合”和“地理空间”测量级别，您可以单击**指定...** 以打开**值**对话框并设置必需的子测量级别。 有关更多信息，请参阅 第 117 页的『设置派生列表值』。

对于地理空间字段，唯一的相关存储类型为**实数**和**整数**（缺省设置为**实数**）。

设置派生列表值

当您从“派生”节点的“公式”**字段类型**下拉列表中选择 **指定...**时，将显示“值”对话框。 在此对话框中，可以设置用于“公式”**字段类型**测量级别“集合”或“地理空间”的子测量级别值。

测量 请选择**集合**或**地理空间**。 如果您选择了任何其他测量级别，那么对话框将显示一条消息，指出没有可编辑的值。

收集

唯一可以为“集合”**测量**级别设置的项是**列表测量**。 缺省情况下，此测量设置为“无类型”，但您可以选择另一个值，以设置列表中的元素的测量级别。 您可以选择下列其中一个选项：

- 无类型
- 分类
- 连续
- 名义
- 有序
- 标志

地理空间

对于“地理空间”**测量**级别，您可以选择下列选项，以设置该列表中的元素的测量级别：

类型 请选择地理空间字段的测量子级别。 可用的子级别由列表字段的深度确定；缺省值如下所示：

- 点（深度为零）
- 线串（深度为 1）
- 多边形（深度为 1）
- 多点（深度为 1）
- 多线串（深度为 2）
- 多多边形（深度为 2）

有关子级别的更多信息，请参阅《SPSS Modeler 的“源”节点、“过程”节点和“输出”节点》指南的『“类型”节点』部分中的“地理空间测量子级别”主题。

有关列表深度的更多信息，请参阅《SPSS Modeler 的“源”节点、“过程”节点和“输出”节点》指南的『“源”节点』部分中的“列表存储以及相关联的测量级别”主题。

坐标系 仅当您将测量级别由非地理空间级别更改为地理空间级别时，此选项才可用。 要对您的地理空间数据应用坐标系，请选中此复选框。 缺省情况下，将显示**工具** > **流属性** > **选项** > **地理空间**窗格中设置的坐标系。 要使用另一个坐标系，请单击**更改**按钮以显示“**选择坐标系**”对话框，并选择与数据匹配的坐标系。

有关坐标系的更多信息，请参阅《SPSS Modeler 用户指南》的『使用流』部分中的『设置流的地理空间选项』主题。

派生列表或地理空间字段

在某些情况下，应该记录为列表项的数据以错误的属性导入到 SPSS Modeler 中。 例如，作为单独的地理空间字段（例如 x 坐标和 y 坐标或者经度和纬度）而导入，或者作为 `.csv` 文件中的各个行而导入。 在这种情况下，必须将各个字段组合成单个列表字段；完成此任务的一种方法是使用“派生”节点。

注: 组合地理空间数据时，您必须知道哪个是 x（或经度）字段，哪个是 y（或纬度）字段。 您必须对该数据进行组合，以使生成的列表字段的元素顺序为 [x, y] 或 [经度, 纬度]，这是地理空间坐标的标准格式。

下列步骤显示派生列表字段的简单示例。

1. 在流中，将“派生”节点连接到“源”节点。
2. 在“派生”节点的“设置”选项卡上，从**派生为**列表中选择**公式**。
3. 在**字段类型**中，选择**集合**（对于非地理空间列表）或**地理空间**。 缺省情况下，SPSS Modeler 使用“最佳猜测”方法来设置正确的列表详细信息；您可以选择**指定 ...** 以打开**值**对话框。 此对话框可以用于集合以输入有关列表中数据的更多信息，对于地理空间，可以用于设置数据类型并指定数据的坐标系。

   注: 对于地理空间，您指定的坐标系必须与数据的坐标系完全匹配。 如果不是完全匹配，地理空间功能可能会产生不正确的结果。
4. 在**公式**窗格中，输入用于将数据组合成正确列表格式的公式。 并且，还可以单击计算器按钮以打开表达式构建器。

   用于派生列表的公式的一个简单示例是 [x, y]，其中 x 和 y 是数据源中的单独字段。 创建的新派生字段是一个列表，其中每个记录的值都是该记录的并置 x 和 y 值。

   注: 以此方式组合成列表的字段必须具有同一种存储类型。

有关列表和列表深度的更多信息，请参阅第 8 页的『列表存储以及相关联的测量级别』。

设置导出标志选项

Derive Flag 节点用于指明特定条件，如高血压或客户帐户停用。 对于每条记录都会创建一个标志字段，当条件为真时，会在字段中添加代表真的标志值。

True 值。 指定针对满足以下指定条件的记录要在标志字段中包括的值。 缺省值为 T。

False 值。 对于那些不满足以下指定条件的记录，指定其标志字段中的值。 缺省值为 F。

True 值条件： 指定某个 CLEM 条件，用于评估每条记录的某些值，并为记录赋予真值或假值（定义如上）。 请注意，对于非假数字值，会将真值赋予记录。

注意：要返回空字符串，您应该输入一对引号，并且中间不包含任何内容，如 `""`。 例如，空字符串通常可用作假值，以使真值在表中更为明显。 类似地，如果希望某个字符串值在其他情况下被视为数值，应使用引号

示例

在 IBM SPSS Modeler 12.0 之前的版本中，可用逗号来分隔值将多个响应导入一个字段中。 例如：

```
museum_of_design,institute_of_textiles_and_fashion
museum_of_design
archeological_museum
$null$
national_art_gallery,national_museum_of_science,other
```

要准备此数据进行分析，您可以使用 `hassubstring` 函数为每个响应生成一个单独的标志字段，其表达式如下：

```
hassubstring(museums,"museum_of_design")
```

设置派生名义选项

Derive Nominal 节点用于执行一组 CLEM 条件，以确定每条记录满足的条件。 当每条记录满足某个条件时，会将一个值（指示满足哪组条件）添加到新的导出字段。

缺省值。 指定不满足任何条件时要使用的值。

将字段设置为。 指定满足某个特定条件时要在新字段中输入的值。 列表中的每个值都有一个关联条件，该条件由用户在相邻列中指定。

如果此条件为 true。 为集合字段中要列出的每个成员指定条件。 使用表达式构建器在可用的函数和字段中进行选择。 可以使用箭头和删除按钮对条件进行重新排序或删除。

条件的工作原理是对数据集中特定字段的值进行检验。 检验每个条件时，都会为新字段分配上述指定值，以指示满足哪个条件（如果有）。 如果不满足任何条件，则会使用缺省值。

设置导出状态选项

Derive State 节点与 Derive Flag 节点相当类似。 标志节点根据当前记录对单个 条件的满足情况设置值，而派生状态节点可以根据字段对两个独立 条件的满足方式更改该字段的值。 这意味着满足每个条件时，该值都会发生更改（打开或关闭）。

初始状态。 选择初始时要为新字段的每条记录指定 **On** 还是 **Off**。 请注意，此值可能在满足每个条件时发生更改。

“On”值。 指定满足 On 条件时新字段的值。

切换为“On”的条件。 指定会在条件为真时将状态更改为“开”的 CLEM 条件。 单击计算器按钮可打开表达式构建器。

“Off”值。 指定满足 Off 条件时新字段的值。

切换为“Off”的条件。 指定会在条件为假时将状态更改为“关”的 CLEM 条件。 单击计算器按钮可打开表达式构建器。

注意：要指定空字符串，您应该输入一对引号，并且中间不包含任何内容，如 ""。 类似地，如果希望某个字符串值在其他情况下被视为数值，那么应使用引号。

设置导出计数选项

Derive Count 节点用于对数据集中数字字段的值应用一系列条件。 满足每个条件时，导出的计数字段的值会根据集合增量的大小而相应增加。 这种类型的 Derive 节点对于时间序列数据十分有用。

初始值。 设置执行新字段时使用的值。 初始值必须是数字常数。 使用箭头按钮可以增加或减少该值。

增量条件。 指定某个 CLEM 条件，满足该条件时将根据“增量为”中指定的数值更改派生值。 单击计算器按钮可打开表达式构建器。

增量。 设置用于增加计数的值。 可以使用数字常量或 CLEM 表达式的结果。

重置条件。 指定某个条件，满足该条件时会将派生值重置为初始值。 单击计算器按钮可打开表达式构建器。

设置导出条件选项

Derive Conditional 节点使用一系列 If-Then-Else 语句派生新字段的值。

If。 指定一个 CLEM 条件，在执行时会逐一为每条记录评估该条件。 如果条件为真（对于数值的情况为非假），那么会为新字段赋予下面通过 Then 表达式指定的值。 单击计算器按钮可打开表达式构建器。

Then。 指定上述 If 语句为真（或非假）时新字段的值或 CLEM 表达式。 单击计算器按钮可打开表达式构建器。

Else。 指定上述 If 语句为假时新字段的值或 CLEM 表达式。 单击计算器按钮可打开表达式构建器。

使用导出节点对值进行重新编码

Derive 节点还可用于对值进行重新编码，例如，通过将具有分类值的字符串字段转换为数值名义（集合）字段。

1. 对于“派生为”，选择字段的类型（名义、标志等）， 视情况而定。
2. 指定对值进行重新编码的条件。 例如，您可以将值设置为 1 if `Drug='drugA'`、2 if `Drug='drugB'` 等等。

填充节点

填充节点用于替换字段值和更改存储类型。 您可以选择根据指定的 CLEM 条件（例如 `@BLANK(FIELD)`）来替换值。 或者，也可以选择将所有空白值或空值替换为特定值。 “填充”节点通常与“类型”节点结合使用，用于替换缺失值。 例如，您可以通过指定诸如 `@GLOBAL_MEAN` 之类的表达式来使用字段的平均值填充空白。此表达式将为所有空白值填充通过设置全局量节点计算的均值。

填写字段。 使用字段选择器（文本字段右边的按钮）从数据集中选择要检查并替换其值的字段。 缺省行为是根据“条件”和“替换”将值替换为下面指定的表达式。 另外，也可以使用下面的“替换”选项选择替代的替换方法。

注: 选择要替换为用户定义值的多个字段时，字段类型相似（均为数字或均为符号）是很重要的。

替换。 选择此选项可使用下列其中一种方法来替换所选字段的值:

- **基于条件**。 此选项将激活“条件”字段和表达式构建器，使用它们可创建用作将值替换为指定值的条件的表达式。
- **始终**。 替换所选字段的所有值。 例如，可以通过此选项使用以下 CLEM 表达式将 income 的存储类型转换为字符串：`(to_string(income))`。
- **空白值**。 替换所选字段中的所有用户指定的空白值。 标准条件 `@BLANK(@FIELD)` 用于选择空白值。 注意：您可以使用源节点的“类型”选项卡或使用“类型”节点定义空白值。
- **空值**。 替换所选字段中的所有系统空值。 标准条件 `@NULL(@FIELD)` 用于选择空值。
- **空白值和空值**。 替换所选字段中的空白值和系统空值。 当您不确定是否已将空定义为缺失值时，此选项将十分有用。

条件。 此选项在选中**根据条件**选项后可用。 使用此文本框指定用于评估所选字段的 CLEM 表达式。 单击计算器按钮可打开表达式构建器。

替换为。 指定某个 CLEM 表达式来为所选字段赋予新值。 此外，也可以通过在文本框中键入 `undef` 将值替换为空值。 单击计算器按钮可打开表达式构建器。

注: 如果所选字段为字符串，那么应将其替换为字符串值。 使用缺省值 0 或其他数字值作为字符串字段的替换值将产生错误。

请注意，使用以下各项可能会更改行顺序:

- 通过 SQL 回送在数据库中执行
- 通过远程 IBM SPSS Analytic Server 执行
- 使用在嵌入式 IBM SPSS Analytic Server 中运行的函数
- 派生列表（例如，请参阅第 118 页的『派生列表或地理空间字段』）
- 调用任何空间函数

使用填充节点进行存储类型转换

使用“填充”节点的“替换”条件，可以轻松转换单个或多个字段的字段存储类型。 例如，借助以下 CLEM 表达式，使用转换函数 `to_integer` 可以将 *income* 从字符串转换为整数：`to_integer(income)`。

可以使用表达式构建器查看可用的转换函数并自动创建 CLEM 表达式。 在“函数”下拉列表中，选择**转换**可查看存储类型转换函数的列表。 可用的转换函数如下:

- `to_integer(ITEM)`

- to_real(ITEM)
- to_number(ITEM)
- to_string(ITEM)
- to_time(ITEM)
- to_timestamp(ITEM)
- to_date(ITEM)
- to_datetime(ITEM)

转换日期和时间值。 请注意，转换函数（以及需要特定类型输入（如日期或时间值）的任何其他函数）取决于“流选项”对话框中指定的当前格式。 例如，如果要将值为 *Jan 2003* 、 *Feb 2003* 等的字符串字段转换为日期存储类型，请选择 **MON YYYY** 作为流的缺省日期格式。

“派生”节点中也有可用的转换函数，用于派生计算过程中的临时转换。 另外，还可以使用“派生”节点来执行其他操作，例如使用分类值对字符串字段进行重新编码。 有关更多信息，请参阅主题 第 120 页的『使用导出节点对值进行重新编码』。

重新分类节点(C)

“重新分类”节点可实现一组分类值到另一组分类值的转换。 对于折叠类别或者进行数据重新分组以执行分析而言，重新分类非常有用。 例如，可以将 *Product* 的值重新分类为三组，如 *Kitchenware* 、 *Bath and Linens* 和 *Appliances* 。 通常情况下，此操作按分组值直接通过“分布”节点执行，并且生成一个“重新分类”节点。 有关更多信息，请参阅主题 第 180 页的『使用分布节点』。

可以对一个或多个符号字段执行重新分类。 您也可以选择为现有字段替换新值或生成新字段。

何时使用重新分类节点

在使用重新分类节点之前，请考虑是否有更适用于当前任务的其他字段操作节点：

- 要采用自动方法将数值范围变换为集合（如等级或百分位数），应使用“分级”节点。 有关更多信息，请参阅主题 第 124 页的『分级节点(B)』。
- 要将数值范围手动分类为集合，应使用“派生”节点。 例如，如果要将工资值压缩至特定的工资范围类别，应使用“派生”节点手动定义每个类别。
- 要根据分类字段（如 *Mortgage_type* ）的值创建一个或多个标志字段，应使用设为标志节点。
- 要将分类字段转换为数值范围，可以使用“派生”节点。 例如，可以将 *No* 和 *Yes* 值分别转换为 0 和 1。 有关更多信息，请参阅主题 第 120 页的『使用导出节点对值进行重新编码』。

为重新分类节点设置选项

重新分类节点的使用分为以下三个步骤：

1. 首先，选择要对多个字段还是单个字段进行重新分类。
2. 下面，选择是在现有字段内重新编码还是创建新字段。
3. 然后，根据需要使用重新分类节点对话框中的动态选项映射集合。

   方式。 选择**单个**可对一个字段进行重新分类。 选择 **多个** 将激活若干选项，它们可实现同时转换多个字段。

   重新分类到。 选择**新字段**将保留原始名义字段，并派生包含重新分类的值的新字段。 选择 **现有字段** 将使用新的分类覆盖原始字段中的值。 此选项实质上是一种“填充”操作。

   指定模式和替换选项后，必须选择转换字段并使用对话框下半部分的动态选项指定新的分类值。 这些选项会依据前面所选模式的不同而变化。

   对字段重新分类。 使用右边的字段选择器按钮选择一个（“单个”模式）或多个（“多个”模式）分类字段。

   新字段名称。 为包含重新编码值的新名义字段指定名称。 如果前面选择了**新字段**，此选项仅在“多个”模式下可用。 如果选择了 **现有字段** ，则会保留原始字段名。 采用“多个”模式时，此选项将被其它控件替

换，以指定向每个新字段添加的扩展名。 有关更多信息，请参阅主题 第 122 页的『对多个字段进行重新分类』。

对值重新分类。 使用此表，可以实现从旧集合值到此处指定的集合值的明确映射。

- **原始值**。 此列列出选择字段的现有值。
- **新值**。 使用此列可输入新的类别值或从下拉列表中选择类别值。 使用分布图中的值自动生成“重新分类”节点时，这些值将包括在该下拉列表中。 这样，您可以将现有值快速映射至已知值集合。 例如，医疗保健组织有时会根据网络或语言环境对诊断进行不同分组。 经过合并或采集，所有各方都需要采用一致方式对新的或现有数据进行重新分类。 可以将值的主列表读入 IBM SPSS Modeler，对 *Diagnosis* 字段运行条形图，然后直接从该图生成字段的重新分类（值）节点，而无需手动键入冗长列表中的每个目标值。 此过程将使所有目标 Diagnosis 值显示在“新值”下拉列表中。

4. 单击 **获取** 读取前面选择的一个或多个字段的原始值。
5. 单击 **复制** 针对尚未映射的字段将原始值粘贴至 新值 列。 未映射的原始值将添加到下拉列表中。
6. 单击 **清除新值** 将擦除 新值 列中的所有指定值。 注意：此操作不会将值从下拉列表中擦除。
7. 单击 **自动** 可自动生成代表每个原始值的连续整数。 只能生成整数值，不能生成实数值（如 1.5、2.5 等）。

例如，可以自动生成代表产品名的连续产品标识，或代表大学课程的课程编号。 此功能对应于 IBM SPSS Statistics 中集合的自动重新编码转换。

对于未指定的值，使用。 此选项用于在新字段中填充未指定的值。 可以选择保留原始值（选择 **原始值** ），也可以指定缺省值。

对多个字段进行重新分类

要一次映射多个字段的类别值，请将模式设置为 **多个** 。 这时“重新分类”对话框中将显示新的设置，下面介绍这些设置。

重新分类字段。 使用右边的“字段选择器”按钮选择要转换的字段。 使用字段选择器，可以同时选中所有字段或属于相似类型的字段，如名义或标志。

字段名称扩展。 同时对多个字段进行重新编码时，指定添加到所有新字段的扩展名比指定各个字段名称更为高效。 指定扩展名（例如，`_recode`），并选择是否将此扩展名附加或前置到原始字段名称。

重新分类字段的存储类型和测量级别

“重新分类”节点总是通过重新编码操作创建名义字段。 在某些情况下，这种方法可能会在使用**现有字段**重新分类模式时更改字段的测量级别。

新字段的存储类型（数据的存储方式，而不是使用方式）基于“设置”选项卡的以下选项进行计算：

- 如果将未指定的值设置为使用缺省值，存储类型将按以下方式决定：检查新值和缺省值，并确定适当的存储类型。 例如，如果所有值均可解析为整数，字段将获得整数存储类型。
- 如果将未指定的值设置为使用原始值，存储类型将以原始字段的存储类型为基准。 如果所有值均可解析为原始字段的存储类型，那么将保留该存储类型；否则，存储类型将通过查找同时包含旧值和新值的最适合的存储类型来确定。 例如，使用再分类 4 => 0, 5 => 0 重新分类整数集 { 1, 2, 3, 4, 5 } 将生成新整数集 { 1, 2, 3, 0 }，而使用再分类 4 => “Over 3”, 5 => “Over 3” 将生成字符串集 { “1”, “2”, “3”, “Over 3” }。

注意：如果原始类型是非实例化类型，那么新类型也将是非实例化类型。

匿名化节点

通过使用匿名化节点，您可以在处理要包含在节点的下游模型中的数据时对字段名称和/或字段值进行掩饰。 这样，可以随意分发所生成的模型（例如，分发至技术支持部门），而未授权用户无法查看机密数据（例如，员工记录或患者的医疗记录）。

您可能需要对其他节点进行更改，具体取决于匿名化节点在流中的位置。 例如，如果通过使用“选择”节点在上游中插入一个匿名化节点，那么该“选择”节点中的选择标准作用于现已匿名化的值时，这些标准需要进行更改。

用于匿名化的方法取决于多种因素。 对于字段名称以及除“连续”测量级别外的所有字段值，数据将替换为以下形式的字符串：

```
prefix_Sn
```

其中 *prefix_* 是用户指定的字符串或缺省字符串 anon_，*n* 是从 0 开始并在遇到每个唯一值时递增（例如，anon_S0、anon_S1 等）的整数值。

类型为“连续”的字段值必须进行变换，因为数值范围处理的是整数或实数值，而不是字符串。 因此，只能通过将范围变换为不同范围对字段值进行匿名化，从而掩饰原始数据。 范围内的值 *x* 的变换按以下方法执行：

```
A*(x + B)
```

其中：

A 是比例因子，必须大于 0。

B 是要为值增加的转换偏移量。

示例

对于年龄字段，如果比例因子 *A* 设置为 7 而转换偏移量 *B* 设置为 3，那么年龄的值将转换为：

```
7*(AGE + 3)
```

匿名化节点的设置选项

您可在此处选择要在较下游位置对其值进行掩饰的字段。

请注意，必须先在匿名化节点的上游对数据字段进行实例化，然后才能执行匿名化操作。 您可以通过在类型节点或在源节点的“类型”选项卡上单击**读取值**按钮对数据进行实例化。

字段。 列出当前数据集中的字段。 如果有任何字段名称已匿名化，那么会在此处显示匿名化名称。

测量。 字段的测量级别。

匿名化值。 选择一个或多个字段，并单击此列，然后选择**是**以使用缺省前缀 **anon_** 对字段值进行匿名化；选择**指定**以显示一个对话框，您可在其中输入自己的前缀，或者对于类型为连续的字段值，也可以指定字段值的转换将采用随机值还是用户指定值。 请注意，不能在同一操作中对连续和非连续字段类型进行指定，必须分别针对每种字段类型执行此操作。

查看当前字段。 选择此选项可查看当前连接到匿名化节点的数据集的字段。 缺省情况下，此选项处于选中状态。

查看未使用的字段设置。 选择此选项可查看曾连接到该节点但已断开连接的数据集的字段。 将节点从一个流复制到另一个流时，或保存并重新加载节点时，此选项十分有用。

指定如何对字段值进行匿名化

通过使用“替换值”对话框，您可以选择对匿名化字段值使用缺省前缀还是自定义前缀。 在此对话框中单击**确定**可针对所选字段将“设置”选项卡中的“对值进行匿名化”设置更改为**是**。

字段值前缀。 匿名化字段值的缺省前缀为 **anon_**；如果要使用其他前缀，请选择**自定义**并输入您自己的前缀。

“变换值”对话框仅针对类型为“连续”的字段显示，可用于指定字段值的变换使用随机值还是用户指定值。

随机。 选择此选项将对转换采用随机值。 **设置随机种子**在缺省情况下处于选中状态；请在**种子**中指定一个值，或使用缺省值。

固定。 选择此选项可为变换指定您自己的值。

- **比例**。 字段值在变换中的乘数。 最小值为 1；最大值通常为 10，但有时需要减小该值以避免溢出。

- **变换值**。 在变换中将为字段值增加的数字。 最小值为 0；最大值通常为 1000，但有时需要减小该值以避免溢出。

对字段值进行匿名化

已在“设置”选项卡上选择进行匿名化的字段，其值将在以下情况下进行匿名化：

- 运行包含匿名化节点的流时
- 对值进行预览时

要对值进行预览，请单击“对值进行匿名化”选项卡中的**对值进行匿名化**按钮。 然后，从下拉列表中选择一个字段名称。

如果测量级别为“连续”，那么将显示以下内容：

- 原始范围的最小值和最大值
- 用于变换值的方程式

如果测量级别不是“连续”，那么屏幕将显示该字段的原始值和匿名化值。

如果屏幕显示黄色背景，那么这表示所选字段的设置自上次对值进行匿名化以来已发生更改，或者表示已对匿名化节点的上游数据进行更改，导致匿名化值发生错误。 此时将显示当前值集合；再次单击**对值进行匿名化**按钮可根据当前设置生成一组新值。

匿名化值。 为所选字段创建匿名化值并在表中显示这些值。 如果针对类型为“连续”的字段使用随机播种，那么每次单击此按钮都会创建一组不同的值。

清除值。 从表中清除原始值和匿名化值。

分级节点(B)

使用 Binning 节点，可以根据一个或多个现有连续（数值范围）字段的值自动创建新的名义字段。 例如，可以将连续收入字段转换为包含若干等宽收入组的新的分类字段，或转换为与均值之间的偏差。 或者，也可以选择一个“主管”分类字段，以保持两个字段之间原始关联的强度。

分级的实用性源于以下几个原因：

- **算法要求**。 某些特定算法（如朴素贝叶斯、Logistic 回归）要求分类输入。
- **性能**。 如果减少输入字段的不同值数量，算法（如多项 Logistic）的性能可能会提高。 例如，对每个分级使用中位数或均值，而不使用原始值。
- **数据隐私**。 敏感类个人信息（如工资）可采用范围的报告形式，而不使用实际工资数字，以保护个人隐私。

提供了一些分级方法。 一旦创建新字段分级后，即可根据割点创建“衍生”节点。

何时使用分级节点

在使用分级节点之前，请考虑是否有更适用于当前任务的其他技术：

- 要为类别手动指定特定（如特定的预定义工资范围），请使用 Derive 节点。 有关更多信息，请参阅主题 第 115 页的『“派生”节点』。
- 要为现有集合创建新类别，请使用重新分类节点。 有关更多信息，请参阅主题 第 121 页的『重新分类节点(C)』。

缺失值处理

分级节点处理缺失值的方法如下：

- **用户指定的空白值**。 转换过程中将包括指定为空白值的缺失值。 例如，若使用 Type 节点指定 –99 表示空白值，那么会在分级过程中包括此值。 要在分级过程中忽略空白值，应使用 Filler 节点将空白值替换为系统空值。
- **系统缺失值 ($null$)**。 在分级转换期间将忽略空值，并在转换后保持空值。

“设置”选项卡提供了有关适用技术的选项。 “视图”选项卡将显示针对先前通过节点的数据建立的割点。

为分箱节点设置选项

使用分级节点，可以采用以下技术自动生成分级（类别）：

- 固定宽度分级
- 分位数（相等计数或总和）
- 均值和标准差
- 等级
- 相对于分类“主管”字段的最优化

对话框的下部分会根据所选的分级方法动态变化。

分级字段。 此处将显示待转换的连续（数值范围）字段。 使用分级节点，可以同时对多个字段进行分级。 使用右侧的按钮可添加或删除字段。

分箱方法。 选择用于确定新字段分级（类别）的分割点的方法。 后续主题描述每个观测值中提供的选项。

分级阈值 指定如何计算分级阈值。

- **始终重新计算**。 运行节点时，将始终计算分割点和分级分配。
- **从“分级值”选项卡中读取（如果可用）**。 仅在必要时（例如，添加新数据后）计算分割点和分级分配。

以下主题将介绍可用分级方法的选项。

固定宽度分级

选择 **固定宽度** 作为分级方法时，对话框中会显示一组新的选项。

名称扩展。 指定要用于所生成字段的扩展。*_BIN* 是缺省扩展。 您还可以指定将扩展部分添加到字段名的开头（ **前缀** ）还是末尾（ **后缀** ）。 例如，可以生成名为 *income_BIN* 的新字段。

分级宽度。 指定用于计算分级“宽度”的值（整数或实数）。 例如，可以使用缺省值 10 对字段 *Age* 进行分级。 由于 *Age* 的范围为 18–65，因此生成的分级如下表所示。

表 *25:* 范围为 *18 - 65* 的年龄分级

分级 1	分级 2	分级 3	分级 4	分级 5	分级 6
>=13 到 <23	>=23 到 <33	>=33 到 <43	>=43 到 <53	>=53 到 <63	>=63 到 <73

分级间隔起点的计算方法为：扫描到的最低值减去分级宽度的一半（指定值）。 例如，在上面显示的分级中，使用 13 作为间隔的起点，依据的计算方法如下：18 *[*最低数据值*]* - 5 *[0.5 × (*分级宽度 *10)]* = 13。

分级数量。 使用此选项可指定用于确定新字段的固定宽度分级（类别）数的整数。

在流中执行分级节点后，即可通过单击分级节点对话框中的 **预览** 选项卡来查看已生成的分级阈值。 有关更多信息，请参阅主题 第 128 页的『预览生成的分级』。

分位数（相等计数或总和）

分位数分级方法用于创建名义字段，这些字段可用于将扫描到的记录分割为百分位数（或四分位数、十分位数等）组，使每个组包含相同数量的记录，或使每个组中值的总和相等。 记录根据指定的分级字段值按升序排列，因此所选分级变量的值最低的记录将获得等级 1，下一组记录等级为 2，依此类推。 每个分级的阈值将根据所用的数据和分位方法自动生成。

分位数名称扩展。 指定用于使用标准 p 分位数生成的字段的扩展名。 缺省扩展名为 *_TILE* 加上 *N* ，其中 *N* 是分位数。 您还可以指定将扩展部分添加到字段名的开头（ **前缀** ）还是末尾（ **后缀** ）。 例如，可以生成名为 *income_BIN4* 的新字段。

定制分位数扩展名。 指定用于定制分位数范围的扩展名。 缺省值为 *_TILEN* 。 请注意，此处的 *N* 将不会被定制数字替换。

可用的 p 分位数如下：

- **四分位数**。 生成 4 个分级，每个包含 25% 的观测值。

- **五分位数**。 生成 5 个分级，每个包含 20% 的观测值。
- **十分位数**。 生成 10 个分级，每个包含 10% 的观测值。
- **二十分位数**。 生成 20 个分级，每个包含 5% 的观测值。
- **百分位数**。 生成 100 个分级，每个包含 1% 的观测值。
- **定制 N**。选择此选项可指定分级数。 例如，值为 3 将产生 3 个划分类别（2 个割点），每个包含 33.3% 的观测值。

请注意，如果数据中的离散值少于指定的分位数，那么不会使用任何分位数。 在这种情况下，新的分布很可能反映数据的原始分布。

分位方法。 指定用于为分级分配记录的方法。

- **记录计数**。 尽量为每个分级分配相等数目的记录。
- **值的总和**。 为分级分配记录时，尽量使每个分级中值的总和相等。 例如，以销售业绩为目标时，此方法可用于根据每条记录的值为十分位数组分配预期业绩，最高分级获得价值最高的预期业绩。 例如，某制药公司可根据所开处方的数量将医师分入十分位数组。 虽然每个十分位数都将包含大致相同的脚本数，但贡献这些脚本的个人数量将不尽相同，编写最多脚本的个人集中在十分位数 10 中。 请注意，此方法假定所有值都大于零，如果不是这样，可能会产生意外的结果。

同数。 当分割点两侧的值相同时，将产生结条件。 例如，如果是分配十分位数，且超过 10% 的记录的分级字段具有相同值，那么除非对阈值进行向上或向下的强制转换，否则无法将这些记录全部分配至同一分级。 可以将结上移至下一个分级，也可以保留在当前分级中，但必须将其解决，使具有相同值的所有记录位于同一分级内，即使这样会导致某些分级的记录数超过预期值也是如此。 后续分级的阈值可能也会因此发生调整，导致对相同数字集合进行不同的值分配，具体取决于用于解决结的方法。

- **添加到下一个**。 选择此选项可将结值上移至下一个分级。
- **保留在当前**。 将值保留在当前（较低）分级中。 此方法可能会减少创建的分级总数。
- **随机分配**。 选择此选项可将同数值随机分配至一个分级。 这将试图使每个分级中的记录数量相等。

示例：按记录计数分位

下表说明了按记录计数进行分位时如何将简单字段值分为四分位数。 请注意，结果将随选择的结选项而变化。

表 *26:* 按记录计数分位的示例

值	添加到下一个	保留在当前分级中
10	1	1
13	2	1
15	3	2
15	3	2
20	4	3

每个分级的项数的计算方法如下：

```
total number of value / number of tiles
```

在上面的简单示例中，每个分级的所需项数为 1.25（5 个值 / 4 个四分位数）。 值 13（值编号为 2）跨越 1.25 的所需计数阈值，因此将根据所选的结选项进行不同处理。 在**添加到下一个**方式下，它将添加到分级 2 中。 在**保留在当前**方式下，它将保留在分级 1 中，并将分级 4 的值范围推至现有数据值范围之外。 结果是，仅创建三个分级，每个分级的阈值将进行相应调整，如下表中所示。

表 *27:* 分级示例结果

分级	下限	上限
1	>=10	<15

表 *27:* 分级示例结果 *(*继续*)*

分级	下限	上限
2	>=15	<20
3	>=20	<=20

注意：启用并行处理可提高按分位数分级的速度。

观测值排秩

选择 **排序** 作为分级方法时，对话框中会显示一组新的选项。

排秩可创建包含数字字段的排秩值、分数排秩值和百分位数值的新字段，具体取决于下面指定的选项。

等级顺序。 选择**升序**（将最低值标记为 1）或**降序**（将最高值标记为 1）。

等级。 选择此选项将按上面指定的升序或降序对观测值进行排序。 新字段中的值的范围将是 1–*N*，其中 *N* 是原始字段中离散值的数目。 结值将获得其排序值的平均值。

分数排序。 选择此选项将对观测值进行排序，其中新字段的值等于排序值除以非缺失观测值的权重和。 分数排序值介于 0–1 之间。

分数排序百分比。 每个排序值除以具有有效值的记录数再乘以 100。 分数排序百分比值介于 1–100 之间。

分机号。 对于所有排序选项，还可以创建定制扩展名，并指定将其添加到字段名的开头（**前缀**）还是末尾（**后缀**）。 例如，可以生成名为 *income_P_RANK* 的新字段。

均数/标准差

选择**均数/标准差**作为分级方法时，对话框中会显示一组新的选项。

此方法可根据指定字段分布的均数和标准差的值生成具有划分类别的一个或多个新字段。 选择下面要使用的偏差数。

名称扩展。 指定要用于所生成字段的扩展名。 *_SDBIN* 是缺省扩展名。 您还可以指定将扩展部分添加到字段名的开头（ **前缀** ）还是末尾（ **后缀** ）。 例如，可以生成名为 *income_SDBIN* 的新字段。

- **+/– 1 标准偏差**。 选择此选项将生成三个分级。
- **+/– 2 标准偏差**。 选择此选项将生成五个分级。
- **+/– 3 标准偏差**。 选择此选项将生成七个分级。

例如，选择 +/–1 标准差将产生三个分级，计算方法如下表所示。

表 *28:* 标准差分级示例

分级 1	分级 2	分级 3
x < (Mean – Std. Dev)	(Mean – Std. Dev) <= x <= (Mean + Std. Dev)	x > (Mean + Std. Dev)

在正态分布中，68% 的观测值落入与均数相距不到一个标准差的范围内，95% 落入两个标准差的范围内，99% 落入三个标准差的范围内。 但请注意，根据标准差创建带状类别可能会使某些分级定义超出实际数据范围，甚至超出可能的数据值范围（例如，负值工资范围）。

最优分级

如果要分箱的字段与另一个分类字段强关联，则可选择分类字段作为“主管”字段以便以类似于保留两个字段间的原始关联强度的方式创建分箱。

例如，假定已采用聚类分析根据家庭贷款的拖欠率对状态进行分组，那么最高拖欠率将位于第一个聚类中。 在这种情况下，可以选择 过期百分比 和 取消赎取权百分比 作为分级字段和模型生成的作为主管字段的聚类成员资格字段。

名称扩展 指定要用于所生成字段的扩展名，以及是将其添加到字段名开头（**前缀**）还是末尾（**后缀**）。 例如，可以生成名为 pastdue_OPTIMAL 的新字段以及名为 inforeclosure_OPTIMAL 的另一个字段。

主管字段 这是用于构造分箱的分类字段。

预分级字段以提高大型数据集的性能 指示应在最优分级的流程化中使用预处理。 该方法会采用简单的非监督式分级方法将尺度值分组为大量分级，以均值表示每个分级中的值，并在继续监督式分级之前对观测值权重进行相应调整。 在实际应用中，此方法会牺牲一定的精度以换取速度，建议用于大型数据集。 使用此选项时，也可以指定任意变量预处理后的最大分级数。

将观测值计数相对较小的分级与较大的相邻分级进行合并。 如果启用，则指示当该分级大小（观测值的个数）与相邻分级大小的比值小于指定的阈值时，将合并分级；请注意阈值越大合并的分级越多。

剪切点设置

使用“分割点设置”对话框，可以指定最优分级算法的高级选项。 这些选项将指示算法如何使用目标字段计算分级。

分箱端点。 您可以指定应该包含（下端点 <= *x*）还是不包含（下端点 < *x*）下端点或上端点。

第一个和最后一个分箱。 对于第一个和最后一个分级，可以指定该分级应无限制（向正无穷或负无穷延伸）还是按最低或最高数据点进行限制。

预览生成的分级

使用 Binning 节点中的“分级值”选项卡，可以查看已生成分级的阈值。 使用“生成”菜单还可以生成 Derive 节点，该节点可用于将一个数据集中的阈值应用于另一个数据集。

分级字段。 使用下拉列表选择要查看的字段。 为明确起见，显示的字段名采用原始字段名。

磁贴。 使用下拉列表选择用于查看的分位数，如 10 或 100。 仅当分级采用分位数方法（相等计数或总和）生成时，此选项才可用。

分级阈值 此处显示每个已生成分级的阈值，以及每个分级内的记录数。 仅对于最优分级方法，每个分级中的记录数显示为总数的百分比。 请注意，采用排序分级方法时，阈值不适用。

读取值。 从数据集中读取分级值。 请注意，当新数据通过流时，阈值也将被覆盖。

生成派生节点

可以根据当前阈值使用“生成”菜单创建 Derive 节点。 将已建立的一组数据的分级阈值应用于另一组数据时，此操作十分有用。 此外，如果这些分割点已知，那么使用大型数据集时导出操作比分级操作更为高效（即更迅速）。

RFM 分析节点

通过近因、频率和货币 (RFM) 分析节点，您可以检查客户最近一次购买您产品或服务的时间（近因）、客户购买的频率 （频率）以及客户支付的所有交易金额（货币），确定可能成为最佳客户的数量。

RFM 分析的原理是曾经购买过产品或服务的客户更有可能会再次进行购买。 分类的客户数据会分为多个分级，其中分级标准可根据您的需要进行调整。 在每个分级中，会分配给客户一个评分；然后将这些评分组合在一起，从而得到 RFM 的总分值。 此评分表示为每个 RFM 参数创建的各分级中的客户成员资格。 这种已分级的数据可以充分满足您的需求，例如识别购买频率最高的高价值客户；另外，它还可以在流中进行传递以便进一步建模和分析。

不过需要注意的是，尽管分析 RFM 评分并对这些评分进行排序的功能非常有用，但使用时还必须注意一些特定的因素。 可以对排名最高的目标客户进行一些促销活动；但过度诱导这些客户可能会适得其反，导致他们在重复的交易过程中出现反感或不进行购买。 另外，我们还需要牢记：不应忽视评分低的客户，因为他们经过培养可能会成为更好的客户。 相反，根据市场反馈，仅具有高分值的客户不一定能带来好的预期销售业绩。 例如，近因中分级为 5 的客户（即最近购买过产品或服务的客户）对有些销售人员（如销售汽车或电视等昂贵且使用期较长的产品的人员）来说可能并不是真正的最佳目标客户。

注意：根据数据存储的方式，可能需要在使用“RFM 分析”节点之前先使用“RFM 汇总”节点将数据转换为可用格式。 例如，输入数据必须是客户格式，一行一个客户；如果客户的数据是事务处理数据，请使用“RFM 汇总”节点在上游派生近因、频率和货币字段。 有关更多信息，请参阅主题 第 63 页的『RFM “汇总”节点』。

将 IBM SPSS Modeler 中的“RFM 汇总”节点和 RFM 分析节点设置为使用独立分级；即，它们分别按近因、频率、货币值对数据进行排序和分级，而无需考虑它们的值或其他两种标准。

RFM 分析节点设置

近因。 使用字段选择器（文本框右侧的按钮）选择近因字段。 它有可能是日期、时间戳记或简单的数值。 请注意，如果日期或时间戳表示的是最近交易的日期，则将最高值视为最近；如果指定一个数值，它会表示自最近交易以来过去的时间，并将最低值视为最近。

注意：如果流中“RFM 汇总”节点的位置在“RFM 分析”节点之前，那么应将“RFM 汇总”节点生成的近因、频率和货币字段选作“RFM 分析”节点的输入。

频率。 使用字段选择器选择要使用的频率字段。

货币。 使用字段选择器选择要使用的货币字段。

分级数。 为三种输出类型分别选择要创建的分级数。 缺省值是 5。

注意：分级的最小值为 2，最大值为 9。

权重。 缺省情况下，计算分值时会将近因数据的重要性视为最高，其次是频率，最后是货币。 如果需要，可以修改影响上述一个或多个字段的权重，来更改重要性级别。

RFM 评分的计算方法如下：（近因分值 x 近因权重）+（频率分值 x 频率权重）+（货币分值 x 货币权重）。

同数。 指定如何分级相同的评分。 选项为：

- **添加到下一个。** 选择此选项可将结值上移至下一个分级。
- **保留在当前。** 将值保留在当前（较低）分级中。 此方法可能会减少创建的分级总数。 （这是缺省值。）

分级阈值 指定在执行节点时是始终重新计算 RFM 分值和分级分配，还是仅在需要时进行计算（如在添加了新数据时）。 如果选择**如果可用，从“分级值”选项卡读取**，则可以在“分级值”选项卡上编辑不同分级的上、下割点。

执行时，RFM 分箱节点可分级原始近因、频率和货币字段，并将下列新字段添加到数据集：

- 近因评分。 近因排序（分级值）
- 频率评分。 频率排序（分级值）
- 消费金额评分。 货币排序（分级值）
- RFM 评分。 近因、频率和货币评分的加权和。

将离群值添加到结束分级。 如果选中此复选框，那么可将位于较低分级下面的记录添加到较低的分级中，同时将最高分级上面的记录添加到最高分级中；否则，会将空值分配给这些记录。 只有选择**如果可用，从“分级值”选项卡读取**时，才可使用此复选框。

RFM 分析节点分级

通过“分级值”选项卡，您可以查看并在某些情况下修改已生成分级的阈值。

注意：如果在“设置”选项卡中选中了**如果可用，从“分级值”选项卡读取**，那么只能在此选项卡上修改值。

分级字段。 使用下拉列表选择要分级的字段。 可用值是“设置”选项卡上选定的值。

分级值表。 此处显示每个已生成分级的阈值。 如果在“设置”选项卡上选中了**如果可用，从“分级值”选项卡读取**，则可以通过双击相应单元格修改每个分级的上、下割点。

读取值。 从数据集中读取已分级值并填写分级值表。 请注意，如果在“设置”选项卡上选中了**始终再计算**，则当新数据通过流时，分级阈值也将被覆盖。

整体节点

“整体”节点可结合使用两个或两个以上模型块，这样所获得的预测会比通过任意一个模型获得的预测更为准确。 通过结合多个模型的预测，可以避免单个模型的局限性，从而使总体准确性更高。 一般情况下，以这种方式组合的模型所得的结果不但可以与使用单个模型所得的最佳结果相媲美，而且结果通常会更理想。

这种节点结合在“自动分类器”、“自动数值”和“自动聚类”自动建模节点中自动产生。

使用“整体”节点后，可以通过“分析”节点或“评估”节点将这些综合结果的准确性与每个输入模型进行比较。要执行此操作，请确保未选中整体节点中“设置”选项卡上的**过滤出整体模型生成的字段**选项。

输出字段

每个整体节点都可以生成含有综合评分的字段。 该字段名称基于指定的目标字段，其前缀根据具体的字段测量级别（标志、名义（集合）或连续（范围））可以为 *$XF_*、*$XS_* 或 *$XR_*。 例如，若目标字段是一个名为 *response* 的标志字段，则输出字段将为 *$XF_response* 。

置信度或倾向字段。 对于标志和名义字段，附加的置信度或倾向字段将根据整体方法进行创建，详情如下表所示。

表 *29:* 整体方法字段创建

整体方法	**字段名称**
投票 置信度加权投票 原始倾向加权投票 调整倾向加权投票 最高置信度当选	*$XFC_<field>*
平均原始倾向	*$XFRP_<field>*
平均调整的原始倾向	*$XFAP_<field>*

整体节点设置

整体的目标字段。 选择单一字段以用作两个或两个以上上游模型的目标字段。 上游模型可以使用标志、名义或连续目标字段，但其中至少要有两个模型必须共享同一目标字段以便获得综合评分。

过滤掉由整体模型生成的字段。 从输出中移除由各个模型生成的所有附加字段，这些模型均输入到“整体”节点中。 如果只想关注所有输入模型中的综合评分，请选中此复选框。 如果希望使用分析节点或评估节点将综合评分的准确性与各个输入模型评分的准确性进行比较，则请确保取消选中此选项。

可用设置取决于选作目标字段的字段的测量级别。

连续目标

对于连续目标，将会对其评分求平均值。 这是求综合评分唯一可用的方法。

求评分或估计值的平均值时，“整体”节点使用标准误差计算得出测量值或估计值与真值之间的差值，并显示这些估计值的接近程度。 缺省对新模型生成标准误差计算；但是，您可以为现有模型取消选择复选框，例如当要对其重新生成时。

类别目标

类别目标可支持包括 **投票** 在内的许多方法，其工作原理是计算每种可能的预测值的选择次数并选择总数最高的值。 例如，如果 5 个模型中有 3 个模型预测为是，另外 2 个预测为否，那么是将以 3 比 2 的票数获胜。 或者，根据每个预测的置信度或倾向值，投票可以是**加权**形式。 然后对加权求和，并再次选择总数最高的值。 最终预测的置信度是指取胜值的加权总和除以整体模型中包含的模型数量。

所有分类字段。 标志和名义字段均支持下列方法：

- 投票
- 置信度加权投票
- 最高置信度当选

仅限标志字段。 对于仅限标志字段的情况，还支持以下多种基于倾向的方法：

- 最初倾向加权投票
- 调整倾向加权投票
- 平均原始倾向

- 平均调整的倾向

投票同数。 根据投票方法，可以指定解决投票同数的方法。

- **随机选择**。 随机选择其中一个同数值。
- **最高置信度**。 选择使用最高置信度进行预测的同数值。 请注意，该置信度值无需与所有预测值的最高置信度值相同。
- **原始倾向或经过调整的倾向（仅限标志字段）**。 使用最大绝对倾向预测的同数值，其中绝对倾向的计算方法如下：

```
abs(0.5 - propensity) *
2
```

或者，对于调整后的倾向，绝对倾向计算方法如下：

```
abs(0.5 - adjusted propensity) * 2
```

分区节点

“分区”节点用于生成分区字段，将数据分割为单独的子集或样本，以供模型构建的训练、测试和验证阶段使用。 通过用某个样本生成模型并用另一个样本对模型进行测试，可以预判此模型对类似于当前数据的大型数据集的拟合优劣。

分区节点会生成名义字段，其角色设置为**分区**。 此外，如果数据中已经存在相应的字段，可以使用“类型”节点将其指定为分区。 在这种情况下，不需要单独的“分区”节点。 可以将任何具有两个或三个值的实例化名义字段用作分区，但不能使用标志字段。 有关更多信息，请参阅主题 第 111 页的『设置字段角色』。

可以在一个流中定义多个分区字段，但如果这么做，那么必须在每个用于分区的建模节点的“字段”选项卡中选择一个分区字段。 （如果仅有一个分区字段，则将在启用分区后自动引入此字段。）

启用分区。 要在分析中使用分区，必须在相应的模型构建或分析节点的“模型选项”选项卡中启用分区。 取消选择此选项可以在不删除字段的条件下禁用分区功能。

要基于其他标准（如数据范围或位置）创建分区字段，还可以使用“派生”节点。 有关更多信息，请参阅主题 第 115 页的『“派生”节点』。

示例。 构建 RFM 流以识别积极响应以往营销活动的最新客户时，销售公司的市场营销部可以使用“分区”节点将数据分割到训练分区和检验分区。

限制:

Modeler Spark 监督式学习节点 (XGBoost-AS， MultilayerPerception-AS) 不支持由分区节点定义的数据分区

分区节点选项

分区字段。 指定由该节点创建的字段的名称。

分区。 可以将数据分区为两个样本（训练和测试）或三个样本（训练、测试和验证）。

- **训练和测试**。 将数据分区为两个样本，使您能够用一个样本训练模型并用另一个样本测试模型。
- **训练、测试和验证**。 将数据分区为三个样本，使您能够用一个样本训练模型，用第二个样本测试并精练模型，然后用第三个样本验证得到的结果。 这种方式会相应减小每个分区的大小，但在使用超大型数据集时最为适用。

分区大小。 指定每个分区的相对大小。 如果分区大小之和小于 100%，那么未包含在分区中的记录将被废弃。 例如，如果用户拥有一千万条记录，并已指定 5% 的训练分区大小和 10% 的测试分区大小，那么在运行该节点之后，大约会有五十万条训练记录和一百万条测试记录，其余记录则被丢弃。

值。 指定用于表示数据中每个分区样本的值。

- **使用系统定义的值（“1”、“2”和“3”）**。 使用整数表示每个分区；例如，位于训练样本中的所有记录的分区字段值均为 1。 这样可确保数据能够在不同语言环境之间移动，而且如果分区字段在其他位置进行重新

实例化（例如从数据集读回数据），将保留排列顺序（因此 1 仍将表示训练分区）。 但是，这种值需要一定的解释。

- **将标签附加到系统定义的值**。 将整数与标签组合；例如，训练分区记录的值为 1_Training。 这样，查看数据的人可能识别出具体的值，并且数据可以保留排列顺序。 但是，这种值仅适用于给定的语言环境。
- **使用标签作为值**。 使用不带整数的标签；例如，**Training**。 这使您能够通过编辑标签来指定值。 但是，这也使数据特定于语言环境，而分区列的重新实例化会使值具有自然排列顺序，而不对应其“语义”顺序。

种子。 仅当选中了**可重复分区分配**时才可用。 根据随机数百分比对记录进行抽样或分区时，此选项允许在另一会话中复制相同的结果。 通过指定随机数生成器所使用的起始值，可以确保在每次执行节点时都会分配相同的记录。 输入所需的种子值，或单击 **生成** 按钮自动生成一个随机值。 如果未选中该选项，则每次执行节点时会生成不同的抽样。

注: 对从数据库中读取的记录使用**种子**选项时，可能需要在抽样前使用“排序”节点以确保每次执行节点时都获得相同的结果。 这是因为随机种子依赖于记录的顺序，而在关系数据库中不能保证记录具有这种顺序。 有关更多信息，请参阅主题 第 64 页的『排序节点』。

使用唯一字段来分配分区。 仅当选中了**可重复分区分配**时才可用。 （仅适合第 1 层数据库）选中此复选框以使用 SQL 回送分配记录到分区。 从下拉列表中，选择具有唯一值的字段（例如标识字段）以确保以随机且可重复的方式分配记录。

“数据库源”节点的描述中对数据库分层进行了说明。 有关更多信息，请参阅主题 第 13 页的『“数据库源”节点』。

生成选择节点

使用“分区”节点中的“生成”菜单，可以自动为每个分区生成一个“选择”节点。 例如，可以选择训练分区中的所有记录，以便仅使用此分区获得进一步的求值或分析。

设为标志节点

“设为标志”节点用于根据为一个或多个集合字段定义的分类值派生标志字段。 例如，您的数据集可能包含一个名义字段 *BP*（血压），其值为 *High*、*Normal* 和 *Low*。 为简化数据操作，可以创建一个代表高血压的标志字段，用于指示患者是否患有高血压。

为设为标志节点设置选项

设置字段。 列出测量级别为名义（集合）的所有数据字段。 从列表中选择一个字段，以显示集合中的值。 您可以在这些值中进行选择，以创建标志字段。 请注意，必须先使用上游源节点或“类型”节点对数据进行完全实例化，然后才能查看可用的名义字段（及其值）。 有关更多信息，请参阅主题 第 103 页的『类型节点』。

字段名称扩展。 选择此选项将启用用于指定扩展名的控件，该扩展名将作为后缀或前缀添加到新的标志字段。 缺省情况下，会通过将原始字段名与字段值组合为标签自动创建新的字段名，如 *Fieldname_fieldvalue* 。

可用的集合值。 此处显示前面选择的集合中的值。 选择要为其生成标志的一个或多个值。 例如，如果名为 *blood_pressure* 的字段中的值为 *High* 、 *Medium* 和 *Low* ，则可以选择 *High* 并将其添加到右侧的列表中。 此操作会为具有表示高血压的值的记录创建一个带标签字段。

创建标志字段。 此处列出新创建的标志字段。 可以指定使用字段名扩展控件命名新字段的选项。

True 值。 指定设置标志时节点所用的真值。 缺省情况下，此值为 **T** 。

False 值。 指定设置标志时节点所用的假值。 缺省情况下，此值为 **F** 。

汇总键。 选择此选项将根据下面指定的关键字段对记录进行分组。 选中**按关键字汇总字段**时，只要任何 记录被设为真，便会“打开”组中的所有标志字段。 使用字段选择器可指定将用于汇总记录的关键字段。

重新结构化节点

“重构”节点可用于根据名义字段或标志字段的值生成多个字段。 新生成的字段可包含来自另一个字段或数值标志（0 和 1）的值。 此节点的功能与“设为标志”节点类似， 但更加灵活。 使用这种节点，可以使用另一个

字段的值创建任意类型的字段（包括数值标志）。 随后，您可以对其他下游节点执行汇总或其他操作。（设为标志节点允许您在一个步骤中汇总字段，因此如果要创建标志字段，使用设为标志节点更为方便。）

例如，下列数据集包含一个名义字段 *Account*，该字段的值为 *Savings* 和 *Draft*。 每个帐户均记录了期初余额和当前余额，而且有些客户在每种类型中均有多个帐户。 假设您希望了解每个客户是否拥有特定的帐户类型，如果有，每种帐户类型中有多少资金。 可以使用重新结构化节点为每个 *Account* 值生成一个字段，并选择 *Current_Balance* 作为值。 这样会用给定记录的当前余额填充每个新字段。

表 *30:* 重新结构化之前的数据示例

CustID	帐户	Open_Bal	Current_Bal
12701	汇票	1000	1005.32
12702	储蓄	100	144.51
12703	储蓄	300	321.20
12703	储蓄	150	204.51
12703	汇票	1200	586.32

表 *31:* 重新结构化之后的数据示例

CustID	帐户	Open_Bal	Current_Bal	Account_Draft_ Current_Bal	Account_Savings _Current_Bal
12701	汇票	1000	1005.32	1005.32	$null$
12702	储蓄	100	144.51	$null$	144.51
12703	储蓄	300	321.20	$null$	321.20
12703	储蓄	150	204.51	$null$	204.51
12703	汇票	1200	586.32	586.32	$null$

将重新结构化节点与“汇总”节点一起使用

在许多情况下，可能需要将“重构”节点与“汇总”节点配对使用。 在上一个示例中，一个客户（标识为 12703）有三个帐户。 可以使用“汇总”节点计算每种帐户类型的总余额。 关键字段为 *CustID* ，且汇总字段是重新结构化字段 *Account_Draft_Current_Bal* 和 *Account_Savings_Current_Bal* 。 下表显示了结果。

表 *32:* 重新结构化并汇总之后的数据示例

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

为重新结构化节点设置选项

可用字段。 列出测量级别为名义（集合）或标志的所有数据字段。 从列表中选择一个字段，以显示集合（或标志）中的值；随后在这些值中进行选择，以创建重构字段。 请注意，必须先使用上游源节点或“类型”节点对数据进行完全实例化，然后才能查看可用的字段（及其值）。 有关更多信息，请参阅主题 第 103 页的『类型节点』。

可用值。 此处显示前面选择的集合中的值。 选择要生成重新结构化字段的一个或多个值。 例如，如果名为 *Blood Pressure* 的字段中的值为 *High* 、 *Medium* 和 *Low* ，您可以选择 *High* 并将其添加到右侧的列表中。 此操作会为值为 *High* 的记录创建一个具有指定值的字段（请参阅下文）。

创建重构字段。 此处列出新创建的重构字段。 缺省情况下，会通过将原始字段名与字段值组合为标签自动创建新的字段名，如 *Fieldname_fieldvalue* 。

包含字段名称。 取消选择此选项将禁止将原始字段名用作新字段名称的前缀。

使用其他字段中的值。 指定一个或多个字段，其值将用于填充重构的字段。 使用字段选择器选择一个或多个字段。 对于选择的每个字段，都会创建一个新字段。 值字段名称将追加到重构的字段名称之后；例如，*BP_High_Age* 或 *BP_Low_Age*。 每个新字段都会继承原始值字段的类型。

创建数字值标志。 选择此选项可使用数字值标志（0 表示假，1 表示真）填充新字段，而不是使用另一个字段的值。

转置节点

缺省情况下，列为字段，而行为记录或观测值。 如有必要，可使用转置节点交换行和列中的数据，使字段变为记录、记录变为字段。 例如，如果有时间序列数据，其中每个序列均为一行而不是一列，则可以在分析之前转置数据。

设置“转置”节点的选项

在**转置方法**下拉菜单中，选择想要“转置”节点执行的方法：**字段和记录**、**记录到字段**或**字段到记录**。 以下部分中描述了所有这三种方法的设置。

限制: 仅在 Windows 64 位、Linux 64 位和 Mac 上支持**记录到字段**和**字段到记录**方法。

字段和记录

可以根据指定的前缀自动生成新字段名称，也可以从数据中的现有字段读取新字段名称。

使用前缀。 此选项将根据指定的前缀（`Field1` 和 `Field2` 等）自动生成新字段名称。 您可以根据需要定制前缀。 如果使用此选项，那么必须指定要创建的字段数，而与原始数据中的行数无关。 例如，如果 **新字段数** 设为 100，前 100 行以外的所有数据都将被丢弃。 如果原始数据中的行数少于 100，那么有些字段将为空。 （可以根据需要增加字段数，但此设置的目的是避免将一百万条记录转置为一百万个字段，因为这会产生无法管理的结果。）

例如，假定数据中包含按行显示的序列，并且每个月有一个单独的字段（列）。 您可以转置此数据，使每个序列包含在一个单独的字段中，每一行表示一个月。

从字段中读取。 从现有字段读取字段名称。 使用此选项，新字段数将由数据决定，最多可达到指定的最大值。 选定字段的每个值都变为输出数据中的一个新字段。 选定的字段可具有任何存储类型（整数、字符串、日期等），但为避免出现重复的字段名称，选定字段的每个值都必须唯一（换言之，值的数目应与行数匹配）。 如果遇到重复的字段名，将显示警告消息。

- **读取值**。 如果选定的字段尚未实例化，那么选择此选项将填充新字段名称的列表。 如果字段已实例化，那么不必执行此步骤。
- **要读取的最大值数目**。 从数据中读取字段名称时，需要指定上限以避免创建过多的字段。 （如上所述，将一百万条记录转置为一百万个字段会产生无法管理的结果。）

例如，如果数据中的第一列指定了每个序列的名称，您可以将这些值用作转置数据中的字段名。

转置。 缺省情况下，仅对连续（数值范围）字段（存储类型为整数或实数）进行转置。 另外，您也可以选择数字字段的子集，或改为转置字符串字段。 但是，所有转置的字段都必须具有相同的存储类型，可以是数字或字符串，但不能同时是这两者，因为混合输入字段会在每个输出列中生成混合的值，而这会违反一个字段中的所有值必须具有相同存储类型的原则。 其他存储类型（日期、时间、时间戳）不能进行转置。

- **所有数字**。 转置所有数字字段（存储类型为整数或实数）。 输出中的行数必须与原始数据中的数字字段数匹配。
- **所有字符串**。 转置所有字符串字段。
- **定制**。 允许您选择数字字段的子集。 输出中的行数必须与所选的字段数匹配。 该选项仅适用于数字字段。

行标识名称。 指定由节点创建的行标识字段的名称。 此字段的值由原始数据中的字段名称确定。

提示: 将时间序列数据从行转置为列时，如果原始数据中有一行（如日期、月或年）带有每个测量周期的标签，请确保将这些标签作为字段名称读入 IBM SPSS Modeler（如前面的示例所示，将原始数据中的月或日期分别显示为字段名称），而不是在第一行数据中包含标签。 这样将避免在每一列中混合标签和值（这会将数值强制读取为字符串，因为一列中不能混合存储类型）。

记录到字段

字段。 “字段”列表包含进入“转置”节点的所有字段。

索引。 使用“索引”部分来选择要用作索引字段的字段。

字段。 使用“字段”部分来选择要用作字段的字段。

值。 使用“值”部分来选择要用作值字段的字段。

聚集函数。 如果某索引对应多个记录，必须将这些记录汇总为一条记录。 使用**汇总函数**下拉列表来指定如何使用以下某个函数来汇总记录。 请注意，汇总影响所有字段。

- **平均值。** 返回每个关键字段组合的均值。 该均值是对集中趋势的测量，它是算术平均值（总和除以观测值数）。
- **总和** 返回每个关键字段组合的合计值。 总和是指所有具有非缺失值的观测值中值的总计。
- **最短** 返回每个关键字段组合的最小值。
- **最大值。** 返回每个关键字段组合的最大值。
- **中位数。** 返回每个关键字段组合的中值。 中位数是对集中趋势的测量，但对于远离中心的值不敏感（这与均值不同，均值容易受到少数极大或极小值的影响）。 也称为第 50 个百分位数或第二个四分位数。
- **计数。** 返回每个关键字段组合的非空值计数。

字段到记录

字段。 “字段”列表包含进入“转置”节点的所有字段。

索引。 使用“索引”部分来选择要用作索引字段的字段。

值。 使用“值”部分来选择要用作值字段的字段。 如果不选择任何值字段，那么所有未分配的数字字段将用作值。 但如果数字字段不可用，那么将使用所有未分配的字符串字段。

历史记录节点

“历史记录”节点最常用于顺序数据，例如时间序列数据。 这种节点用于创建包含先前记录中字段的数据的新字段。 使用“历史记录”节点时，可能需要使用按特定字段预先排序的数据。 可以使用排序节点执行此操作。

为历史记录节点设置选项

所选字段。 使用“字段选择器”（文本框右边的按钮）选择需要使用其历史记录的字段。 每个所选字段将用于创建数据集中所有记录的新字段。

偏移量。 指定要从中抽取历史字段值的当前记录之前的最新记录。 例如，如果“偏移量”设为 3，则当每条记录通过此节点时，之前第三条记录的字段值将包括在当前记录中。 使用“范围”设置可指定从中提取记录的记录后退范围。 使用箭头可调整偏移量值。

范围。 指定要从中抽取值的以前记录的数目。 例如，如果“偏移量”设为 3 且“范围”设为 5，那么通过该节点的每条记录将针对“选定字段”列表中指定的每个字段添加五个字段。 这意味着当节点在处理记录 10 时，将从记录 7 到记录 3 添加字段。 使用箭头来调整范围值。

历史记录不可用时。 选择下列其中一个选项以用于处理没有历史记录值的记录。 这通常是指数据集顶端的前几条记录，它们没有可用作历史记录的先前记录。

- **废弃记录。** 选择此选项将废弃对于所选字段没有可用历史记录值的记录。
- **保留未定义的历史记录。** 选择此选项将保留没有可用历史记录值的记录。 将使用未定义的值来填充历史记录字段，显示为 `$null$`。

- **填入值**。 指定要用于没有可用历史记录值的记录的值或字符串。 缺省的替换值为系统空值 *undef* 。 使用字符串 `$null$` 来显示空值。

为实现正确执行，在选择替换值时请记住以下规则：

- 所选字段应属于同一存储类型。
- 如果所有选定字段的存储类型均为数字，替换值必须解析为整数。
- 如果所有选定字段的存储类型均为实数，替换值必须解析为实数。
- 如果所有选定字段的存储类型均为符号，替换值必须解析为字符串。
- 如果所有选定字段的存储类型均为日期/时间，替换值必须解析为日期/时间字段。

如果上述任一条件不成立，则会在执行历史记录节点时收到错误警告。

字段重排节点

使用“字段重排”节点，可以定义用于显示下游字段的自然顺序。 此顺序将影响字段在各种位置（例如表、列表和字段选择器）的显示方式。 例如，使用大型数据集时，此操作有助于使所需字段更直观。

设置字段重排选项

对字段进行重新排序的方法有两种：定制排序和自动排序。

自定义排序

选择**自定义顺序**可启用一个包含字段名和类型的表格，您可在其中查看所有字段并使用箭头按钮创建自定义顺序。

要对字段进行重新排序，请执行下列操作：

1. 在表中选择一个字段。 采用按住 Ctrl 并单击的方法可选择多个字段。
2. 使用简单的箭头按钮可将字段上移或下移一行。
3. 使用行箭头按钮可将字段移至列表底部或顶部。
4. 通过将分隔线（标为 **[其他字段]** ）上移或下移，指定此处未包括的字段的顺序。

有关 [其他字段] 的详细信息

其他字段。 [其他字段] 分割线的用途是将表格分为两半。

- 显示在分隔线以上的字段（以它们出现在表中的顺序）排序后，将在此节点下游字段的所有自然顺序的顶端显示。
- 显示在分隔线以下的字段（以它们出现在表中的顺序）排序后，将在此节点下游字段的所有自然顺序的底端显示。

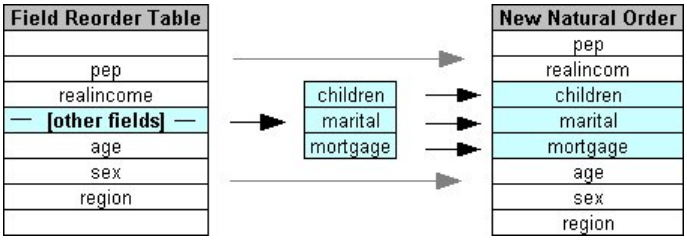


图 *6:* 说明“其他字段”在新字段顺序中的结合方式的图表

- 未出现在字段重排表中的所有其他字段将在这些“顶端”和“底端”字段之间显示，按分隔线的位置标示。

其他定制排序选项包括：

- 通过单击每个列标题（ **类型** 、 **名称** 和 **存储类型** ）上方的箭头可按升序或降序对字段进行排序。 按列排序时，未在此指定的字段（按 **[其他字段]** 行标识）将排在其自然顺序的最后。
- 单击 **清除未使用的** 可将所有未使用的字段从字段重排节点中删除。 未使用的字段在表中以红色字体显示。 这表示该字段已在上游操作中被删除。

- 指定任意新字段（显示有闪电图标，表示新的或未指定的字段）的排序方式。 单击 **确定** 或 **应用** 时，该图标将消失。

注意：如果应用定制顺序后在上游添加了字段，那么会将新字段追加到定制列表的底部。

自动排序

选择 **自动排序** 以指定排序参数。 对话框选项将动态变化，以提供用于自动排序的选项。

排序依据。 选择三种方式之一，对读入“重排”节点的字段进行排序。 箭头按钮将指示顺序为升序还是降序。 选择一种以进行更改。

- 名称
- Type
- 存储

应用自动排序后在字段重排节点上游添加的字段将根据所选排序类型被自动置于适当位置。

时间间隔节点

SPSS Modeler V17.1 和更早版本中的原始“时间间隔”节点与 Analytic Server (AS) 不兼容，并在 SPSS Modeler R18.0 中已不推荐使用。

替换“时间间隔”节点包含对原始“时间间隔”节点的一系列更改。 该新节点可以与 Analytic Server 一起使用，也可以由 SPSS Modeler 单独使用。

“时间间隔”节点用于指定时间间隔并派生用于估算或预测的新时间字段。 支持全部范围的时间间隔，从秒到年。

使用此节点可以派生新的时间字段；新字段的存储类型与您选择的输入时间字段相同。 此节点将生成下列各项：

- 在“字段”选项卡上指定为**时间字段**的字段以及选择的前缀/后缀。 缺省情况下，前缀为 `$TI_`。
- 在“字段”选项卡上指定为**维度字段**的字段。
- 在“字段”选项卡上指定为**要汇总的字段**的字段。

另外，还可以生成若干其他字段，具体取决于选择的时间间隔或周期（例如测量值所处的分钟或秒）。

时间间隔 - 字段选项

使用“时间间隔”节点中的“字段”选项卡选择从中派生新时间间隔的数据。

字段 显示此节点的所有输入字段及其测量类型图标。 所有时间字段都具有“连续”测量类型。 请选择要用作输入的字段。

时间字段 显示要从中派生新时间间隔的输入字段；只允许使用单个连续字段。 “时间间隔”节点将此字段用作用于转换时间间隔的汇总键。 新字段的存储类型与选择的输入时间字段相同。 如果您选择了整数字段，那么该字段将被视为时间索引。

维度字段 您可以选择性地在此处添加字段，以便创建单个基于字段值的时间序列。 举一个简单的例子，对于地理空间数据，可以使用点字段作为维度。 在此示例中，来自“时间间隔”节点的数据输出排序成点字段中每个点值的时间序列。

当您使用平面化多维数据（类似于 TM1 节点生成的数据）或者为了支持类似地理空间之类的更加复杂的数据类型时，维度是理想选择。 从本质上来说，您可以考虑使用**维度字段**作为 SQL 查询中的**分组依据**子句的等效项，或者类似于“汇总”节点中的**关键字段**；但是，**维度字段**在本质上更加复杂，因为它能够处理比只有传统行和列数据更加复杂的数据结构。

要汇总的字段 选择在更改时间字段的时间段过程中要汇总的字段。 只有此处选择的字段才可用于“构建”选项卡上的**所指定字段的定制设置**表。 在离开此节点的数据中，将过滤掉所有未在此处包括的字段。 这意味着，将从该数据中过滤掉**字段**列表中的所有余下字段。

时间间隔 - 构建选项

使用“构建”选项卡可以指定用于更改时间间隔的选项，以及指定如何根据数据的测量类型对其中的字段进行汇总。

对数据进行汇总时，现有的所有日期、时间或时间戳记字段都将被生成的字段取代，并从输出中删除。 其他字段将根据此选项卡中指定的选项进行汇总。

时间间隔 请选择用于构建序列的时间间隔和周期长度。

缺省设置 请选择要应用于不同类型的数据的缺省汇总。 缺省汇总将根据测量级别进行应用；例如，连续字段按合计进行汇总，而名义字段则使用众数。 您可以为三种不同的测量级别设置缺省汇总：

- **连续** 可用于连续字段的函数包括**合计**、**均值**、**最小值**、**最大值**、**中位数**、**第一个四分位数**和**第三个四分位数**。
- **名义** 选项包括**众数**、**最小值**和**最大值**。
- **标志** 选项为**如果任意一项为 true 则为 true** 和**如果任意一项为 false 则为 false**。

所指定字段的定制设置 您可以指定各个字段的缺省汇总设置的例外情况。 请使用右侧的图标在表中添加或移除字段，或者单击相应列中的单元格以更改用于该字段的汇总函数。 无类型字段将从列表中排除，且不能添加到表中。

新字段名称扩展 请指定要对此节点所生成的所有字段应用的**前缀**或**后缀**。

“重新投影”节点

对于地理空间数据或地图数据，最常用的两种指示坐标的方法是投影坐标系和地理坐标系。 在 IBM SPSS Modeler 中，表达式构建器空间函数、“空间-时间预测”(STP) 节点和“地图可视化”节点之类的项使用投影坐标系，因此必须对导入的所有使用地理坐标系进行记录的数据执行重新投影。 在有可能的情况下，将在使用（而不是导入）地理空间字段（任何具有地理空间测量级别的字段）时自动对其进行重新投影。 存在任何无法自动进行重新投影的字段时，您可以使用“重新投影”节点来更改其坐标系。 以此方式进行重新投影意味着您可以解决由于使用了不正确的坐标系而出错的情况。

以下列表显示了可能必须进行重新投影以更改坐标系的示例情况：

- 追加 如果尝试针对地理空间字段追加两个具有不同坐标系的数据集，那么 SPSS Modeler 显示以下错误消息：*<Field1>* **和** *<Field2>* **的坐标系不兼容。 请将其中某个或全部字段重新投影到同一个坐标系。**

  <Field1> 和 *<Field2>* 是导致错误的地理空间字段的名称。
- *If/else* 表达式 如果使用的表达式包含 if/else 语句，其中表达式的两部分中含地理空间字段或返回类型，但是使用不同的坐标系，那么 SPSS Modeler 显示以下错误消息：**条件表达式包含不兼容的坐标系：***<arg1>* **和** *<arg2>*。

  <arg1> 和 *<arg2>* 是返回使用不同坐标系的地理空间类型的 then 或 else 自变量。
- 构造一列地理空间字段 要创建由许多地理空间字段组成的列表字段，向列表表达式提供的所有地理空间字段自变量必须采用同一个坐标系。 如果不是，那么将显示以下错误消息：*<Field1>* **和** *<Field2>* **的坐标系不兼容。 请将其中某个或全部字段重新投影到同一个坐标系。**

有关坐标系的更多信息，请参阅《SPSS Modeler 用户指南》的『使用流』部分中的『设置流的地理空间选项』主题。

为“重新投影”节点设置选项

字段

地理字段
缺省情况下，此列表为空。 您可以将地理空间字段从**要重新投影的字段**列表移入此列表，以确保不对这些字段进行重新投影。

要重新投影的字段
缺省情况下，此列表包含所有作为此节点的输入的地理空间字段。 此列表中的所有字段都将重新投影到**坐标系**区域中设置的坐标系。

坐标系

流缺省值
选择此选项表示使用缺省坐标系。

指定
如果选择此选项，那么可以使用**更改**按钮来显示**“选择坐标系”**对话框，并选择要用于进行重新投影的坐标系。

有关坐标系的更多信息，请参阅《SPSS Modeler 用户指南》的『使用流』部分中的『设置流的地理空间选项』主题。

第 5 章 图形节点

通用图形节点功能

数据挖掘过程的多个阶段都会使用图形和图表探索导入到 IBM SPSS Modeler 中的数据。 例如，可将“散点图”或“分布”节点连接到数据源，以了解数据类型和数据分布。 然后可以执行记录和字段操作，以准备下游建模操作的数据。 图形的另一个常见用途是检查新导出字段的分布和它们之间的关系。

“图形”选项板含有以下节点：

“图形板”节点在单个节点中提供许多不同类型的图形。 使用此节点，可以选择要探索的数据字段，然后从适用于选定数据的字段中选择 一个图形。 节点将自动过滤出适用于字段选项的所有图形类型。

散点图节点可显示数字字段间的关系。 可通过使用点（散点）或线创建散点图。

“分布”节点显示符号（分类）值（例如抵押类型或性别）的出现次数。 通常，您可以使用“分布”节点来显示数据中的不平衡，然后可以在创建模型前使用“均衡”节点来纠正此类不均衡。

“直方图”节点显示数字字段的值的出现次数。 此节点经常用来在进行数据操作和模型构建前探索数据。 与“分布”节点相似，“直方图”节点经常用来揭示数据中的不平衡。

“收集”节点显示一个数字字段的值相对于另一个数字字段的值的分布。 （它创建类似于直方图的图形。） 图示说明值不断变化的变量或字段时，它 是有用的。 使用 3-D 图形表示时，还可以使用按分类显示分布的符号轴。

“多重散点图”节点创建在单个 *X* 字段上显示多个 *Y* 字段的散点图。 *Y* 字段被绘制为彩色的线；每条线相当于“样式”设置为**线**且“X 模式”设置为**排序**的散点图节点。 在探索多个变量随时间推移的变化情况时，多重散点图非常有用。

Web 节点说明两个或两个以上符号（分类）字段的值之间的关系强度。 此图使用不同粗细的线条来表示连接强度。 例如，您可以使用 Web 节点来探索在电子商务站点上购买一组商品之间的关系。

“时间散点图”节点显示一组或多组时间序列数据。 通常，您将首先使用“时间间隔”节点来创建 *TimeLabel* 字段，该字段将用于标注 *x* 轴。

“评估”节点有助于评估和比较预测模型。 评估图表显示模型预测特定结果的优劣程度。 它根据预测值和预测置信度对记录进行排序。 它将记录分成若干个相同大小的组（**分位数**），然后从高到底为每个分位数划分业务标准值。 在散点图中，将以单独的线条显示多个模型。

“地图可视化”节点可以接受多个输入连接，并在地图上将地理空间数据显示为一系列层。 每个层都是单个地理空间字段；例如，底层可能是国家或地区的地图，在其之上可能存在一个道路层、一个河流层和一个城镇层。

E-Plot (Beta) 节点显示数字字段之间的关系。 它与“绘图”节点类似，但是其选项不同，并且其输出使用特定于此节点的新图形界面。 使用 beta 级节点可运用新图形功能。

t-分布随机邻域嵌入 (t-SNE) 是用于可视化高维数据的工具。 其将数据点亲缘关系转换为可能性。 此 t-SNE 节点在 SPSS Modeler 中使用 Python 进行实现并且需要 `scikit-learn`© Python 库。

将图形节点添加到流后，可双击节点以打开用于指定选项的对话框。 绝大多数图形都含有一些独特的选项，这些选项会显示在一个或多个选项卡上。 除此以外，还有若干通用于所有图形的选项卡选项。 以下章节包含有关这些通用选项的更多信息。

配置图形节点的选项后，可通过对话框运行该节点或将它作为流的组成部分来运行。 可在已生成图形窗口中根据选择或数据区域生成“派生”（集合和标记）和“选择”节点，有效地将数据划分为多个“子集”。 例如，可使用此强大功能来识别和排除离群值。

审美原则、重叠、面板和动画

重叠和审美原则

外观（和重叠）向可视化添加维数。 外观效果（分组、聚类或堆积）取决于直观表示类型、字段（变量）类型和图形元素类型以及统计值。 例如，颜色的分类字段可能用于对散点图中的点分组或在堆积条形图中创建堆积。 用于颜色的连续数字范围也可以用于指示散点图中每一个点的范围值。

您应该使用外观和重叠进行试验以找到一个满足需要的值。 以下描述也许能够帮助您选出正确的一个。

注：并非所有外观或重叠都可以用于所有可视化类型。

- **颜色**。 当用分类字段定义颜色时，它将根据单个类别拆分可视化，每个类别一种颜色。 当颜色是连续数值范围时，根据范围字段的值颜色各不相同。 如果图形元素（例如，条形图或箱图）代表多个记录/个案，且一个范围字段用于颜色，则颜色根据范围字段的平均值各不相同。
- **形状**。 分类字段定义形状，其将可视化拆分为不同形状的元素，每个类别一个。
- **透明度**。如果透明度由一个分类字段定义，该分类/字段将根据每个类别对直观表示进行分割，每个类别一个透明度级别。 当透明度为连续数字范围时，它会根据范围字段的值发生变化。 如果图形元素（例如，条形图或箱图）代表多个记录/个案，且一个范围字段用于透明度，则颜色根据范围字段的平均值各不相同。 处于最大值时，图形元素为完全透明。 在最小值处，则完全不透明。
- **数据标签**。数据标签由一个任意类型的字段定义，其值用于创建附加到图形元素的标签。
- **大小**。 当用分类字段定义大小时，它将根据单个类别拆分可视化，每个类别一种大小。 当大小是连续数值范围时，根据范围字段的值大小各不相同。 如果图形元素（例如，条形图或箱图）代表多个记录/个案，且一个范围字段用于大小，则大小根据范围字段的平均值各不相同。

面板和动画

面板。面板，也称为刻面，创建一个图形表。 虽然在镶面字段中为每一个类别生成了一个图形，但所有的面板都同时显示。 镶面对于检查可视化是否取决于镶面字段的条件非常有用。 例如，您可以按性别生成直方图面板以确定频率分布在男性和女性中是否相等。 即，您可以检查薪水是否取决于性别差异。 选择一个用于镶面的分类字段。

动画。动画与面板类似，因为从动画字段的值中创建了多个图形，但是这些图形不一起显示。 相反，您使用探索模式的控件以动画来显示输出并变换通过单个图形的序列。 此外，与镶面不同，动画并不需要分类字

段。 您可以指定一个连续字段，其值自动被拆分到范围中。 可以在探索模式中使用动画控件改变范围的大小。 并不是所有可视化都提供动画。

使用“输出”选项卡

对于所有图形类型，均可为已生成图形的文件名和显示指定以下选项。

注: 分布节点图具有其他设置。

输出名称。 指定运行节点时产生的图形名称。 **自动** 根据生成输出的节点选择名称。 （可选）可以选择**定制**以指定其他名称。

输出到屏幕。 选择此选项将在新窗口中生成并显示图形。

输出到文件。 选择此选项可将输出另存为文件。

- **输出图形**。 选择此选项将以图形格式产生输出。 仅在“分布”节点中可用。
- **输出表**。 选择此选项将以表格式产生输出。 仅在“分布”节点中可用。
- **文件名**。 指定用于所生成图形或表的文件名。 使用省略符按钮 **(...)** 指定具体文件和位置。
- **文件类型**。 在下拉列表中指定文件类型。 对于所有图形节点，除了具有**输出表**选项的分布节点外，可用的图形文件类型如下：

  - 位图 (.bmp)

  - PNG (.png)

  - 输出对象 (.cou)

  - JPEG (.jpg)

  - HTML (.html)

  - ViZml 文档 (.xml)，可在其他 IBM SPSS Statistics 应用程序中使用。

  对于分布节点上的**输出表**选项，可用的文件类型如下：

  - 制表符分隔的数据 (.tab)

  - 逗号分隔的数据 (.csv)

  - HTML (.html)

  - 输出对象 (.cou)

分页输出。 如果将输出保存为 HTML，那么将启用此选项以使您可以控制每个 HTML 页面的大小。 （仅应用于分布节点。）

每页行数。 如果选择了**编页码输出**，那么将启用此选项，以便您能够确定各个 HTML 页面的长度。 缺省情况下，设置为 400 行。 （仅应用于分布节点。）

使用“注释”选项卡

用于所有节点，此选项卡提供的选项可用于重命名节点、提供定制的工具提示及存储长的注解。

3D 图形

IBM SPSS Modeler 中的散点图和集合图能够在第三坐标轴上显示信息。 这样就可以更加灵活地对数据进行可视化，以选择子集或派生用于建模的新字段。

创建 3D 图形后，即可单击图形并拖动鼠标旋转图形，以便从任何角度查看图形。

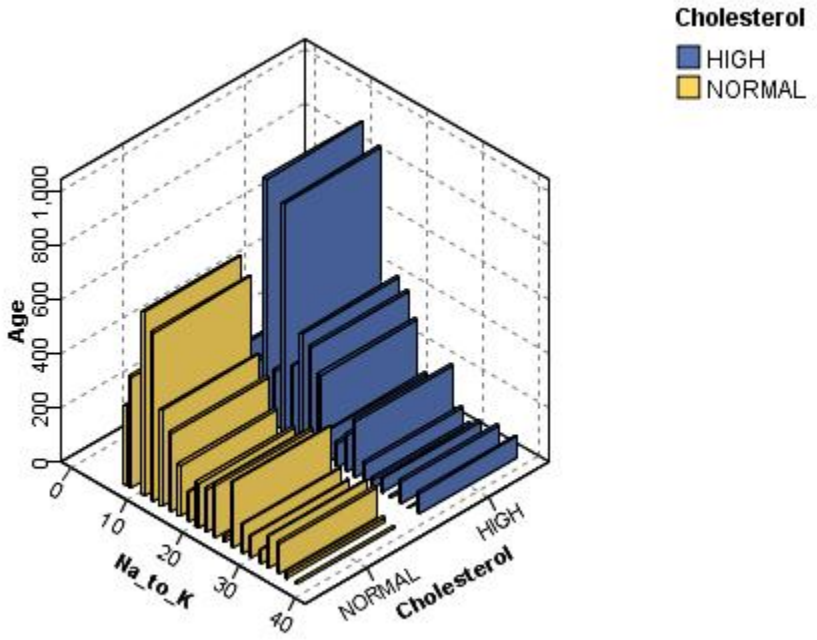


图 *7:* 带有 *x*、*y* 和 *z* 轴的集合图

在 IBM SPSS Modeler 中创建三维图形有两种方法：在第三坐标轴上绘制信息（真实的三维图形）和以三维效果显示图形。 散点图和收集都可以使用这两种方法。

在第三坐标轴上绘制信息散点图

1. 单击图形节点对话框中的**散点图**选项卡。
2. 单击三维按钮以启用 *z* 轴选项。
3. 使用“字段选择器”按钮以选择 *z* 轴的字段。 某些情况下只允许使用符号字段。 “字段选择器”将显示适合的字段。

向图形中添加 3D 效果

1. 创建图形后，即可单击输出窗口中的**图形**选项卡。
2. 单击 3D 按钮以将视图切换为 3D 图形。

图形板节点

通过“图形板”节点，您可以在一个节点中从许多不同的图形输出（条形图、饼图、直方图、散点图、热图等） 中进行选择。 从第一个选项卡开始，选择需要探索的数据字段，然后节点将提供一个适用于数据的图形类型的选项。 节点将自动过滤出适用于字段选项的所有图形类型。 在“详细”选项卡上，可以定义详细的选项或较高级的图形选项。

注: 为了编辑节点或选择图形类型，必须将“图形板”节点连接到包含数据的流。

有两个按钮可用于控制哪个可视化模板（以及样式表和图）可用：

管理。 管理计算机上的可视化模板，及样式表与地图。 您可以导入、导出、重命名和删除本地计算机上的可视化模板，样式表与地图。 有关更多信息，请参阅第 164 页的『管理模板、样式表和地图文件』主题。

位置。 更改可视化模板，样式表与地图的存储位置。 当前位置列出在按钮右侧。 有关更多信息，请参阅第 163 页的『设置模板、样式表和地图位置』主题。

图形板基本选项卡

如果无法确定哪种直观表示类型最适于表示您的数据，请使用“基本”选项卡。 当您选择数据时，然后为您显示适合数据的可视化类型子集。 有关示例，请参阅第 154 页的『图形板示例』。

1. 从列表中选择一个或多个字段（变量）。使用 Ctrl+单击以选择多个字段。

   请注意，字段的测量级别决定了可用的直观表示类型。 您可以在列表中右键单击字段并选择一个选项，以更改测量级别。 有关可用测量级别类型的更多信息，请参阅第 146 页的『字段（变量）类型』。
2. 选择可视化类型。 有关可用类型的描述，请参阅第 148 页的『可用的内置图形板可视化类型』。
3. 对于某些直观表示，您可以选择一个汇总统计。 哪些统计子集可用，取决于该统计是基于计数还是根据连续的字段计算。 可用统计也取决于模板本身。 下一步可能是可用的完整统计列表。
4. 如果要定义多个选项（例如可选审美原则和面板字段），请单击**详细**。 请参阅主题第 147 页的『图形板详细选项卡』，了解更多信息。

根据连续字段计算的汇总统计

- 平均值。集中趋势的测量。 算术平均值，等于总和除以观测值数。
- 中位数。大于或小于一半观测值的值，即 50th 个百分位。 如果有偶数个观测值，则中位数为它们以升序或降序排列时两个中间观测值的平均值。 中位数是集中趋势的一种测量，对离群值不敏感（与平均值不同，平均值会受部分极高或极低值的影响）。
- 众数 *(Mode)*。最常出现的值。 如果多个值共享最大出现频率，则每个值都是一个众数。
- *Minimum*。数值变量的最小值。
- *Maximum*。数值变量的最大值。
- 范围。 最大值与最小值之间的差。
- 平均范围。 范围中间值，即与最小值的差等于与最大值的差的值。
- *Sum*。所有带有非缺失值的观测值的值的合计或总计。
- 累积和。 值的累积总和。 每个图形元素显示一个子组的和加上所有先前组的总和。
- 百分比和。 每个子组中根据求和字段与所有组上的和对比所得的百分比。
- 累积百分比和。 每个子组中根据求和字段与所有组上的和对比所得的累积百分比。 每个图形元素显示一个子组的百分比加上所有先前组的总百分比。
- 偏差。对围绕平均值的离差的测量，值等于与平均值的差的平方和除以个案数减一。 方差按单元计量，即变量自身单元数的平方。
- 标准差。对围绕平均值的离差的测量。 在正态分布中，68% 的观测值落入与均值相距不到一个标准差的范围内，95% 落入两个标准差的范围内。 例如，如果平均年龄值 45，标准差为 10，则 95% 的观测值将介于正态分布的 25 到 65 之间。
- 标准错误 *(Standard Error)*。检验统计量值因样本而异的测量。 这是统计量抽样分布的标准差。 例如，均数的标准误差是样本均数的标准差。
- 峰度 *(Kurtosis)*。存在离群值的程度的测量。 对于正态分布，峰度统计量的值为零。 正峰度值表示数据呈现比正态分布更极端的离群值。 负峰度值表示数据呈现比正态分布极端程度较低的离群值。
- 偏度 *(Skewness)*。分布不对称性的测量。 正态分布是一种对称性分布，其偏度值为 0。 具有显著性正偏度的分布右侧尾部较长。 具有显著负偏态的分布具有向左延伸的长尾。 提示：取大于其标准误差两倍的偏度值指示离开对称的距离。

以下区域统计可能导致每个子组有多个图形元素。 使用间隔、面积或边缘图形元素时，区域统计会导致一个显示范围的图形元素。 所有其他图形元素导致两个独立元素，一个显示范围的开始，一个显示范围的结束。

- **区域：范围**。 最小值与最大值之间的值范围
- **区域：95% 均值置信度区间**。 有 95% 机会包含总体平均值的一系列值。
- **区域：95% 单值置信度区间**。 有 95% 机会包含给定单个个案的预测值的一系列值。
- **区域：平均值上/下 1 个标准差**。 平均值上下 1 个标准差之间的一些列值
- **区域：平均值上/下 1 个标准误差**。。 平均值上下 1 个标准误差之间的一系列值

基于计数的汇总统计

- **计数**。 行/个案数量。
- **累积计数**。 行/个案累积数量。 每个图形元素显示一个子组的计数加上所有先前组的总计数。
- **计数百分比**。 每个子组中行/个案数量对比行/个案的总数百分比。
- **计数累积百分比**。 每个子组中行/个案数量对比行/个案的总数的累积百分比。 每个图形元素显示一个子组的百分比加上所有先前组的总百分比。

字段（变量）类型

图标显示在字段列表中的字段旁边，以指示字段类型和数据类型。 图标同时识别多响应集。

表 *33:* 测量级别图标

测量级别	数字	String	日期	时间
连续		不适用		
排序集合				
设置				

表 *34:* “多响应集”图标

“多响应集”类型	图标
多响应集，多类别	
多响应集，多二分	

测量级别

创建可视化时，字段的测量级别很重要。 以下是对于测量级别的描述。 您可以右键单击字段列表中的字段并选择一个选项来临时更改测量级别。 在多数情况下，您只需考虑字段的两个最广泛的分类，分类字段和连续字段：

分类。 包含有限数量的不同值或类别（例如，性别或宗教）的数据。 分类字段可以为字符串（字母数字）字段，也可以为使用数字编码表示分类的数字字段（例如，0 = 男 和 1 = 女）。 这种数据也称为定性数据。 集合、有序集合和标志都是分类字段。

- 集合。 其值表示不具有内在等级的类别的字段/变量（例如，员工任职的公司部门）。 名义变量的示例包括地区、邮政编码和宗教信仰。 也称为名义变量。
- 有序集合。 其值表示具有某种内在等级的类别的字段/变量（例如，从十分不满意到十分满意的服务满意度水平）。 排序集合的示例包括表示满意度或可信度的态度分数和优先选择评分。 也称为有序变量。
- 标志。 具有两个不同值的字段/变量，例如 Yes 和 No 或者 1 和 2。 也称为二分变量或二元变量。

连续。 以区间或比率刻度度量的数据，其中数据值既表示值的顺序，也表示值之间的距离。 例如，72,195 美元的薪金高于 52,398 美元的薪金，并且这两个值之间的差距为 19,797 美元。 也称为数量、比例或数字范围数据。

分类字段定义可视化中的类别，通常画出单独的图形元素或将图形元素分组。 经常在分类字段的类别内汇总连续字段。 例如，对于性别类别的缺省收入可视化将显示男性的平均收入和女性的平均收入。 也可像散点

图中一样画出连续字段的原始值。 例如，散点图会显示每个个案的当前薪金和起始薪金。 分类字段可用于按性别分组个案。

数据类型

测量级别不是决定其类型的字段的唯一属性。 字段也可保存为某个特定数据类型。 可能的数据类型有字符串（非数值数据，如字母）、数值（实数）和日期。 与测量级别不同，不能暂时更改字段的数据类型。 您必须更改数据在源数据集中的存储方式。

多重响应集

有些数据文件支持一种名为**多重响应集**的特殊“字段”。 多响应集在通常意义上不是真正的“字段”。 多响应集使用多个字段记录对问题的响应，其中响应者可以给出一个以上的答案。 可将多响应集视为分类字段那样处理，您对分类字段执行的大部分操作，也可以对多响应集执行。

多响应集可以是多二分集或多类别集。

多二分集。 多二分集通常包含多个二分字段：仅具有两个可能值（是/否、存在/不存在、选中/未选中性质）的字段。 虽然字段可能不是严格二分，但集合中的所有字段都用相同方式编码。

例如，调查为以下问题提供了五个可能的回答：“对于新闻，您依赖于以下哪些来源？”响应者可以通过选中每个选择旁的框来进行多项选择。 五个回答在数据文件中变成五个字段，代码 0 为 否（未选中）；代码 1 为是（选中）。

多类别集。 多类别集由多个字段组成，都用相同方式编码，常常具有多个可能的响应类别。 例如，某个调查项目为“请列举最能描述您的种族血统的民族，最多三个”。 可能有上百种回答，但为了进行编码，列表限制为 40 个最常见的民族，任何其他回答都归为“其他”类别。 在数据文件中，三个选项变为三个字段，每个具有 41 个类别（40 个编码国家和一个“其他”类别）。

图形板 详细选项卡

当您知道您想创建什么类型的可视化或当您想将可选外观、面板和/或动画添加到可视化中时，请使用“详细”选项卡。 有关示例，请参阅第 154 页的『图形板示例』。

1. 如果您在“基本”选项卡上选择了一个可视化类型，将显示该类型。 否则，从下拉列表中选择一个。 有关直观表示类型的信息，请参阅第 148 页的『可用的内置图形板可视化类型』。
2. 直观表示缩略图右侧紧邻的是用于指定直观表示类型所需字段（变量）的控件。 必须指定所有这些字段。
3. 对于某些直观表示，您可以选择一个汇总统计。 在某些情况下（如条形图），您可以使用一种摘要选项用于透明外观。 有关汇总统计的描述，请参阅第 145 页的『图形板基本选项卡』。
4. 可以选择一个或多个可选审美原则。 这些外观可允许您在直观表示中包括其他字段，从而添加维数。 例如，您可以使用字段改变散点图中的点的大小。 有关可选外观的更多信息，请参阅第 142 页的『审美原则、重叠、面板和动画』。 请注意，脚本中不支持透明外观。
5. 如果您要创建地图可视化，**地图文件**组将显示要使用的一个或多个地图文件。 如果有缺省的地图文件，则显示此文件。 要更换地图文件，单击**选择地图文件**以显示“选择地图”对话框。 在此对话框中还可指定缺省地图文件。 有关更多信息，请参阅第 147 页的『选择用于地图可视化的地图文件』主题。
6. 可以选择一个或多个面板或动画选项。 有关面板和动画选项的更多信息，请参阅第 142 页的『审美原则、重叠、面板和动画』。

选择用于地图可视化的地图文件

如果选择了地图可视化模板，则需要地图文件来定义绘制地图的地理信息。 如果有缺省的地图文件，则会将其用于地图可视化。 要选择其他地图文件，在“详细”选项卡上单击**选择地图文件**以显示“选择地图”对话框。

您可通过“选择地图”对话框选择主地图文件和参考地图文件。 这些地图文件定义绘制地图的地理信息。 您的应用程序在安装时自带有一组标准地图文件。 如果要使用其他 ESRI 形状文件，则首先需要将其转换为 SMZ 文件。 请参阅主题第 164 页的『转换和分发地图 Shapefile』，了解更多信息。 转换地图后，单击“模板选择器”对话框上的**管理...**，以将地图导入至管理系统，这样它将在“选择地图”对话框中可用。

以下列出了在指定地图文件时的考虑要点：

- 所有地图模板需要至少一个地图文件。

- 地图文件通常将地图关键字属性链接到数据关键字。
- 如果模板不需要地图关键字链接到数据关键字，则它需要参考地图文件和字段，后者指定了在参考地图上绘制元素的坐标（例如，经度和纬度）。
- 重叠地图模板需要两个地图：主地图文件和参考地图文件。 首先绘制参考地图，因此它位于主地图文件之后。

有关地图术语（例如属性和特征）的信息，请参阅第 165 页的『有关地图的重要概念』。

地图文件。 您可以选择位于管理系统中的任何地图文件。 其中包括预安装地图文件和导入的地图文件。 有关管理地图文件的更多信息，请参阅第 164 页的『管理模板、样式表和地图文件』。

地图关键字。 指定您要用作将地图文件链接到数据关键字的关键字属性。

保存此地图文件和设置为缺省值。 如果您要将选定地图文件用作缺省值，则选择此复选框。 在指定了缺省地图文件之后，您无需在每次创建地图可视化时指定地图文件。

数据关键字。 此控件所列出的值与“模板选择器”的“详细”选项卡上所列值相同。 在此处提供这些值以便于您针对所选的特定地图文件更改关键字。

在可视化中显示所有地图特征。 在选中此复选框后，地图中的所有特征都会在可视化中呈现，即使并不存在匹配的数据关键字值。 如果您只想查看具有数据的特征，请取消选中此选项。 在**不匹配的地图关键字**列表中所示地图关键字标识的特征不会在直观表示中呈现。

比较地图值和数据值。 地图关键字和数据关键字彼此链接以便创建地图可视化。 地图关键字和数据关键字应该从同一域（例如，国家或地区和区域）中绘制。 单击**比较**以测试数据关键字和地图关键字值是否匹配。显示的图标将通知您比较的状态。 下面介绍了这些图标。 如果在执行比较之后，某些数据关键字值没有匹配的地图关键字值，则这些数据关键字值显示在**不匹配的数据关键字**列表中。 在**不匹配的地图关键字**列表中，您还可以看到哪些地图关键字值没有匹配的数据键值。 如果未选中**在可视化中显示所有地图特征**，那么不会呈现由这些地图关键字值标识的特征。

表 *35:* 比较图标

图标	描述
	未执行比较。 这是您在单击**比较**前的缺省状态。 由于您不知道数据关键字和地图关键值是否匹配，因此您需要继续小心操作。
	已执行比较，数据关键字和地图关键字值完全匹配。对于每个数据关键字值，都具有由地图关键字标识的匹配特征。
	已执行比较，某些数据关键字和地图关键字值不匹配。 对于某些数据关键字值，不存在由地图关键字标识的匹配特征。 您需要继续小心操作。 如果继续，地图可视化将不会包含所有数据值。
	已执行比较，数据关键字和地图关键字值完全不匹配。 如果继续，将不会呈现任何地图，因此您应当选择其他数据关键字或地图关键字。

可用的内置图形板可视化类型

您可以创建多个不同的直观表示类型。 所有以下内置类型都在基本和详细选项卡上可用。 某些模板说明（尤其是地图模板）通过**特殊文本**来标识在“详细”选项卡上指定的字段（变量）。

表 *36:* 可用图形类型

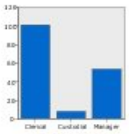
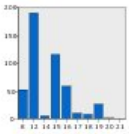
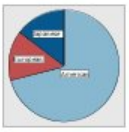
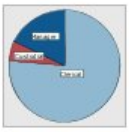
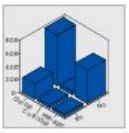
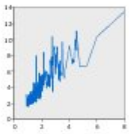
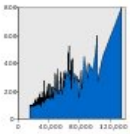
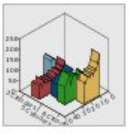
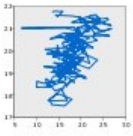
图表图标	描述	图表图标	描述
	**条** 计算连续数值字段的汇总统计，并将分类字段的每个类别结果显示为条形图。 所需项：一个分类字段和一个连续字段。		**计数条形图** 将分类字段的每个类别中行/个案比例显示为条形图。 您还可使用分布图形节点生成此图形。 该节点提供了一些其他选项。 请参阅第 179 页的『分布节点』主题以获取更多信息。 需要：单一分类字段。
	**Pie** 计算连续数字字段的和，并将分布在分类字段的各个类别中的总和比例显示为饼块。 所需项：一个分类字段和一个连续字段。		**计数饼图** 将分类字段的各个类别中的行/观测值比例显示为饼块。 需要：单一分类字段。
	**三维条形图** 计算连续数字字段的汇总统计，并显示两个分类字段之间类别交叉的结果。 需要：一对分类字段和一个连续字段。		**三维饼图** 除附加三维效果之外与饼图相同。 所需项：一个分类字段和一个连续字段。
	**线** 计算一个字段对于另一个字段的每个值的汇总统计，并绘制一条连接值的线。 您还可使用绘图图形节点生成线图图形。 该节点提供了一些其他选项。 请参阅第 169 页的『散点图节点』主题以获取更多信息。 需要：一对任意类型的字段。		**区(A)** 根据各个其他字段值计算字段的汇总统计，并绘制一个区域将这些值连接起来。 线和面之间的区别在于面类似于一条下方带有显示颜色空间的线。 然而，如果您使用颜色外观，这将生成一个线的简单拆分以及面的堆积。 需要：一对任意类型的字段。
	**三维面积图** 显示根据另一个字段的值绘制一个字段的值，并由一个分类字段拆分。 为各个类别绘制分区元素。 所需项：一个分类字段以及一对任意类型的字段。		**Path** 显示根据其他字段的值绘制成的一个字段的值，并按照它们在原始数据集中显示的顺序用线连接。 有顺序之分是路径图和折线图之间的主要区别。 需要：一对任意类型的字段。

表 *36:* 可用图形类型 *(*继续*)*

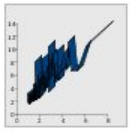
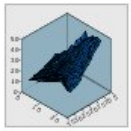
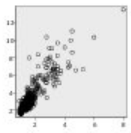
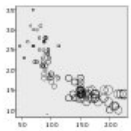
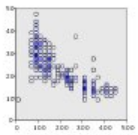
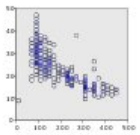
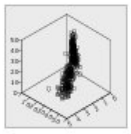
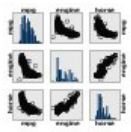
图表图标	描述	图表图标	描述
	**Ribbon** 计算一个字段对于另一个字段的每个值的汇总统计，并绘制一个连接值的 ribbon。 带状图实际上是具有 3-D 效果的折线图。 但不是真正的 3-D 图。 需要：一对任意类型的字段。		**Surface** 显示根据三个字段的值彼此绘制成的字段的值，并以平面连接这些值。 所需项：三个任意类型的字段。
	**散点图** 显示根据另一个字段的值绘制一个字段的值。 此图形能够突出显示字段（如果有）之间的关系。 也可以使用“散点图”节点生成散点图。 该节点提供了一些其他选项。 请参阅第 169 页的『散点图节点』主题以获取更多信息。 需要：一对任意类型的字段。		**气泡图** 与基本散点图类似，显示根据一个字段的值绘制的另一个字段的值。 区别在于第三个字段的值用于改变单个点的大小。 所需项：三个任意类型的字段。
	**分箱化散点图** 与基本散点图类似，显示根据一个字段的值绘制的另一个字段的值。 区别在于，相似的值经过分箱成为组，颜色或大小审美原则用于标明各个分箱中的观测值数量。 需要：一对连续字段。		**六边形分箱化散点图** 请参阅“分组散点图”的说明。区别在于基础分箱的形状不同，即其形状近似于六角形，而不是近似于圆形。 生成的六边形分箱散点图类似于分级散点图。 但是，不同图形各个分级中的值的数目不同，原因在于底层分级的形状不同。 需要：一对连续字段。
	**三维散点图 (3-D Scatterplot)** 显示根据另一个字段的值绘制三个字段的值。 此图形能够突出显示字段（如果有）之间的关系。 您还可使用绘图图形节点生成一个三维散点图。 该节点提供了一些其他选项。 请参阅第 169 页的『散点图节点』主题以获取更多信息。 所需项：三个任意类型的字段。		**散点图矩阵 (SPLOM)** 显示根据各个字段的其他值绘制的一个字段的值。 SPLOM 类似于散点图的表。 SPLOM 也包含了各个字段的直方图。 需要：两个或两个以上连续字段。

表 *36:* 可用图形类型 *(*继续*)*

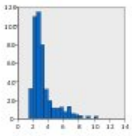
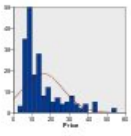
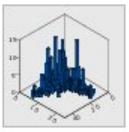
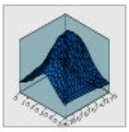
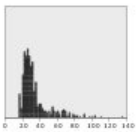
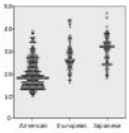
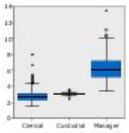
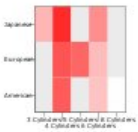
图表图标	描述	图表图标	描述
	**直方图** 显示字段的频率分布。 直方图可以帮您确定分布类型并查看分布是否偏斜。 也可以使用直方图节点生成此图形。 该节点提供了一些其他选项。 请参阅第 182 页的『直方图选项卡』主题以获取更多信息。 需要：单一任意类型字段。		**带有正态分布的直方图** 显示具有正态分布叠加曲线的连续字段的频率分布。 需要：单一连续字段。
	**三维直方图** 显示一对连续字段的频率分布。 需要：一对连续字段。		**三维密度** 显示一对连续字段的频率分布。 这与三维直方图相似，二者唯一的区别使用平面替代条形来显示分布。 需要：一对连续字段。
	**点阵图** 显示各个观测值/行，并将其堆叠在 *x* 轴的不同数据点上。 此图形在显示数据分布上类似于直方图，但是显示每个个案/行，而非特定分箱的汇总计数（值范围）。 需要：单一任意类型字段。		**二维点图** 对于分类字段的每个类别，显示单个个案/行，并将其堆积在 *y* 轴上的不同数据点上。 所需项：一个分类字段和一个连续字段。
	**箱图** 为分类字段的各个类别计算连续字段的五个统计值（最小值、第一四分位数、中位数、第三四分位数和最大值）。 结果显示为箱图/架构元素。 长方形图有助于您查看类别中连续数据分布的变化情况。 所需项：一个分类字段和一个连续字段。		**热图** 计算连续字段的平均值，以交叉两个分类字段之间的类别。 需要：一对分类字段和一个连续字段。

表 *36:* 可用图形类型 *(*继续*)*

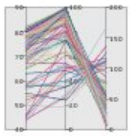
图表图标	描述	图表图标	描述
	**平行** 为各个字段创建平行轴，并绘制一条线来连接数据中各个行/观测值的字段值。 需要：两个或两个以上连续字段。		**计数的等值线图** 计算分类字段（**数据关键字**）每个类别的计数，并绘制地图，其中以颜色饱和度表示地图特征（对应于类别）中的计数。 所需项：一个分类字段。 关键字与**数据键**类别匹配的地图文件。
	**平均值/中位数/和分区图** 计算分类字段（**数据关键字**）每个类别的连续字段（**颜色**）的平均值、中位数或和，并绘制地图，其中以颜色饱和度表示地图特征（对应于类别）中的计算统计。 所需项：一个分类字段和一个连续字段。 关键字与**数据键**类别匹配的地图文件。		**值的等值域图** 绘制地图，其中以颜色表示地图特征（对应于由另一个分类字段（**数据关键字**）定义的值）的分类字段（**颜色**）的值。 如果每个功能部件的“颜色”字段存在多个分类值，那么将使用众数值。 所需项：一对分类字段。 关键字与**数据键**类别匹配的地图文件。
	**计数分区图上的坐标** 这与计数的等值域图相类似，区别在于前者具有两个额外的连续字段（**经度**和**纬度**），用于标识等值域图上绘制点的坐标。 需要：一个分类字段和一对连续字段。 关键字与**数据键**类别匹配的地图文件。		**均值/中位数/总和的等值域图上的坐标** 这类似于平均值/中位数/和分区图，不同之处在于有两个附加连续字段（**经度**和**纬度**），用于确定分区地图上绘制点的坐标。 所需项：一个分类字段和三个连续字段。 关键字与**数据键**类别匹配的地图文件。
	**值的等值域图上的坐标** 这与值的等值域图相类似，区别在于前者具有两个额外的连续字段（**经度**和**纬度**），用于标识等值域图上绘制点的坐标。 所需项：一对分类字段和一对连续字段。 关键字与**数据键**类别匹配的地图文件。		**地图上的计数条形** 为每个地图特征（**数据关键字**）计算分类字段（**类别**）的每个类别中行/个案比例，绘制地图并在每个地图特征中心位置绘制条形图。 所需项：一对分类字段。 关键字与**数据键**类别匹配的地图文件。

表 *36:* 可用图形类型 *(*继续*)*

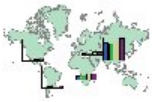
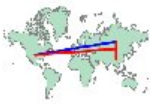
图表图标	描述	图表图标	描述
	**地图上的图条** 为每个地图特征（**数据关键字**）计算连续字段（**值**）的汇总统计，并将分类字段（**类别**）的每个类别结果显示为位于每个地图特征中心位置的条形图。 需要：一对分类字段和一个连续字段。 关键字与**数据键**类别匹配的地图文件。		**地图上的计数饼图** 显示各个地图特征（**数据键**）的分类字段（**类别**）的各个类别中的行/ 观测值比例，并在各个地图特征的中央位置绘制一个由比例分区组成的饼图。 所需项：一对分类字段。 关键字与**数据键**类别匹配的地图文件。
	**地图上的饼图** 为每个地图特征（**数据关键字**）计算分类字段（**类别**）的每个类别中连续字段（**值**）的和，绘制地图并在每个地图特征中心位置将和绘制为饼图分区。 需要：一对分类字段和一个连续字段。 关键字与**数据键**类别匹配的地图文件。		**地图上的线图** 根据各个地图特征（**数据关键字**）的其他字段 (**X**) 的值计算连续字段 (**Y**) 的汇总统计，并在各个地图特征的中央位置绘制一个连接这些值的折线图。 所需项：一个分类字段以及一对任意类型的字段。 关键字与**数据键**类别匹配的地图文件。
	**参考地图上的坐标** 使用连续字段（**经度**）和**纬度**）绘制地图和散点，这些字段用于标识散点的坐标。 需要：一对范围字段。 地图文件。		**参考地图上的箭头** 使用标识各个箭头起始点（**起点经度**和**起点纬度**）和结束点（**结束经度**和**结束纬度**）的连续字段绘制地图和箭头。 数据中的每个记录/观测者都会在地图中生成一个箭头。 需要：四个连续字段。 地图文件。
	**点重叠地图** 绘制参考地图，并在上面重叠另一个点地图，点特征采用分类字段（**颜色**）来着色。 所需项：一对分类字段。 其关键字与**数据关键字**类别匹配的点地图文件。 参考地图文件。		**多边形重叠地图** 绘制参考地图并在其上方放置一个多边形地图，多边形特征由分类字段（**颜色**）进行着色。 所需项：一对分类字段。 关键字与**数据关键字**类别匹配的多边形地图文件。 参考地图文件。

表 *36:* 可用图形类型 *(*继续*)*

图表图标	描述	图表图标	描述
	**线重叠地图** 绘制参考地图，并在上面重叠另一个线地图，线特征采用分类字段（**颜色**）来着色。 所需项：一对分类字段。 其关键字与**数据关键字**类别匹配的线地图文件。 参考地图文件。		

创建地图可视化

对于很多可视化，您只需执行两项选择：感兴趣的字段（变量）和用于显示这些字段的模板。 无需其他选择或操作。 地图可视化至少需要一个附加步骤：选择用于定义地图可视化的地理信息的地图文件。

创建简单地图的基本步骤包括：

1. 在“基本”选项卡上选择感兴趣的字段。 有关各个地图直观表示所需要的字段类型和数目的信息，请参阅第 148 页的『可用的内置图形板可视化类型』。
2. 选择地图模板。
3. 单击“详细”选项卡。
4. 检查**数据关键字**和其他必需下拉列表是否设置为正确字段。
5. 在“地图文件”组中，单击**选择地图文件**。
6. 通过“选择地图”对话框来选择地图文件和地图关键字。 地图关键字值必须与**数据关键字**指定字段的值匹配。 可以使用**比较**按钮来比较这些值。 如果选择了重叠地图模板，则还需要选择参考地图。 参考地图并不通过关键字关联到数据。 它用作主地图的背景。 有关“选择地图”对话框的更多信息，请参阅第 147 页的『选择用于地图可视化的地图文件』。
7. 单击**确定**关闭“选择地图”对话框。
8. 在“图形板模板选择器”中，单击**运行**以创建地图直观表示。

图形板示例

本节包括了几个不同示例，以演示可用选项。 示例同时提供信息用于解释结果产生的可视化。

这些示例使用名为 *graphboard.str* 的流，其引用名为 *employee_data.sav*、*customer_subset.sav* 和 *worldsales.sav* 的数据文件。 这些文件可在任何 IBM SPSS Modeler Client 安装的 *Demos* 文件夹中找到。 此目录可通过 Windows“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。 *graphboard.str* 文件位于 *streams* 文件夹中。

建议您以显示的顺序阅读示例。 后续示例基于先前示例进行构建。

示例：带有汇总统计的条形图

我们将创建条形图，该条形图汇总了集合/分类变量的各个类别的连续数字字段/变量。 特别是，我们将创建显示男性和女性平均工资的条形图。

此示例和以下若干示例使用员工数据，这是包含了公司员工相关信息的虚拟数据集。

1. 添加一个指向 *employee_data.sav* 的 Statistics 文件源节点。
2. 添加“图形板”节点并打开它进行编辑。
3. 在“基本”选项卡上，选择性别和当前工资。 （按住 Ctrl 并单击可选择多个字段/变量。）
4. 选择**条形图**。
5. 从“摘要”下拉列表中选择**均值**。

6. 单击**运行**。
7. 在结果显示中，单击“显示字段和值标签”工具栏按钮（工具栏中心的两组中的第二个）。

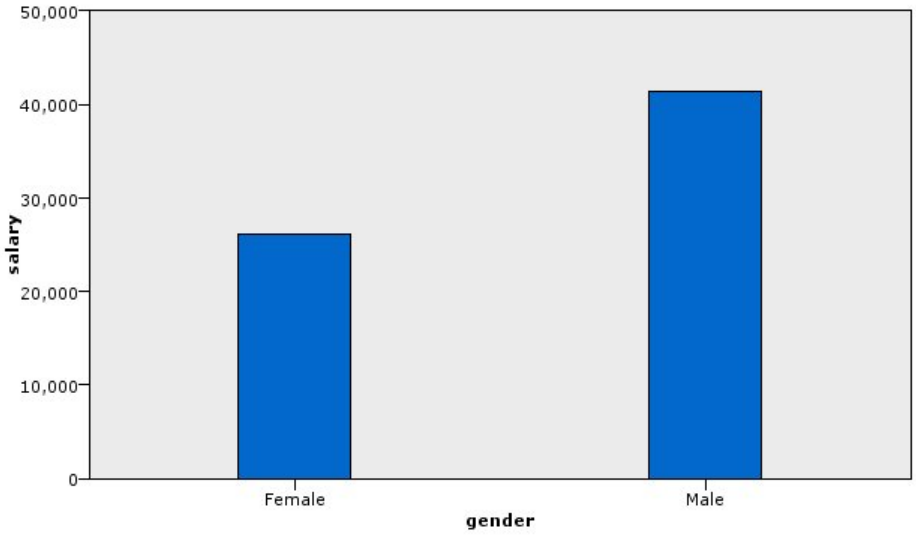


图 *8:* 含汇总统计的条形图

我们可以观察到以下内容：

- 根据条形图的高度，很明显男性的平均薪水高于女性的平均薪水。

示例：含汇总统计的堆积条形图

现在我们将创建堆积条形图，以了解男性与女性的平均工资差异是否取决于工作类型。 可能女性从事某些工作类型的平均人数较男性多。

注: 本示例使用 *Employee data*。

1. 添加“图形板”节点并打开它进行编辑。
2. 在“基本”选项卡上，选择 *Employment Category* 和 *Current Salary*。 （使用 Ctrl+单击可选择多个字段/变量。）
3. 选择**条形**。
4. 从“汇总”列表中，选择 **平均值** 。
5. 单击“详细”选项卡。 请注意，此处反映了您在上一选项卡中所做的选择。
6. 在“可选审美原则”组中，从“颜色”下拉列表选择 性别 。
7. 单击**运行**。

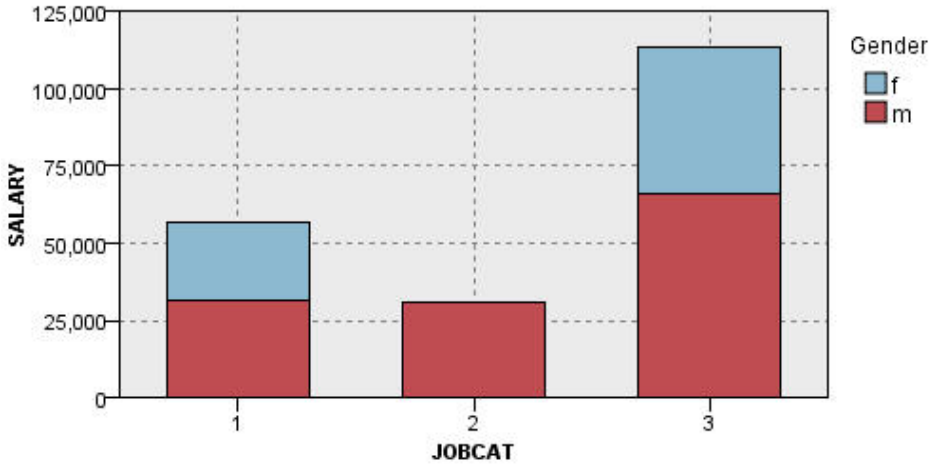


图 *9:* 堆积条形图

我们可以观察到以下内容：

- 与比较男性和女性平均工资的条形图相比，复式条形图中各个工作类型的平均工资之间的差异没有那么明显。 可能各个组中男性和女性的数目不同。 可以通过创建计数条形图来检查这个问题。
- 无论哪个工作类型，男性的平均工资始终要高于女性的平均工资。

示例：带面板的直方图

我们将创建一个按性别排列面板的直方图，以便可以比较男性和女性薪水的频率分布。 频率分布显示有多少个案/行位于特定薪水范围内。 面板直方图可以帮助我们进一步分析性别之间的薪水差异。

注：本示例使用 *Employee data*。

1. 添加一个图形板节点并将其打开用于编辑。
2. 在“基本”选项卡上，选择当前薪水。
3. 选择 **直方图** 。
4. 单击“详细”选项卡。
5. 在“面板和动画”组中，从“面板通过”下拉列表中选择 *gender*。
6. 单击**运行**。

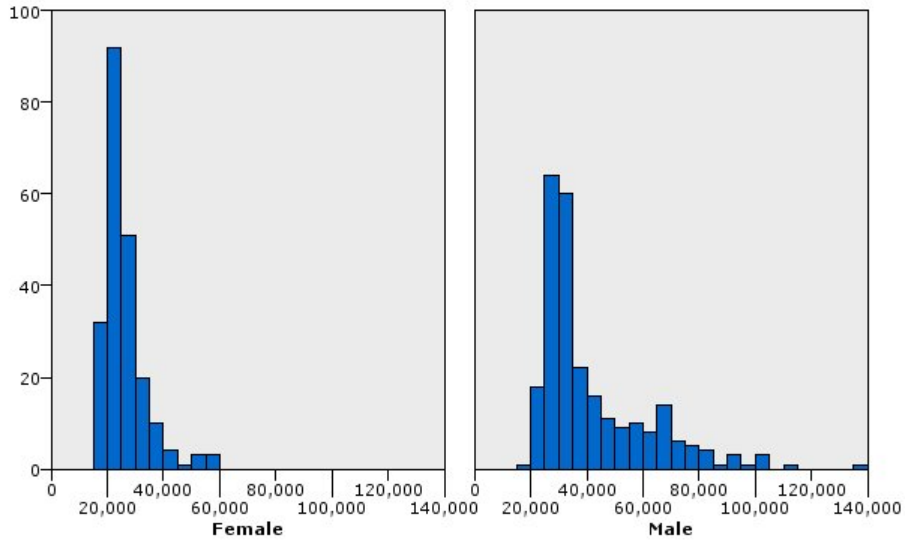


图 *10:* 带面板的直方图

我们可以观察以下内容：

- 两个频率分布都不是正态分布。 即，直方图不与钟型曲线类似，因为只有数据呈正态分布才会与其类似。
- 较高的条形图位于每个图形的左侧。 因此，对于男性和女性，更多的人薪水较低。
- 男性和女性的薪水频率分布不相等。 注意直方图的形状。 薪水较高的男性比薪水较高的女性更多。

示例：带面板点图

与直方图一样，点图显示连续数值范围的分布。 不同之处在于，直方图显示数据分箱范围的计数，而点图显示数据中的各个行/观测值。 因此，与直方图相比，点图提供更多粒度。 事实上，在分析频率分布时，可能开始使用点图更好。

注：本示例使用 *Employee data*。

1. 添加“图形板”节点并打开它进行编辑。
2. 在“基本”选项卡上，选择当前薪水。
3. 选择 **点阵图** 。
4. 单击“详细”选项卡。
5. 在“面板和动画”组中，从“面板通过”下拉列表中选择 *gender*。
6. 单击**运行**。
7. 最大化生成的输出窗口以更清楚地查看图。

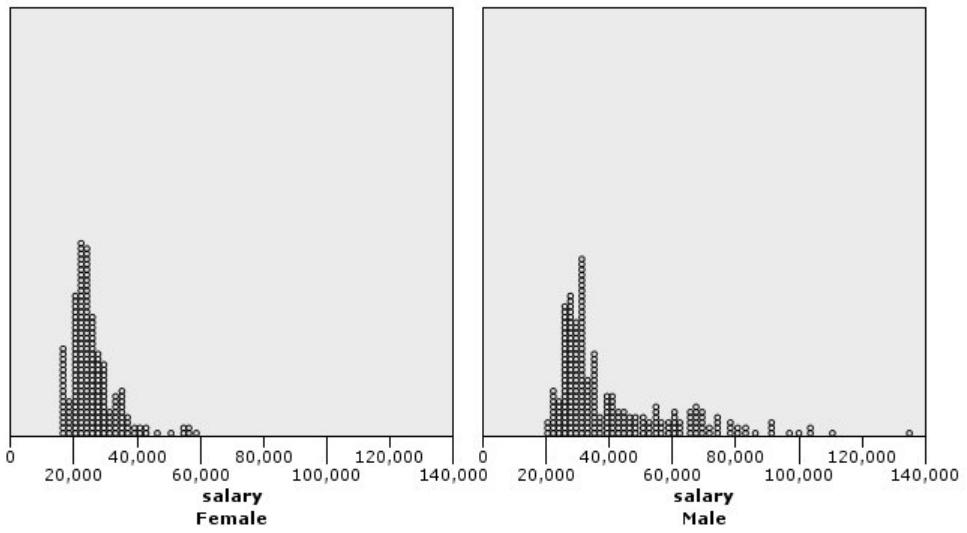


图 *11:* 带面板的点图

通过与直方图进行比较（请参阅第 156 页的『示例：带面板的直方图』），我们可以观察到以下内容：

- 在女性直方图中出现的峰值 20,000 在点图中不太急剧。 有许多案例/行集中在该值周围，但其中大多数值更接近于 25000。 此粒度级别在直方图中不明显。
- 虽然男性的直方图表明男性的平均薪资在 40000 后逐渐下降，但点阵图显示，该值在该值之后分布相当均匀，达到 80000。 在这一范围内，任何一个工资值都有三个或三个以上的男性赚取这一级别薪金。

示例：箱图

箱图是另一个可查看数据分布情况的可视化工具。 箱图包含几个统计测量，我们将在创建可视化后对其进行探索。

注：本示例使用 *Employee data*。

1. 添加“图形板”节点并打开它进行编辑。
2. 在“基本”选项卡上，选择性别和当前工资。 （按住 Ctrl 并单击可选择多个字段/变量。）

3. 选择**箱图**。
4. 单击**运行**。

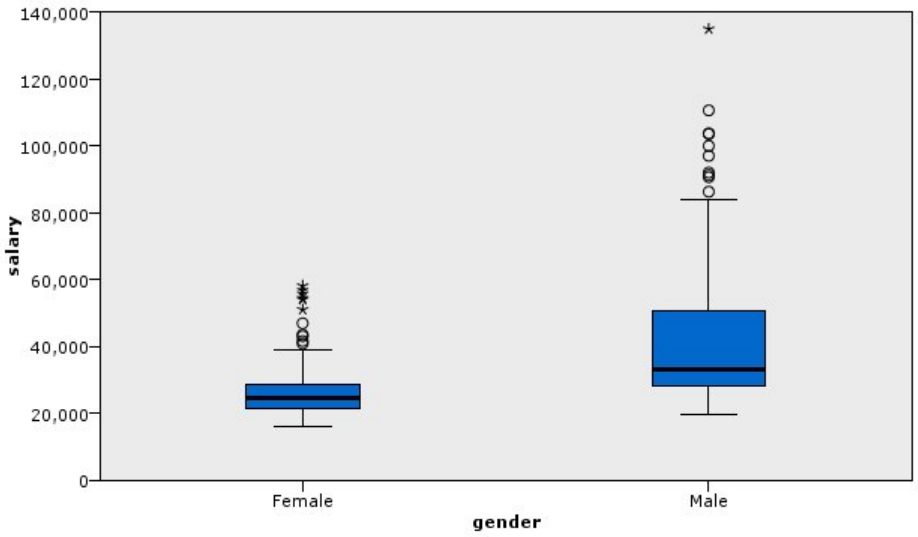


图 *12:* 箱线图

现在我们来探索箱图的各个部分：

- 箱图中间的深色线是 *salary* 的中位数。 一半观测值/行的值大于中位数，另一半值小于中位数。 与均值相同，中位数是集中趋势的测量。 与平均值不同的是，中位数受到值极大的观测值/行的影响较小。 在本例中，中位数小于平均值（与 第 154 页的『示例：带有汇总统计的条形图』 相比）。 平均值与中位数之间的差值表明，有一小部分的观测值/行的值非常大，从而提高了均值。 即，有几个赚取高薪的员工。
- 箱图的底部表示第 25 个百分位。 25% 的个案/行的值低于第 25 个百分位。 箱图的顶部代表第 75 个百分位。 25% 的观测值/行的值大于第 75 个百分位数。 这意味着 50% 的个案/行在箱图内。 女性对应的盒比男性对应的盒短得多。 这表示女性的工资变化小于男性的工资变化。 框的底部和顶部通常称为**折叶点**。
- 从盒内延伸出的 T 条形称为**内限**或**须**。 这些条形会延伸到盒高度的 1.5 倍，如果该范围内没有观测值/行的值，那么会延伸到最小值或最大值。 如果数据呈正态分布，大约 95% 或数据期望在内围之间。 在该示例中，女性对应的内界限延伸长度小于男性对应的延伸长度，这同样也说明了女性的 工资 变化幅度小于男性的工资变化幅度。
- 点是**离群值**。 这些值是指落在内界限外的值。 离群值是极值。 星号或星形表示**极端异常值**。 这些代表拥有超过箱图高度三倍的值的个案/行。 女性和男性都对应有几个离群值。 前面提过平均值比中位数大。 是这些离群值导致了平均值较大。

示例：饼图

我们现在将使用不同的数据集以探索一些其他的可视化类型。 数据集为 *customer_subset*，这是包含客户相关信息的虚拟数据文件。

我们将首先创建一个饼图以查看在不同地理区域中的客户比例。

1. 添加指向 *customer_subset.sav* 的“统计文件”源节点。
2. 添加一个图形板节点并将其打开用于编辑。
3. 在“基本”选项卡上，选择地理指示符。
4. 选择 **计数饼图** 。
5. 单击**运行**。

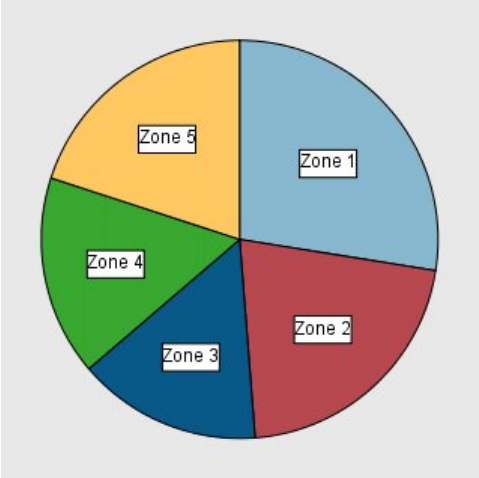


图 *13:* 饼图

我们可以观察以下内容：

- 区域 1 的客户比其他每个区域的客户更多。
- 其他区域的客户均匀分布。

示例：热图

现在，我们将创建分类热图，以查看不同地理区域以及不同年龄组的客户的平均收入。

注：本示例使用 *customer_subset*。

1. 添加“图形板”节点并打开它进行编辑。
2. 在“基本”选项卡中，依次选择地理指示、年龄类别和家庭收入（以千美元为单位）。（按住 Ctrl 并单击可选择多个字段/变量。）
3. 选择**热图**。
4. 单击**运行**。
5. 在结果输出窗口中，单击“显示字段和值标签”工具栏按钮（工具栏中心的两组中的右边按钮）。

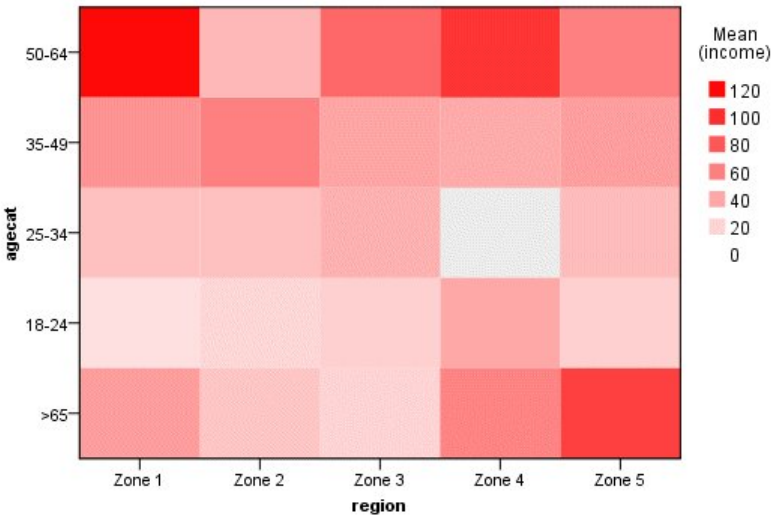


图 *14:* 分类热图

我们可以观察到以下内容：

- 热图就像使用颜色而非数字来代表单元格值的表格。 亮深红色表示较大的值，而灰色则表示较小的值。 每个单元格的值都是各个变量对的连续字段/变量的平均值。

- 除区域 2 和区域 5 外，年龄在 50 到 64 岁之间的客户群体的平均家庭收入比其他群体的家庭收入高。
- 区域 4 中没有年龄在 25 到 34 之间的客户。

示例：散点图矩阵 (SPLOM)

我们将创建一个带有几个不同变量的散点图矩阵，以便我们可以确定数据集中变量之间是否存在任何关系。

注：本示例使用 *customer_subset*。

1. 添加一个图形板节点并将其打开用于编辑。
2. 在“基本”选项卡上，选择年龄（以年计）、家庭收入（以千计）和信用卡债务（以千计）。（按住 Ctrl 并单击可选择多个字段/变量。）
3. 选择 **SPLOM**。
4. 单击**运行**。
5. 最大化输出窗口以更清楚地查看矩阵。

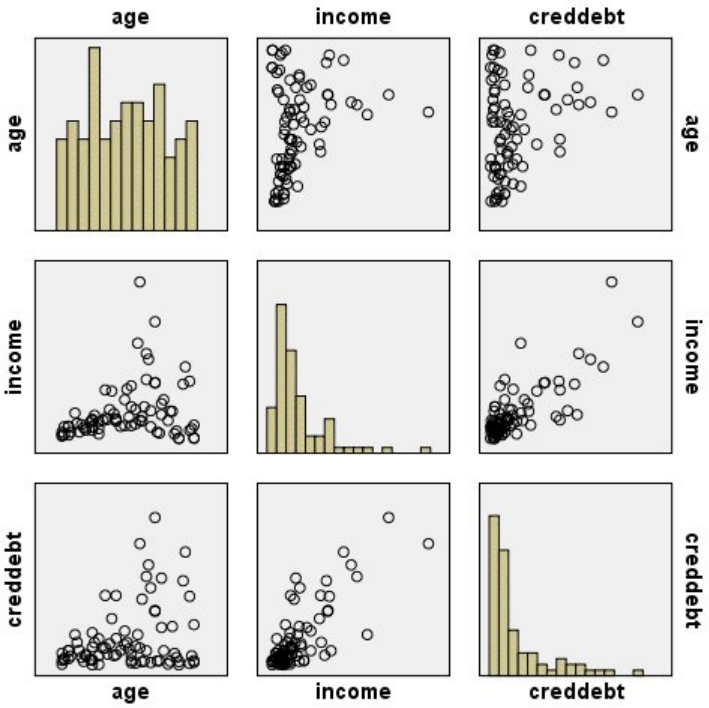


图 *15:* 散点图矩阵 *(SPLOM)*

我们可以观察以下内容：

- 对角线上显示的直方图显示每个变量在 SPLOM 中的分布。 *age* 的直方图显示在左上单元格中，*income* 的直方图显示在中间单元格中，*creddebt* 的直方图则显示在右下单元格中。 这些变量都不是正态分布。即，没有直方图与钟型曲线类似。 还请注意，*income* 和 *creddebt* 的直方图正偏斜。
- *age* 和其他变量看起来没有任何关系。
- *income* 和 *creddebt* 之间存在线性关系。 即，*creddebt* 随着 *income* 的增加而增加。 您可能想要创建这些变量以及其他相关变量的单个散点图以进一步探索关系。

示例：和分区图（着色地图）

现在，我们将创建一个地图直观表示图形。 然后，在随后的实例中，我们将创建此可视化图形的变体。 数据集是 *worldsales*，这是一个包含按不同大洲和产品列出的销售收入的假设数据文件。

1. 添加“图形板”节点并打开它进行编辑。
2. 在“基本”选项卡上，选择洲和收入。（按住 Ctrl 并单击可选择多个字段/变量。）

3. 选择**总和的等值线图**。
4. 单击“详细”选项卡。
5. 在“可选外观”组中，从“数据标签”下拉列表中选择洲。
6. 在“地图文件”组中，单击**选择地图文件**。
7. 在“选择地图”对话框中，检查**地图**是否设置为洲，并且**地图关键字**是否设置为 *CONTINENT*。
8. 在“比较地图”和“数据值”组中，单击**比较**以确保地图关键字与数据键匹配。 在此示例中，所有数据关键字值已与地图关键字和特征匹配。 我们还可看到这里不存在大洋洲的数据。
9. 在“选择地图”对话框中，单击**确定**。
10. 单击**运行**。

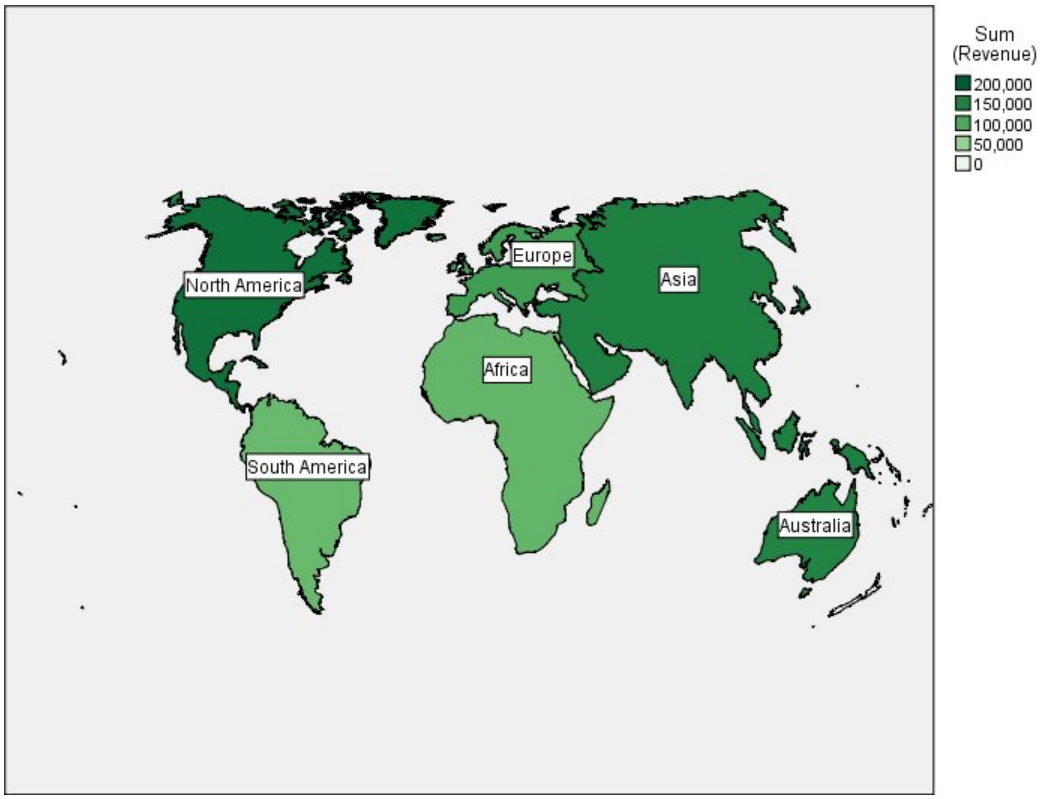


图 *16:* 总和的等值线图

从该地图可视化中，我们很容易看到北美洲的收入最高；南美洲和非洲的收入最低。 每个大洲均已标出，因为我们使用大洲作为数据标签外观。

示例：地图上的条形图

本例显示在每个大洲中销售收入按产品分类的情况。

注：本例使用 *worldsales*。

1. 添加一个图形板节点并将其打开用于编辑。
2. 在“基本”选项卡上，选择洲、产品和收入。 （按住 Ctrl 并单击可选择多个字段/变量。）
3. 选择**地图的图条**。
4. 单击“详细”选项卡。

   如果使用特定类型的多个字段，则必须检查以确认每个字段被分配到正确条形内。
5. 从“类别”下拉列表中选择产品。
6. 从“值”下拉列表中选择收入。
7. 从“数据键”下拉列表中选择洲。

8. 从“摘要”下拉列表中，选择和。
9. 在“地图文件”组中，单击**选择地图文件**。
10. 在“选择地图”对话框中，检查**地图**是否设置为洲，并且**地图关键字**是否设置为 *CONTINENT*。
11. 在“比较地图”和“数据值”组中，单击**比较**以确保地图关键字与数据键匹配。 在本例中，所有数据关键字值均具有匹配的地图关键字和特征。 我们还可看到这里不存在大洋洲的数据。
12. 在“选择地图”对话框中，单击**确定**。
13. 单击**运行**。
14. 最大化结果输出窗口以更清楚地查看显示。

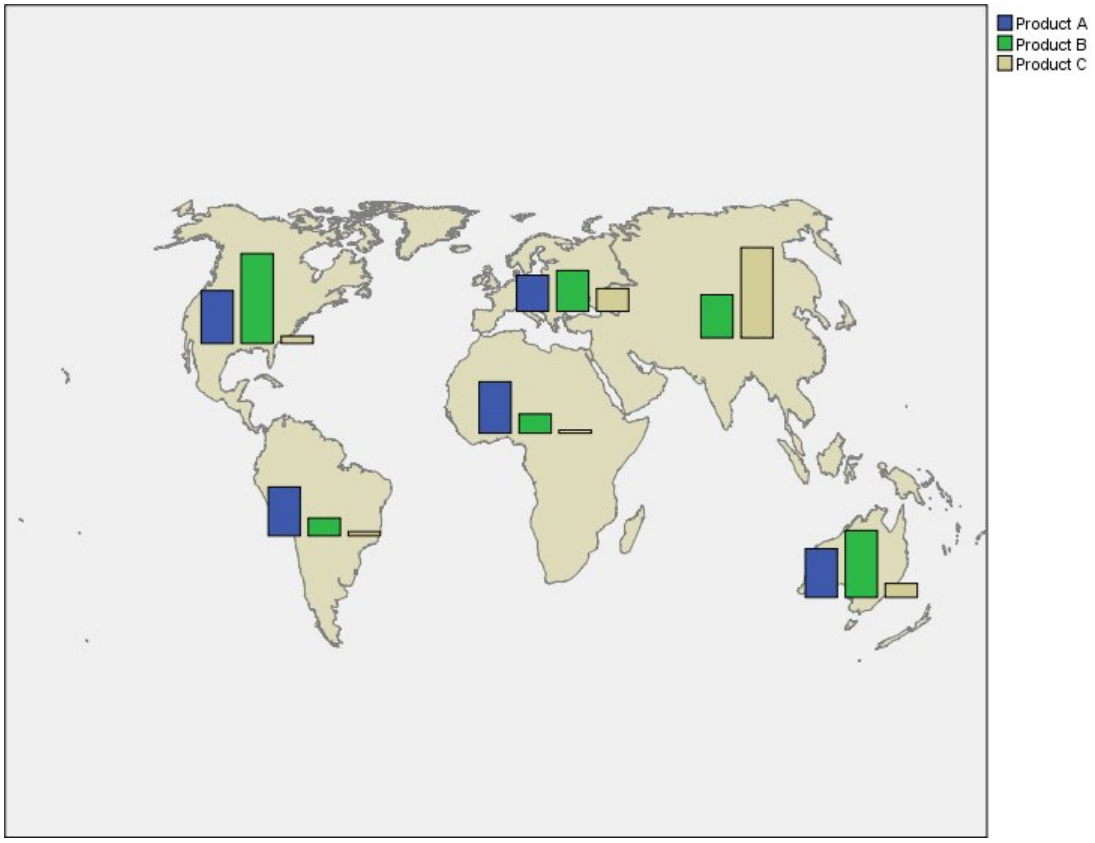


图 *17:* 地图上的条形图

我们可以观察以下内容:

- 在南美洲和非洲中不同产品的总收入分布非常相似。
- 产品 *C* 在除亚洲以外的任何地方都产生了最低收入。
- *Product A* 在亚洲中没有收入或收入最低。

图形板“外观”选项卡

可以在创建图形前指定外观选项。

一般外观选项

标题。 输入要用作图形标题的文本。

子标题。 输入要用作图形子标题的文本。

文字说明。 输入要用作图形文字说明的文本。

采样。 为较大数据集指定一种方法。 可以指定最大数据集大小，或使用缺省记录数。 如果选择**抽样**选项，那么处理大数据集时的性能将会提高。 另外，您也可以选择 **使用所有数据**，但必须要注意，这一选项可能大幅降低软件的执行效率。

样式表外观选项

有两个按钮可用于控制哪个可视化模板（以及样式表和图）可用：

管理。 管理计算机上的可视化模板，及样式表与地图。 您可以导入、导出、重命名和删除本地计算机上的可视化模板，样式表与地图。 有关更多信息，请参阅第 164 页的『管理模板、样式表和地图文件』主题。

位置。 更改可视化模板，样式表与地图的存储位置。 当前位置列出在按钮右侧。 有关更多信息，请参阅第 163 页的『设置模板、样式表和地图位置』主题。

以下示例显示外观选项在图形上的位置。 （请注意，并非所有图形都使用所有这些选项。）

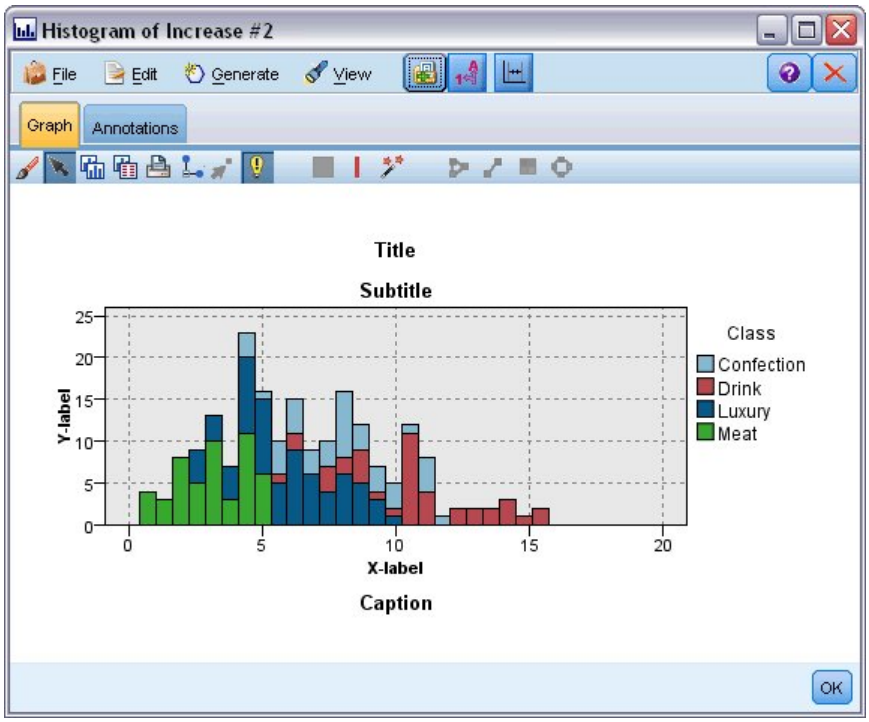


图 *18:* 各个图形外观选项的位置

设置模板、样式表和地图位置

直观表示模板、直观表示样式表和地图文件保存在特定本地文件夹中或保存在 IBM SPSS Collaboration and Deployment Services Repository 中。 选择模板、样式表和地图时，只会显示在此位置中内置的模板、样式表和地图。 通过将所有模板、样式表和地图文件保存在一个位置，IBM SPSS 应用程序可方便地访问它们。 有关添加更多模板、样式表和地图文件到此位置的信息，请参阅 第 164 页的『管理模板、样式表和地图文件』。

如何设置模板、样式表和地图文件的位置

1. 在模板或样式表对话框中，单击**位置...**以显示“模板、样式表和地图”对话框。
2. 选择模板、样式表和地图文件的缺省位置选项：

   本地计算机。 模板、样式表和地图文件位于本地计算机上的特定文件夹中。 在 Windows XP 上，此文件夹为 *C:\Documents and Settings\<user>\Application Data\SPSSInc\Graphboard*。 无法更改此文件夹。

   IBM SPSS Collaboration and Deployment Services Repository。 模板、样式表和地图文件位于 IBM SPSS Collaboration and Deployment Services Repository 中由用户指定的文件夹中。 要找到该特定文件夹，单击**文件夹**。 有关更多信息，请参阅第 164 页的『将 IBM SPSS Collaboration and Deployment Services Repository 用作模板、样式表和地图文件位置』。

3. 单击**确定**。

将 IBM SPSS Collaboration and Deployment Services Repository 用作模板、样式表和地图文件位置

直观表示模板和样式表可保存在 IBM SPSS Collaboration and Deployment Services Repository 中。 此位置是 IBM SPSS Collaboration and Deployment Services Repository 中的特定文件夹。 如果将其设为缺省位置，那么此位置中的任何模板、样式表和地图文件都可用于选择。

如何将 IBM SPSS Collaboration and Deployment Services Repository 中的文件夹设为模板、样式表和地图文件的位置

1. 在带有位置按钮的对话框中，单击**位置...**。
2. 选择 IBM SPSS Collaboration and Deployment Services Repository。
3. 单击**文件夹**。

   注：如果您尚未连接到 IBM SPSS Collaboration and Deployment Services Repository，会提示您输入连接信息。
4. 在“选择文件夹”对话框中，选择保存模板、样式表和地图文件的文件夹。
5. （可选）可以从**检索标签**中选择一个标签。 将只显示具有该标签的模板、样式表和地图文件。
6. 如果您正在查找包含某个特殊模板或样式表的文件夹，您可能希望在“搜索”选项卡上搜索模板、样式表或地图文件。 “选择文件夹”对话框会自动选择找到的模板、样式表或地图文件所在的文件夹。
7. 单击**选择文件夹**。

管理模板、样式表和地图文件

您可以使用“管理模板、样式表和地图”对话框管理您的计算机上本地位置中的模板、样式表和地图文件。 此对话框允许您导入、导出、重命名和删除您的计算机上本地位置中的直观表示模板、样式表和地图文件。

在您选择模板、样式表或地图的某个对话框中单击**管理...** 。

管理模板、样式表和地图对话框

“模板”选项卡列出所有本地模板。 “样式表”选项卡列出所有本地样式表，并显示具有样本数据的示例直观表示。 您可以选择其中一个样式表来将其样式应用到示例直观表示。 请参阅相关主题第 225 页的『应用样式表』以了解更多信息。 “地图”选项卡列出所有本地地图文件。 此选项卡还显示地图关键字（包括示例值）、注释（如在创建地图时有提供）和地图预览。

以下按钮位于当前启用的任何选项卡上。

导入。 从文件系统导入直观表示模板、样式表或地图文件。 导入模板、样式表或地图文件供 IBM SPSS 应用程序使用。 如果另一个用户发送给您一个模板、样式表或地图文件，您可以导入该文件，然后在您的应用程序中使用。

导出。 将直观表示模板、样式表或地图文件导出至文件系统。 当您希望将模板、样式表或地图文件发送给另一个用户时，可将其导出。

重命名。 重命名所选直观表示模板、样式表或地图文件。 您不能将名称改为正在使用的名称。

导出地图关键字。 将地图关键字作为以逗号分隔的值 (CSV) 文件导出。 此按钮仅在“地图”选项卡上启用。

删除。 删除所选直观表示模板、样式表或地图文件。 您可以通过在 Windows 和 Linux 上按住 Ctrl 并单击来选择多个模板、样式表或地图文件。 删除操作无法撤销，所以请小心操作。

转换和分发地图 Shapefile

图形板模板选择器允许您从可视化模板和 SMZ 文件的组合创建地图可视化。 SMZ 文件与 ESRI shapefile（SHP 文件格式）在这一点非常相似：它们包含用于绘制地图的地理信息（例如，国家或地区边界），但针对地图直观表示进行了优化。 图形板模板选择器预先安装有一定数量的 SMZ 文件。 如果您要将现有 ESRI shapefile 用于地图可视化，首先需要通过地图转换实用程序将此 shapefile 转换为 SMZ 文件。 地图转换实用程序支持包含单层的点、多义线或多边形（形状类型 1、3 和 5）ESRI shapefile。

除了转换 ESRI shapefile 外，地图转换实用程序还允许您修改地图的细节层次、更改特征标签、合并特征和移动特征，以及执行其他可选更改。 还可以使用地图转换实用程序来修改现有的 SMZ 文件（包括预先安装的文件）。

编辑预先安装的 SMZ 文件

1. 从管理系统中导出 SMZ 文件。 请参阅主题第 164 页的『管理模板、样式表和地图文件』，了解更多信息。
2. 使用地图转换实用程序可打开并编辑导出的 SMZ 文件。 建议您使用其他名称来保存此文件。 请参阅主题第 165 页的『使用地图转换实用程序』，了解更多信息。
3. 将修改后的 SMZ 文件导入管理系统。 请参阅主题第 164 页的『管理模板、样式表和地图文件』，了解更多信息。

地图文件的其他资源

多个专用和公共源中都提供了 SHP 文件格式的地理空间数据，此数据可用于支持绘制地图需要。 如果要查找免费数据，请访问当地政府 Web 站点。 本产品中的多个模板都基于从 GeoCommons (http://www.geocommons.com) 和 美国人口调查局 (http://www.census.gov) 获取的公开可用数据。

重要声明：非 IBM 产品的相关信息来自这些产品的供应商，及其发布的公告或其他公开来源。 IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。 有关非 IBM 产品性能的问 题应当向这些产品的供应商提出。 本信息中对任何非 IBM Web 站点的引用都只是为了方便起见才提供的， 不以任何方式充当对那些 Web 站点的保证。 那些 Web 站点中的资料不是本 IBM 程度资料的一部分，除非在此 IBM 程序随附的声明文件中另有声明，否则使用那些资料站点带来的风险将由您自行承担。

有关地图的重要概念

通过了解有关 shapefile 的一些重要概念，能够帮助您有效地使用地图转换实用程序。

shapefile 提供用于绘制地图的地理信息。 地图转换实用程序支持三种类型的 shapefile：

- **点**。 该 shapefile 标识点的位置，例如城市。
- **多义线**。 该 shapefile 标识路径及其位置，例如河流。
- **多边形**。 该 shapefile 标识带有边界的区域及其位置，例如国家。

最常用的是多边形 shapefile。 从多边形 shapefile 可以创建分区着色地图。 分区着色地图使用颜色来代表不同多边形（区域）内的值。 点和多义线 shapefile 通常重叠在多边形 shapefile 上。 例如，美国各城市点 shapefile 重叠在美国各州多边形 shapefile 上。

shapefile 由**特征**构成。 特征是独立的地理实体。 例如，特征可以是国家、州、城市等等。 shapefile 还包含有关特征的数据。 这些数据存储在**属性**中。 属性类似于数据文件中的字段或变量。 至少有一个属性为特征的**地图关键字**。 地图关键字可以是标签，例如国家或州名。 地图关键字用于链接到数据文件中的变量/字段，以便创建地图可视化。

注意，您只能在 SMZ 文件中保留一个或多个关键属性。 地图转换实用程序不支持保存其他属性。 这意味着，如果您要在不同层次上汇总，则需要创建多个 SMZ 文件。 例如，如果您要汇总美国各州和地区，那么将需要两个不同的 SMZ 文件：这两个文件分别具有标识州和地区的关键字。

使用地图转换实用程序

如何启动地图转换实用程序

从菜单中选择：

工具 > 地图转换实用程序

在地图转换实用程序中有四个主要屏幕（步骤）。 它们还分别包含相应的子步骤，以便详细控制对地图文件的编辑操作。

第 1 步 - 选择目标和源文件

您首先需要选择源地图文件和转换后地图文件的目标位置。 对于 shapefile，您将需要 *.shp* 和 *.dbf* 文件。

选择要用于转换的 .shp (ESRI) 或 .smz 文件。 浏览到计算机上的现有地图文件。 这是您要将其转换并保存为 SMZ 文件的文件。 shapefile 的 *.dbf* 文件必须存储在与 *.shp* 文件相同的位置，并且二者的基本文件名应相同。 需要 *.dbf* 文件，因为它包含 *.shp* 文件的属性信息。

为转换后地图文件设置目标位置和文件名。 为将要从原始地图源创建的 SMZ 文件输入路径和文件名。

- **导入模板选择器**。 除了在文件系统中保存文件外，您还可以将地图添加到模板选择器的管理列表中。 如果选择此选项，地图将自动出现在您的计算机上安装的 IBM SPSS 产品的模板选择器中。 如果您不立即导入模板选择器，则需要以后手动导入它。 有关将地图导入至模板选择器管理系统的更多信息，请参阅第 164 页的『管理模板、样式表和地图文件』。

步骤 2 - 选择地图关键字

现在，您将选择要在 SMZ 文件中包括哪些地图关键字。 接着，您可以更改某些将影响地图呈现的选项。 在地图转换实用程序的后续步骤中包含预览地图。 您选择的呈现选项将用于生成地图预览。

选择主要地图关键字。 选择作为主要关键字的属性，此属性用于标识和标注地图中的特征。 例如，世界地图的主要关键字可以是标识国家或地区名称的属性。 主要关键字还将数据链接到地图特征，因此，请确保您所选择的属性的值（标签）将与数据中的值匹配。 选择属性后将显示示例标签。 如果需要更改这些标签，您可以在稍后的步骤中进行。

选择要包括的其他关键字。 除主要地图关键字外，请检查任何其他要包括在生成的 SMZ 文件中的关键字属性。 例如，某些属性可能包含翻译后的标签。 如果您希望使用其他语言编码的数据，则可能需要保留这些属性。 注意，您只能选择那些与主要关键字表示同一特征的其他关键字。例如，如果主键是美国州的全名，那么只能选择表示美国州的备用键， 例如，州缩写。

自动平滑化地图。 带有多边形的 Shapefile 通常包含太多数据点和太多统计地图直观表示的详细信息。 过量的详细信息可能会分散并负面影响性能。 您可以降低详细信息的级别并通过平滑功能来推广地图。 结果是，地图的外观将更加轮廓鲜明并且呈现速度越快。 地图自动平滑化时，最大角度为 15 度，而要保留的百分比为 99。 有关这些设置的信息，请参阅第 166 页的『使地图平滑』。 注意，您在另一个步骤中还有机会应用额外的平滑化。

除去同一特征中接触多边形之间的边界。 某些特征可能包含在感兴趣主特征内部具有边界的子特征。 例如，世界大洲地图可能包含每个大洲内所包含的国家或地区的内部边界。 如果选择此选项，那么地图中将不会显示内部边界。 在世界大洲地图示例中，选择此选项将移除国家或地区边界，同时保留大洲边界。

第 3 步 - 编辑地图

您已指定地图的基本选项，现在可以编辑更多特定选项。 这些修改是可选的。 地图转换实用程序的此步骤将指导您执行关联任务，并且显示了地图的预览，以便您能够验证更改。 某些任务可能不可用，具体取决于 shapefile 类型（点、这些或多边形）和坐标系。

每项任务在地图转换实用程序的左侧具有以下公共控件。

显示地图上的标签。 缺省情况下，不会在预览中显示特征标签。 您可以选择显示这些标签。 尽管标签有助于标识特征，但它们可能会干扰预览地图上的直接选择。 请在需要时（例如，编辑特征标签时）启用此选项。

对地图预览着色。 缺省情况下，预览地图将以一种纯色显示各个分区。 所有特征具有相同颜色。 您可以选择为每个地图特征指定相应的颜色。 此选项可能有助于区分地图中的不同特征。 合并特征并且希望查看新特征在预览中的表示方法时，此选项尤其有用。

每个任务在地图转换实用程序的右侧还具有以下公共控件。

撤销。 单击**撤销**可还原至上一状态。 最多可以撤销 100 次更改。

使地图平滑

带有多边形的 Shapefile 通常包含太多数据点和太多统计地图直观表示的详细信息。 过量的详细信息可能会分散并负面影响性能。 您可以降低详细信息的级别并通过平滑功能来推广地图。 结果是，地图的外观将更加轮廓鲜明并且呈现速度越快。 此选项对于点和多义线地图不可用。

最大角度。 最大角度（必须是 1 到 20 之间的值）指定趋于线性的点集的平滑容许误差。 此值越大，线性平滑的容许误差就越大，并且随后将删除多个点，从而产生更加一般化的地图。 要应用线性平滑，地图转换实

用程序将检查地图中每个三点集合形成的内角。 如果 180 减去此角度所得的值小于指定值，那么地图转换实用程序将删除中间点。 换而言之，地图转换实用程序检查的是这三个点形成的线是否几乎为直线。 如果是，地图转换实用程序会将两个端点之间处理为直线，并丢弃中间点。

保留百分比。 要保留的百分比，它必须是 90 到 100 之间的值，确定对地图进行平滑化时要保留的陆地面积量。 此选项仅影响那些具有多个多边形的特征，也就是包括岛屿的特征。 如果特征的总面积减去某个多边形所得的值大于原始面积的指定百分比，那么地图转换实用程序将从地图中删除此多边形。 地图转换实用程序绝不除去特征的所有多边形。 也就是说，不论应用多大的平滑化量，特征至少有一个多边形。

在您选择最大角度和保留百分比后，单击**应用**。 预览将通过平滑更改来更新。 如果需要再次对地图执行平滑，请重复此操作，直到达到所需平滑度级别。 注意，平滑存在限制。 如果您重复平滑化，您将达到某个无法再对地图应用平滑化的位置。

编辑特征标签

您可以根据需要（例如，为了与预期数据匹配）编辑特征标签，并且重新将标签放在地图中。 在创建地图直观表示之前，您应该查看标签，即使您不认为需要更改这些标签也是如此。 由于缺省情况下在预览中不显示标签，因此您可能需要选择**显示地图上的标签**以显示它们。

关键字。 选择包含您要检查和/或编辑的特征标签的关键字。

特征。 此列表显示所选关键字中包含的特征标签。 要编辑标签，请在列表中双击它。 如果在地图上显示了标签，还可以直接在地图预览中双击特征标签。 如果要将标签与实际数据文件进行比较，则单击**比较**。

X/Y。 这些文本框列出地图上所选特征标签的当前中心点。 单位显示在地图的坐标中。 它们可以是局部笛卡尔坐标（例如，美国国家平面坐标系）或地理坐标（其中 **X** 为经度，**Y** 为纬度）。 输入标签新位置的坐标。 如果标签已显示，还可以在地图上单击并拖动标签。 文本框将通过新位置进行更新。

比较。 如果某个数据文件包含需要与特定关键字的特征标签匹配的数据值，则单击**比较**以显示“与外部数据源比较”对话框。 在此对话框中，您可以打开数据文件，并将其值直接与地图关键字的特征标签值进行比较。

“与外部数据源比较”对话框

The Compare to an External Data Source dialog box allows you to open a tab-separated values file (with a *.txt* extension), a comma-separated values file (with a *.csv* extension), or a data file formatted for IBM SPSS Statistics (with a *.sav* extension). 在文件打开后，您可以从数据文件中选择字段，并与特定地图关键字中的特征标签进行比较。 然后，您可以纠正地图文件中的任何偏差。

数据文件中的字段。 选择您要将其值与特征标签进行比较的字段。 如果 *.txt* 或 *.csv* 文件的第一行包含每个字段的描述性标签，则选中**使用第一行作为列标签**。 否则，每个字段将按其在数据文件中的位置进行标识（例如，"列 1"、"列 2"，依此类推）。

要比较的关键字。 选择您要将其特征标签与数据文件字段值进行比较的地图关键字。

比较。 在准备好比较值后单击此按钮。

比较结果。 缺省情况下，?比较结果?表仅列出在数据文件中不匹配的字段值。 该应用程序通常通过检查插入或缺失的空格来查找相关的特征标签。 单击地图标签列中的下拉列表，以将地图文件中的特征标签与显示的字段值匹配。 如果地图文件中没有对应的特征标签，请选择保持不匹配。 如果您要查看所有字段值，包括那些已匹配特征标签的字段值，请取消选中**仅显示不匹配个案**。 如果要覆盖一个或多个匹配，您可做此操作。

在您将特征匹配到字段值时，每个特征只能使用一次。 如果您要将多个特征匹配到单个字段值，则可以合并特征，然后将合并后的新特征匹配到字段值。 有关合并特征的更多信息，请参阅第 167 页的『合并特征』。

合并特征

在地图中创建更大的区域时，合并特征非常有用。 例如，如果要转换州地图，那么可以将州（本示例中为特征）合并到更大的北部、南部、东部和西部地区。

密钥。 选择包含特征标签的地图关键字，这些标签将有助于标识要合并的特征。

特征。 单击要合并的第一个特征。 按住 Ctrl 并单击其他要合并的特征。 注意，还会在地图预览中选中这些特征。 除从列表中选择特征外，还可以在地图映射中直接单击以及按住 Ctrl 单击特征。

选择要合并的特征之后，请单击**合并**以显示“命名合并后的特征”对话框，您可以在此对话框中将标签应用于新特征。 合并特征之后，您可能希望选中**对地图预览进行着色**以确保结果是您所期望的结果。

合并特征之后，您可能还希望移动新特征的标签。 您可以在编辑特征标签任务中完成此操作。 有关更多信息，请参阅主题 第 167 页的『编辑特征标签』。

“命名合并后的特征”对话框

“命名合并后的特征”对话框允许您为合并后的新特征指定标签。

“标签”表显示地图文件中每个关键字的信息，并允许您为每个关键字指定标签。

新标签。 为合并后的特征输入新标签，以指定给特定地图关键字。

关键字。 您要为其指定新标签的地图关键字。

旧标签。 将被合并为新特征的特征的标签。

清除靠接多边形的边界。 选中此选项以从已合并的特征中清除边界。 例如，您将州合并为地理区域，此选项将清除各个州之间的边界。

移动特征

您可以在地图中移动特征。 在您要将多个特征放在一起时（例如，大陆和边远小岛），这非常有用。

关键字。 选择包含有助于您确定要移动的特征的特征标签的地图关键字。

特征。 单击您要移动的特征。 注意，特征将在地图预览中被选中。 您还可以直接在地图预览中单击特征。

X/Y。 这些文本框列出地图上特征的当前中心点。 单位显示在地图的坐标中。 它们可以是局部笛卡尔坐标（例如，美国国家平面坐标系）或地理坐标（其中 **X** 为经度，**Y** 为纬度）。 输入特征新位置的坐标。 还可以在地图上单击并拖动特征。 文本框将更新为新的位置。

删除特征

可以删除地图中不需要的特征。 通过删除地图直观表示中不需要的特征以使地图更加简洁时，此选项非常有用。

密钥。 选择包含特征标签的地图关键字，这些标签将有助于标识要删除的特征。

特征。 单击要删除的特征。 如果要同时删除多个特征，请按住 Ctrl 并同时单击其他特征。 注意，还会在地图预览中选中这些特征。 除从列表中选择特征外，还可以在地图映射中直接单击以及按住 Ctrl 单击特征。

删除个别元素

除删除整个特征外，还可以删除组成特征的某些个别元素，例如湖和岛屿。 此选项对于点地图不适用。

元素数目。 单击要删除的元素。 如果要同时删除多个元素，请按住 Ctrl 并同时单击其他元素。 注意，还会在地图预览中选中这些元素。 除从列表中选择元素外，还可以在地图映射中直接单击以及按住 Ctrl 单击元素。 由于元素名称的列表不是描述性列表（每个元素在特征中都分配了一个编号），因此，您应该在地图预览中确认选择，以确保选中了所需元素。

设置投影

地图投影指定三维地球在二维中的表示方式。 所有投影都会产生失真。 不过，根据您在查看全球地图还是局部地图，某些投影可能更为适合。 另外，某些投影保留了原始特征的形状。 保留了形状的投影称为正形投影。 此选项仅对于带有地理坐标（经度和纬度）的地图适用。

与地图转换实用程序中的其他选项不同，您可以在创建地图可视化之后更改投影。

投影。 选择地图投影。 如果您要创建全球或半球地图，请使用局部、*Mercator* 或 *Winkel Tripel* 投影。 对于较小区域，则使用局部、*Lambert* 正形圆锥投影或横轴 *Mercator* 投影。 所有投影都对数据使用 WGS83 ellipsoid。

- 在通过局部坐标系（例如，美国国家平面坐标系）创建地图时，始终使用**局部**投影。 这些坐标系由笛卡儿坐标而不是地理坐标（经度和纬度）定义。 在局部投影中，水平线和垂直线在笛卡儿坐标系中等距排列。 局部投影不是正形投影。

- **Mercator** 投影是用于全球地图的正形投影。 水平线和垂直线是直线并且相互垂直。 注意：墨卡托投影在接近北极和南极时无限延伸，因此，如果地图包括北极或南极，那么将不能使用此投影。 地图接近这些限制时，失真程度最大。
- **Winkel Tripel** 投影是用于全球地图的非正形投影。 尽管它不是正形投影，但它在形状和大小之间提供了良好的平衡。 除了赤道和本初子午线外，所有线条均为曲线。 如果您的全球地图包括北极或南极，这是很好的投影选择。
- 顾名思义，**Lambert 正形圆锥**投影为正形投影，它用于东西方向比南北方向更长的大陆或较小大陆块地图。
- **横轴 Mercator** 是另一种适合大陆或较小大陆块地图的正形投影。 该投影用于南北方向比东西方向更长的大陆块地图。

第 4 步 - 完成

此时，您可以添加注释来描述地图文件，并根据地图关键字创建一个样本数据文件。

地图关键字。 如果在地图文件中存在多个关键字，则选择您要在预览中显示其特征标签的地图关键字。 如果根据地图创建了数据文件，那么这些标签将用于数据值。

注释。 输入描述地图的注释，或者提供可能与用户相关的其他信息，例如原始 shapefile 的源。 该注释将出现在图形板模板选择器的管理系统中。

从特征标签创建数据集。 如果您要从显示的特征标签创建数据文件，选中此选项。 单击**浏览...** 后，您将可以指定位置和文件名。 如果添加了 *.txt* 扩展名，则文件将保存为制表符分隔值文件。 如果添加了 *.csv* 扩展名，则文件将保存为逗号分隔值文件。 如果您添加扩展名 *.sav*，将以 IBM SPSS Statistics 格式保存该文件。 在未指定扩展名时，SAV 为缺省值。

分发地图文件

在地图转换实用程序的第一步中，您选择了保存转换后的 SMZ 文件的位置。 您可能还选择了将地图添加到图形板模板选择器的管理系统。 如果选择了保存至管理系统，那么地图将可供同一计算机上运行的任何 IBM SPSS 产品使用。

要将地图分发给其他用户，您将需要向这些用户发送 SMZ。 然后这些用户使用管理系统来导入地图。 您可以简单地发送您在步骤 1 中指定其位置的文件。 如果您想要发送管理系统中的文件，您首先需要将其导出：

1. 在模板选择器中，单击**管理...**
2. 单击“地图”选项卡。
3. 选择您要分发的地图。
4. 单击**导出...** 并选择要保存文件的位置。

现在，您可以将实际地图文件发送给其他用户。 用户将需要反向执行此过程，并将地图导入到管理系统中。

散点图节点

“散点图”节点可显示各数字字段之间的关系。 可使用点创建图（即散点图），或者也可使用线段。 可通过在对话框中指定一种 X 模式来创建三种类型的线散点图。

X 模式 = 排序

如果将 X 模式设置为 **排序**，则数据将按照绘制在 *x* 轴上的字段的值进行排序。 这将在图形上画出一条贯穿左右的线。 如果将名义字段用作重叠，则将在图形上生成多条不同色调的、贯穿左右的线。

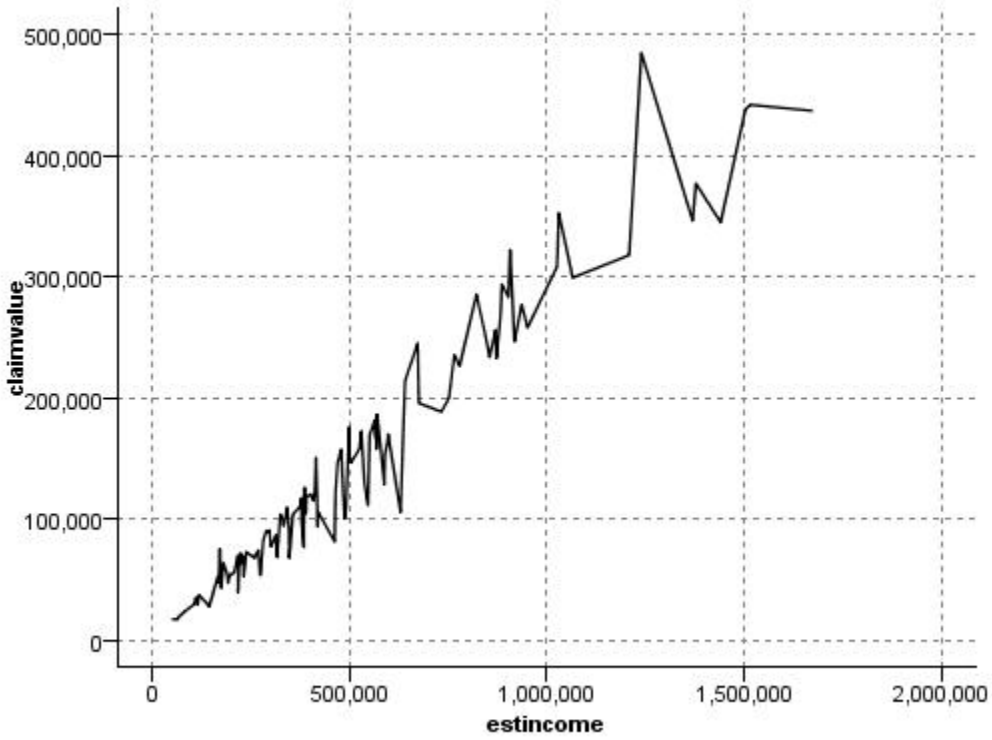


图 *19: X* 模式设置为“排序”的线散点图

X 模式 = 重叠

如果将 X 模式设置为 **重叠**，则将在同一图形上创建多线散点图。 不会对重叠散点图的数据进行排序；只要 *x* 上的值不断增大，数据将绘制在一条线上。 如果值减小，那么将开始一条新线。 例如，如果 *x* 从 0 增大到 100，则 *y* 值将绘制在一条线上。 *x* 降低为 100 以下时，则在第一条线之外，将绘制一条新线。 完成的散点图可能有多个散点图，它们可用于对比多个 *y* 值序列。 此类型的散点图对具有周期时间成分的数据很有用，比如连续 24 小时周期的用电需求。

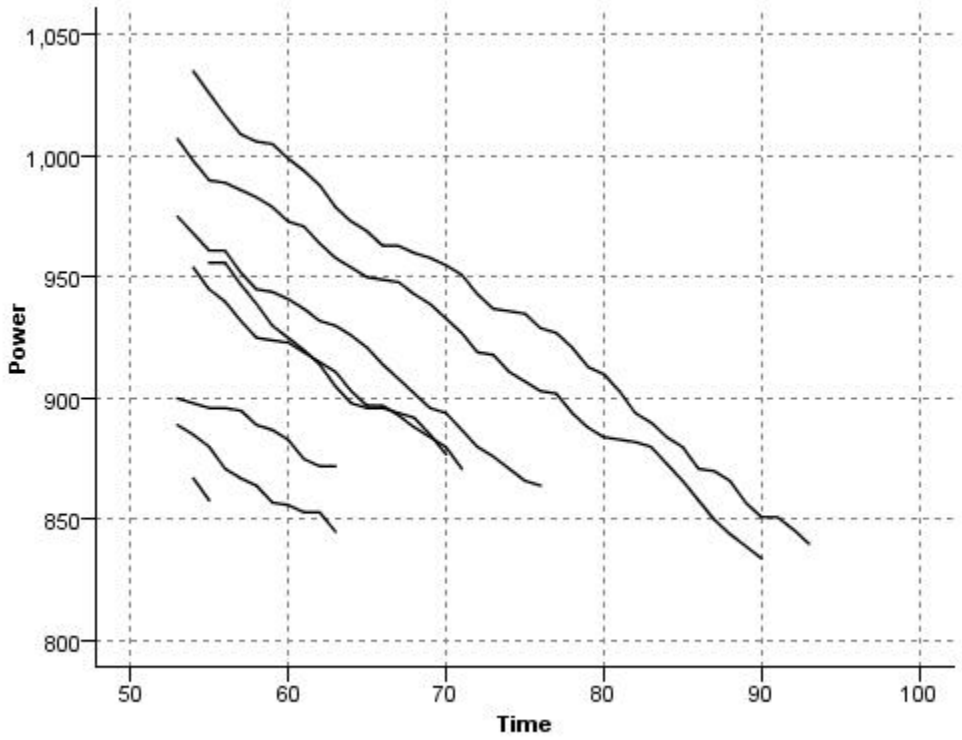


图 *20: X* 模式设置为“重叠”的线散点图

X 模式 = 如所读取

将 X 模式设置为 **如所读取**，散点图的 *x* 和 *y* 值将与从数据源读取的值一样。 对于具有时间序列成分的数据，此选项将有助于您研究与数据顺序关联的趋势和样式。 可能需要在创建此类型的散点图之前对数据进行排序。 它也可以用于对比设置为 **排序** 和 **如所读取** 的 X 模式，以便确定样式对排序的依赖程度。

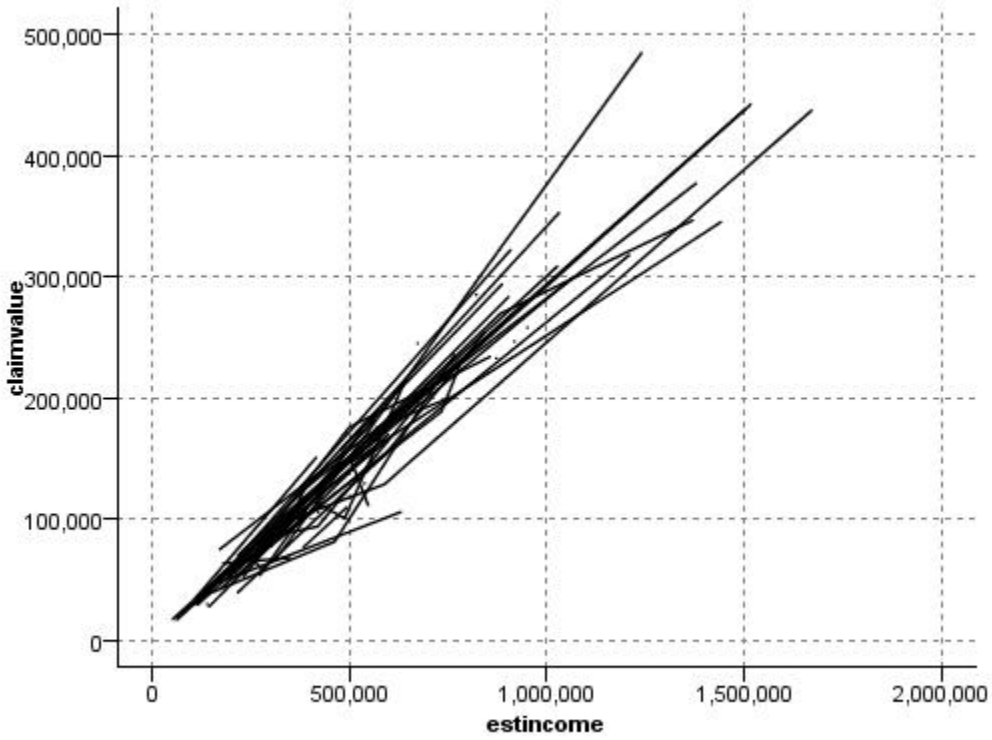


图 *21:* 线散点图起先显示为“排序”，随后以“如所读取”的 *X* 模式再次执行

也可使用“图形板”节点生成散点图和线散点图。 但是，此节点还提供了其他选项。 有关更多信息，请参阅主题 第 148 页的『可用的内置图形板可视化类型』。

散点图节点选项卡

散点图对照 *X* 字段的值，显示 *Y* 的值。 通常而言，这两个字段分别对应于一个自变量和一个因变量。

X 字段。 从列表中选择显示在水平 *x* 轴上的字段。

Y 字段。 从列表中选择显示在垂直 *y* 轴上的字段。

Z 字段。 单击 3D 图表按钮后，可以从列表中选择要显示在 *z* 轴上的字段。

Overlay）。 有几种方式可用于说明数据值类别。 例如，可以使用 *maincrop* 作为颜色重叠来指定补贴申请人种植的主要作物的 *estincome* 和 *claimvalue* 值。 有关更多信息，请参阅主题 第 142 页的『审美原则、重叠、面板和动画』。

覆盖类型。 指定是否显示重叠函数或光滑线。 光滑线和重叠函数总是作为 y 的函数来计算。

- **无**。 不显示任何重叠。
- **更平滑**。 显示平滑的拟合线，该拟合线用局部加权迭代稳健最小二乘回归 (LOESS) 来计算。 此方法可有效计算一系列的回归，每个回归均集中关注散点图内的一个小区域。 此操作将生成一连串“局部”回归线，然后可将这些线连接起来形成一条平滑曲线。

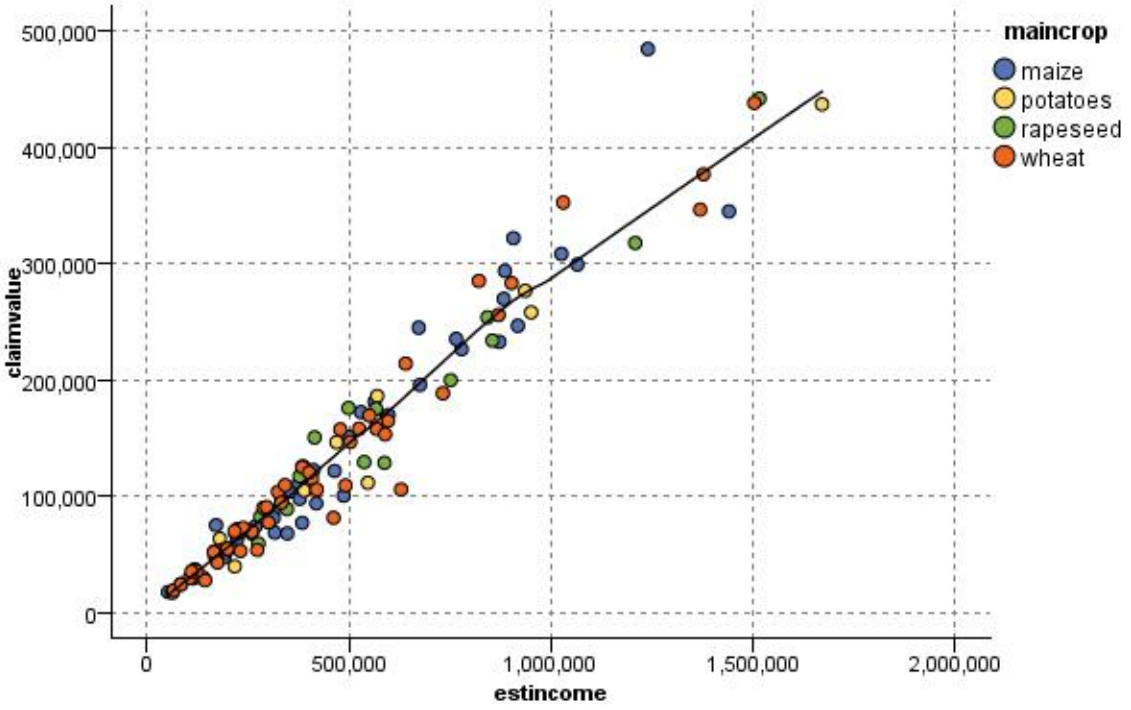


图 *22:* 具有 *LOESS* 光滑重叠的图

- **函数**。 选择此选项可指定一个用于与实际值进行比较的已知函数。 例如，要比较实际值与预测值，则可绘制函数 y = x 作为重叠。 在文本框中指定函数 **y =** 。 缺省函数为 y = x ，但您也可以指定任何类型的函数，如关于 x 的二次函数或任意表达式。

  注意：重叠函数不可用于面板图或动画图。

当您为一个散点图设置了选项后，可以通过单击对话框中的**运行**来运行绘制。 但是，您也许希望使用“选项”选项卡指定更多内容，如进行“分隔”、设置“X 模式”和“样式”。

散点图选项选项卡

样式。 选择**点**或**线**作为绘制样式。 选择 **线** 将激活 **X 模式** 控件。 选择 **点** 将使用加号 (+) 作为缺省的点形状。 创建图形后，可以更改点的形状并改变其大小。

X 模式 如要绘制线散点图，您需要选择“X 模式”以定义图的样式。 选择**排序**、**重叠**或**如所读取**。 如选择 **重叠** 或 **如所读取** ，则必须指定用来抽取前 *n* 个记录样本的数据集大小上限。 否则，将使用缺省的 2000 条。

自动 X 范围。 选择此选项将使用此轴上整个范围内的数据值。 取消选择此项，则将使用由您指定的 **最小值** 和 **最大值** 限定的值的精确子集。 您可以直接键入值或使用箭头。 缺省情况下，将选择自动范围以支持快速构建图形。

自动 Y 范围。选择此选项将使用此轴上整个范围内的数据值。 取消选择此项，则将使用由您指定的 **最小值** 和 **最大值** 限定的值的精确子集。 您可以直接键入值或使用箭头。 缺省情况下，将选择自动范围以支持快速构建图形。

自动 Z 范围。 仅用于在“散点图”选项卡上指定 3-D 图形的情况。选择此选项将使用此轴上整个范围内的数据值。 取消选择此项，则将使用由您指定的 **最小值** 和 **最大值** 限定的值的精确子集。 您可以直接键入值或使用箭头。 缺省情况下，将选择自动范围以支持快速构建图形。

抖动。 又称为**颤动**。在数据集中有许多重复值的情况下，“抖动”对于点图很有用处。 如要将值的分布观察地更加清楚，您可利用“抖动”使点随机分布在实际值周围。

使用先前版本 *IBM SPSS Modeler* 的用户请注意：在 IBM SPSS Modeler 的本发行版中，散点图中使用的抖动值所采用的度量方式与以前不同。 在早期版本中，该值是实际数字，但在本版本中，它是相对于框大小的

比例。 这就意味着，使用早期版本生成的流所具有的颤动值在本版本中可能过大。 在本版本中，任何非零的颤动值都将被转换为 0.2。

要绘制的最大记录数。 为大型数据集指定一种绘制方法。 可以指定数据集大小上限，或使用缺省的 2000 条记录。 如果选择 **分隔** 或 **抽样** 选项，则处理大数据集的性能将显著提高。 另外，您也可以选择 **使用所有数据**，但必须要注意，这一选项可能大幅降低软件的执行效率。

注意：如果“X 模式”设置为**重叠**或**如所读取**，那么上述选项将处于禁用状态且仅使用前 *n* 个记录。

- **分级**。 选择此选项可对所包含记录数超过指定数字的数据集进行分级。 “分级”使图形在实际绘制前被分散在较小的网格中，并计算在每个单元格中将出现的点的数目。 在最终图形中，每个网格中的分级矩心处将出现一个点（该点即代表分级中所有点位置的平均数）。 所绘制符号的大小表示在此区域内点的数目（除非您用大小作为重叠）。 使用矩心及尺寸代表点的数量使分级后的散点图成为表现大数据集的最佳方式。 因为该方式杜绝了在密集区域过量绘制（点的颜色没有区别）的问题，也减少了符号误导的问题（即点的密度出现偏差）。 当某些符号（特别是加号 [+]）部分重叠时，其所产生的密集区域并不是原始数据的真实反映，这一现象称为符号误导。
- **样本**。 选择此选项将随机抽取数量相当于文本框中所输入记录数的数据。 缺省值为 2,000。

散点图外观选项卡

可以在创建图形前指定外观选项。

标题。 输入要用作图形标题的文本。

子标题。 输入要用作图形子标题的文本。

文字说明。 输入要用作图形文字说明的文本。

X 标签。 接受自动生成的 *x* 轴（水平）标签，或者选择**定制**来指定标签。

Y 标签。 接受自动生成的 *y* 轴（垂直）标签，或者选择**定制**来指定标签。

Z 标签。 仅可用于 3D 图形，接受自动生成的 *z* 轴标签，或者选择**定制**来指定定制标签。

显示网格线。 此选项缺省情况下处于选中状态，它显示散点图或图形背后的网格线，以便更轻松地确定区域和带状分界值点。 网格线通常用白色显示，但图形背景为白色时除外；此情况下，网格线用灰色显示。

使用散点图

散点图和多重散点图基本上是 *X* 相对于 *Y* 的散点图。 例如，如果您正在探索农业补贴应用程序中的潜在欺诈行为，那么您可能想要绘制在应用程序上索赔的收入图与神经网络估算的收入图。 可以使用重叠，如农作物类型，来证实索赔（值或数量）与作物类型之间是否存在一定关系。

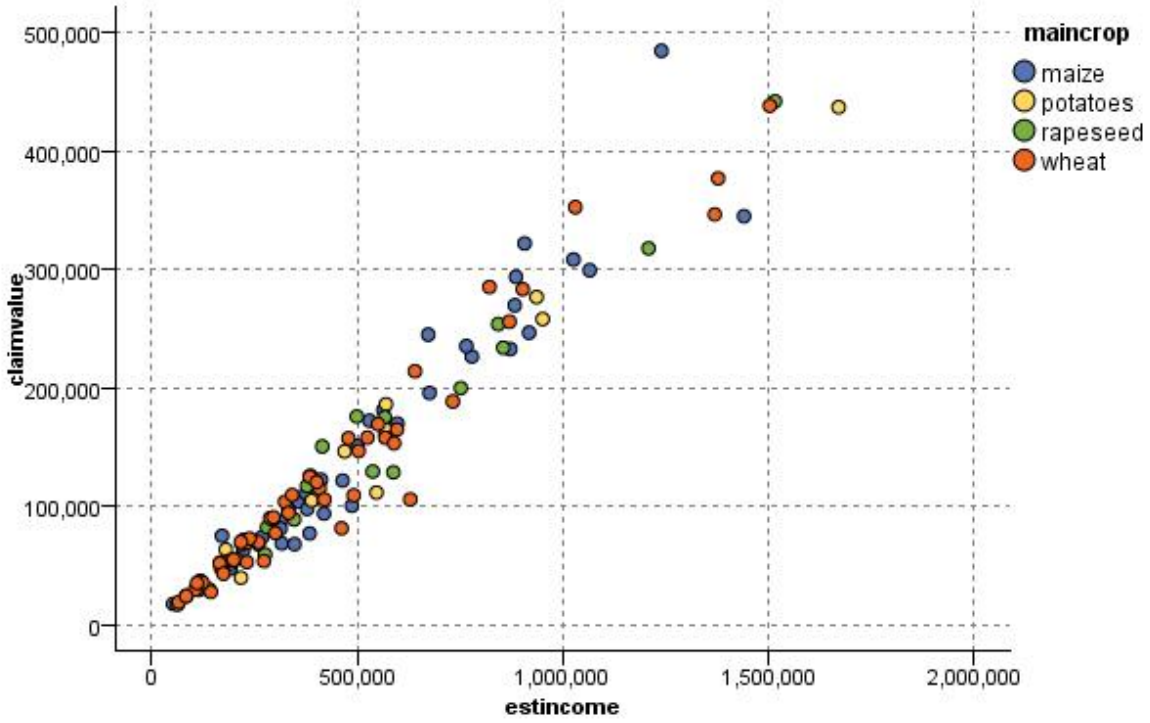


图 *23:* 将主要作物类型作为重叠来绘制估计收入与索赔值之间关系的散点图

由于散点图、多重散点图和评估图表是 *Y* 相对于 *X* 的二维图形显示，通过定义区域，标记元素或绘制条形区可以很容易地与它们进行交互。 也可以为这些区域、条形图或元素所代表的数据生成节点。 有关更多信息，请参阅主题 第 208 页的『探索图形』。

多重散点图节点

多重散点图是散点图的特殊类型，它显示相对于单一 *X* 字段的多个 *Y* 字段。 *Y* 字段将被绘制为彩色线，且每条线都相当于一个“样式”设置为**线**而“X 模式”设置为**排序**的散点图节点。 当您使用时间序列数据并要研究几个变量在一段时间中的变化时，多重散点图十分有用。

多重散点图选项卡

X 字段。 从列表中选择显示在水平 *x* 轴上的字段。

Y 字段。 从列表中选择一个或多个根据 *X* 字段值范围显示的字段。 使用“字段选择器”按钮选择多个字段。 单击“删除”按钮从列表中删除字段。

Overlay）。 有几种方式可用于说明数据值类别。 例如，您可以使用动画重叠显示数据中所有值的多重散点图。 这对包含 10 个类别以上的集非常有用。 当集合中使用的类别超过 15 个 时，您可能会注意到性能下降了。 有关更多信息，请参阅主题 第 142 页的『审美原则、重叠、面板和动画』。

标准化。 选中此选项可将所有 *Y* 值都定标到 0–1 这一范围内，以便在图形上显示。 标准化有助于研究多条线之间的关系，如果不使用标准化，这种关系可能由于每个系列的值范围的差异而变得不明显，建议在同一个图形中绘制多条线时或比较并行面板中的散点图时使用标准化选项。 （当所有的数据值都落在相似范围内时，不必使用标准化选项。）

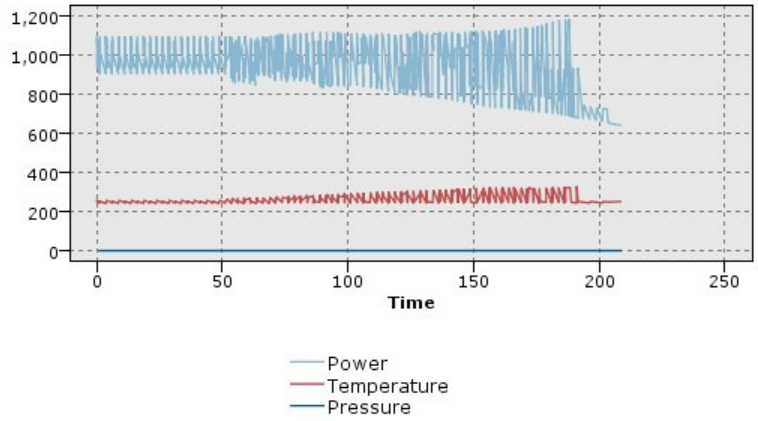


图 *24:* 显示发电厂在一段时间中的变化的标准多重散点图（请注意，如果未选择“标准化”，则压力的散点图不可见）

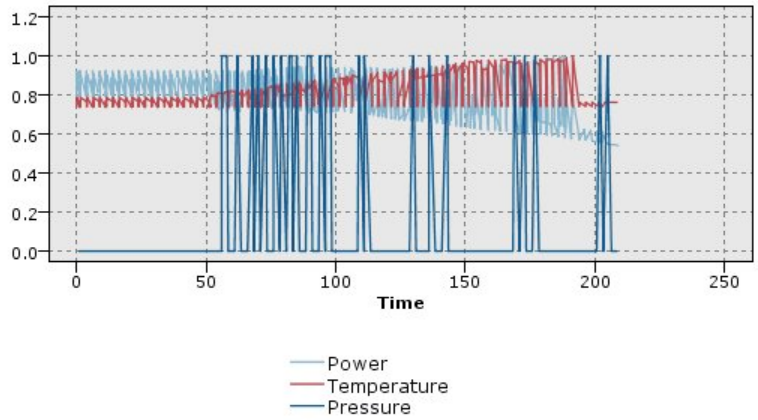


图 *25:* 显示压力散点图的标准化多重散点图

覆盖函数。 选择此选项可指定一个用于与实际值进行比较的已知函数。 例如，要比较实际值与预测值，则可绘制函数 y = x 作为重叠。 在文本框中指定函数 **y =** 。 缺省函数为 y = x ，但您也可以指定任何类型的函数，如关于 x 的二次函数或任意表达式。

注意：重叠函数不可用于面板图或动画图。

当记录数大于以下值时。 为大型数据集指定一种绘制方法。 可以指定数据集大小上限，或使用缺省的 2000 点。 如果选择 **分隔** 或 **抽样** 选项，则处理大数据集的性能将显著提高。 另外，您也可以选择 **使用所有数据** ，但必须要注意，这一选项可能大幅降低软件的执行效率。

注意：如果“X 模式”设置为**重叠**或**如所读取**，那么上述选项将处于禁用状态且仅使用前 *n* 个记录。

- **分级**。 选择此选项可对所包含记录数超过指定数字的数据集进行分级。 “分级”使图形在实际绘制前被分散在较小的网格中，并计算在每个单元格中将出现的连接数。 在最终图形中，每个网格中的分级矩心处将使用一个连接（该连接即代表分级中所有连接点位置的平均数）。
- **样本**。 选择此选项将随机抽取指定记录数的数据样本。

多重散点图外观选项卡

可以在创建图形前指定外观选项。

标题。 输入要用作图形标题的文本。

子标题。 输入要用作图形子标题的文本。

文字说明。 输入要用作图形文字说明的文本。

X 标签。 接受自动生成的 *x* 轴（水平）标签，或者选择**定制**来指定标签。

Y 标签。 接受自动生成的 *y* 轴（垂直）标签，或者选择**定制**来指定标签。

显示网格线。 此选项缺省情况下处于选中状态，它显示散点图或图形背后的网格线，以便更轻松地确定区域和带状分界值点。 网格线通常用白色显示，但图形背景为白色时除外；此情况下，网格线用灰色显示。

使用多重散点图

散点图和多重散点图基本上是 *X* 相对于 *Y* 的散点图。 例如，如果您正在探索农业补贴应用程序中的潜在欺诈行为，那么您可能想要绘制在应用程序上索赔的收入图与神经网络估算的收入图。 可以使用重叠，如农作物类型，来证实索赔（值或数量）与作物类型之间是否存在一定关系。

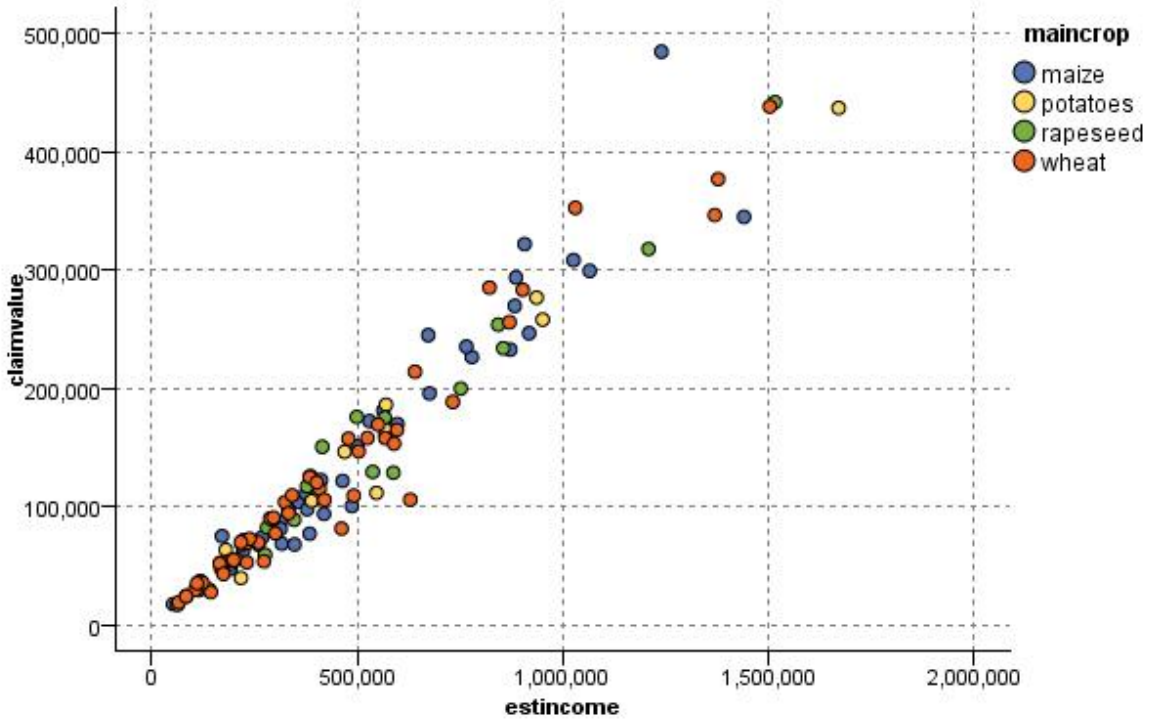


图 *26:* 将主要作物类型作为重叠来绘制估计收入与索赔值之间关系的散点图

由于散点图、多重散点图和评估图表是 *Y* 相对于 *X* 的二维图形显示，通过定义区域，标记元素或绘制条形区可以很容易地与它们进行交互。 也可以为这些区域、条形图或元素所代表的数据生成节点。 有关更多信息，请参阅主题 第 208 页的『探索图形』。

时间散点图节点

您可以使用“时间散点图”节点查看一个或多个绘制的在一段时间内的时间序列。 这些由您绘制的序列包含数字值，并且被假定将在一个时间范围内（其中的周期一致）发生。

在 SPSS Modeler V17.1 和更低版本中，通常在使用时间散点图节点前，应使用时间间隔节点，创建 *TimeLabel* 字段。缺省情况下，在图形中，该字段为 *x* 轴标签。

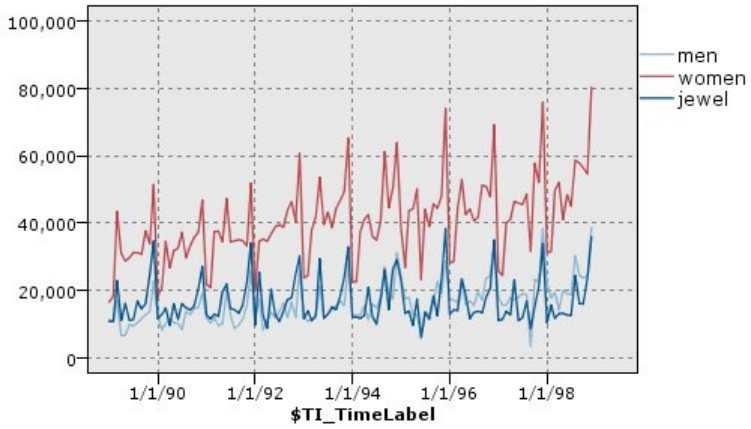


图 *27:* 绘制一段时间内男士及女士服装、珠宝的销售额

创建干预和事件

可以在时间散点图中，使用上下文菜单生成导出（标志或名义）节点，并由此创建事件和干预字段。 例如，您可以将铁路工人罢工这个情况创建为一个事件字段。如果事件发生，那么派生状态为真，如未发生则为假。 对于干预字段，以价格增长作为例，您可以使用派生计数标识增长日期，0 代表旧价格，1 代表新价格。 有关更多信息，请参阅主题 第 115 页的『“派生”节点』。

时间散点图选项卡

散点图。 提供一个绘制时间序列数据的选择。

- **选定系列**。 绘制选定时间序列的值。 如果选择了此选项，则在绘制置信度区间时，将取消选择 **标准化** 。
- **选定的时间序列模型**。 与时间序列模型结合使用，此选项为一个或多个选定的时间序列绘制所有的相关字段（实际和预测值以及置信度区间）。 此选项禁用对话框中的其他选项。 如果绘制置信度区间，则此选项为首选项。

序列。 选择您要绘制的一个或多个带有时间序列数据的字段。 数据必须是数字。

X 轴标签。 选择缺省标签或者单一字段用作散点图中 *x* 轴的标签。 如果选择“缺省值”，那么系统将使用“时间间隔”节点中创建的 TimeLabel 字段作为上游（针对 SPSS Modeler V17.1 和更低版本中创建的流）。如果没有上游“时间间隔”节点，那么将选择有序整数列。

在单独的面板中显示系列。 指定是否将每个序列显示在单独的面板中。 另外，如果您未选择面板，那么所有时间序列都将绘制在同一图形中，而且光滑线也不可用。 如果将所有时间序列绘制在同一个图形中，每个序列都将有不同的颜色代表。

标准化。 选中此选项可将所有 *Y* 值都定标到 0–1 这一范围内，以便在图形上显示。 标准化有助于研究多条线之间的关系，如果不使用标准化，这种关系可能由于每个系列的值范围的差异而变得不明显，建议在同一个图形中绘制多条线时或比较并行面板中的散点图时使用标准化选项。 （当所有的数据值都落在相似范围内时，不必使用标准化选项。）

显示。 选择一个或多个要在散点图中显示的元素。 您可以选择“线”、“点”和 (LOESS) “光滑线”。 只有在您选择将多个序列显示在不同面板中时，“光滑线”才可用。 缺省情况下将选择“线”元素。 请确保在您运行图形节点前，至少选择一种绘制元素；不然，系统将返回错误，告知您未选择可绘制的内容。

限制记录数。 如果要限制绘制的记录数，请选择此选项。 指定记录条数，从数据文件开头读取，这些将在 **要绘制的最大记录数** 选项中绘制。 缺省情况下，此数字设置为 2,000。 如果您要绘制数据文件中倒数 *n* 条记录，您可以事先使用排序节点将记录按时间降序排列。

时间散点图外观选项卡

可以在创建图形前指定外观选项。

标题。 输入要用作图形标题的文本。

子标题。 输入要用作图形子标题的文本。

文字说明。 输入要用作图形文字说明的文本。

X 标签。 接受自动生成的 *x* 轴（水平）标签，或者选择**定制**来指定标签。

Y 标签。 接受自动生成的 *y* 轴（垂直）标签，或者选择**定制**来指定标签。

显示网格线。 此选项缺省情况下处于选中状态，它显示散点图或图形背后的网格线，以便更轻松地确定区域和带状分界值点。 网格线通常用白色显示，但图形背景为白色时除外；此情况下，网格线用灰色显示。

布局。 仅对时间散点图，可指定时间值散点是沿横轴分布还是沿纵轴分布。

使用时间散点图

在创建了“时间散点图”节点后，有几个选项可用来调整图形显示并生成节点以备进一步分析。 有关更多信息，请参阅主题 第 208 页的『探索图形』。

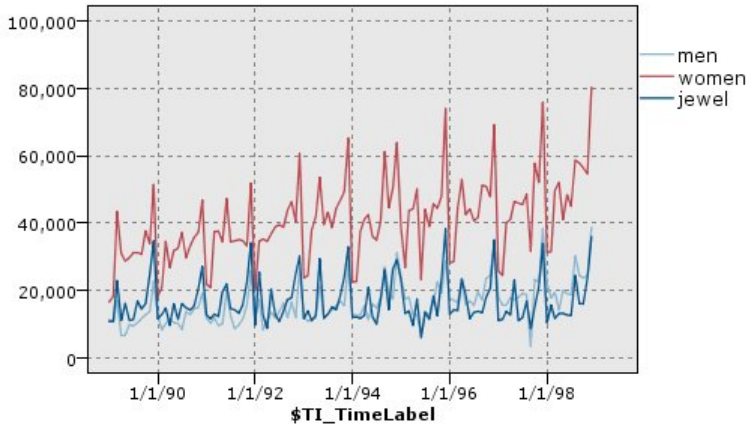


图 *28:* 绘制一段时间内男士及女士服装、珠宝的销售额

在创建了时间散点图、定义了带状区域并查看了结果之后，您可以使用“生成”菜单上的选项及上下文菜单创建“选择”或“派生”节点。 有关更多信息，请参阅主题 第 214 页的『从图形中生成节点』。

分布节点

分布图或表显示数据集中符号（非数字）值的出现情况，比如抵押类型或性别。 “分布”节点的一个典型用法是：在创建模型前，显示数据中可使用“平衡”节点纠正的不平衡度。 您可以使用分布图或表窗口中的“生成”菜单自动生成平衡节点。

也可以使用“图形板”节点生成图形计数条。 但是，此节点还提供了其他选项。 有关更多信息，请参阅主题 第 148 页的『可用的内置图形板可视化类型』。

注意：要显示数据值的出现情况，应该使用“直方图”节点。

分布图选项卡

散点图。 选择分布类型。 选择 **选定字段** 将显示选定字段的分布。 选择 **所有标志字段（真值）** 显示数据集中，值为真的标志字段的分布。

字段。 选择名义或标志字段以为其显示值的分布。 在字段列表中，只会出现类型未被明确设置为数字的字段。

重叠。 选择用作颜色重叠的名义或标志字段，从而显示该字段值在指定字段的每个值中的分布情况。 例如，您可以使用市场活动响应 (*pep*) 作为儿童 (*children*) 数量的重叠字段，说明对活动的热情与家庭大小的关系。 有关更多信息，请参阅主题 第 142 页的『审美原则、重叠、面板和动画』。

按颜色标准化。 选择此选项将按比例缩放图条，因此，所有的图条都将填满图形的整个幅宽。 重叠值相当于每个图条的一部分，使用它可轻松比较不同类别。

排序。 选择在分布图中显示值的方法。 选择 **按字母顺序** 使用字母顺序，或选择 **按计数** 根据出现频率的降序排列。

比例尺。 选择此选项将按比例缩放值分布，因此，计数最大的值将填满整个图的幅宽。 所有其他的图条都将根据此值进行调整。 取消选择此选项，则图条将根据每个值的全部计数进行缩放。

分布外观选项卡

可以在创建图形前指定外观选项。

标题。 输入要用作图形标题的文本。

子标题。 输入要用作图形子标题的文本。

文字说明。 输入要用作图形文字说明的文本。

X 标签。 接受自动生成的 *x* 轴（水平）标签，或者选择**定制**来指定标签。

Y 标签。 接受自动生成的 *y* 轴（垂直）标签，或者选择**定制**来指定标签。

显示网格线。 此选项缺省情况下处于选中状态，它显示散点图或图形背后的网格线，以便更轻松地确定区域和带状分界值点。 网格线通常用白色显示，但图形背景为白色时除外；此情况下，网格线用灰色显示。

使用分布节点

分布节点用于显示数据集中符号值的分布情况。 通常，在使用处理类节点前，将使用分布节点对数据进行检查并更正任何不平衡处。 例如，如果无子女的响应者的实例远多于其它类型的响应者，您可能想要减少此种实例，以使后续数据挖掘操作能够生成更加有用的规则。 分布节点将帮助您检查并找到这些不平衡处。

分布节点与众不同之处在于它能以图形和表两种方式分析数据。

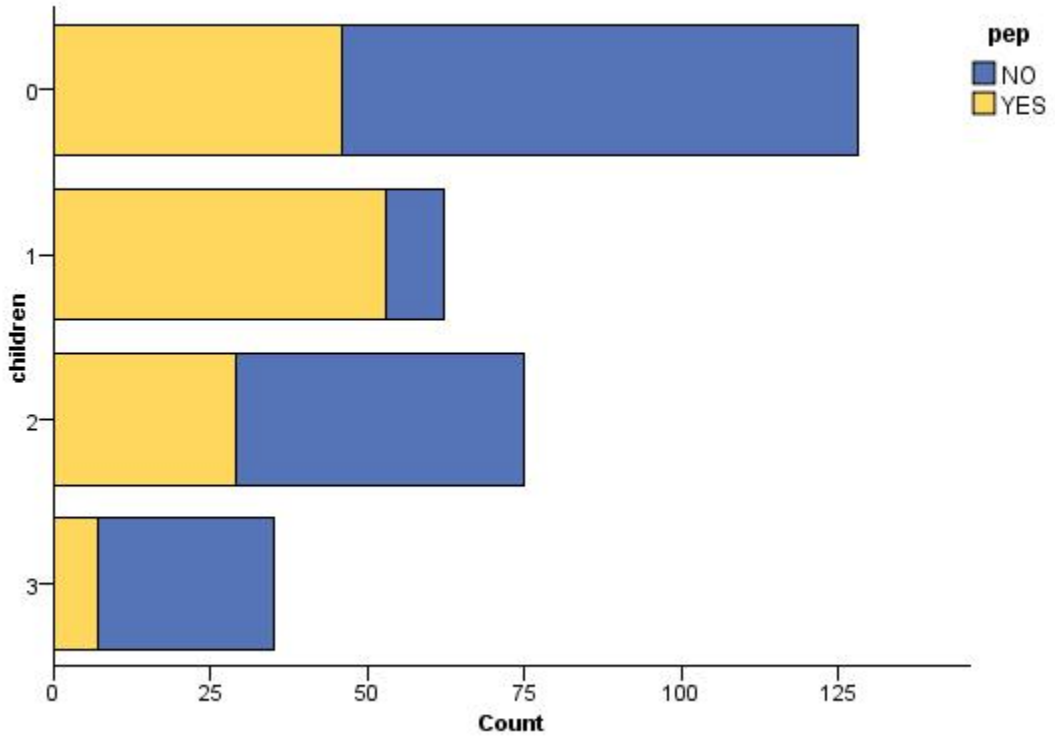


图 *29:* 分布图显示对市场活动有响应的人（有孩子或没孩子）数

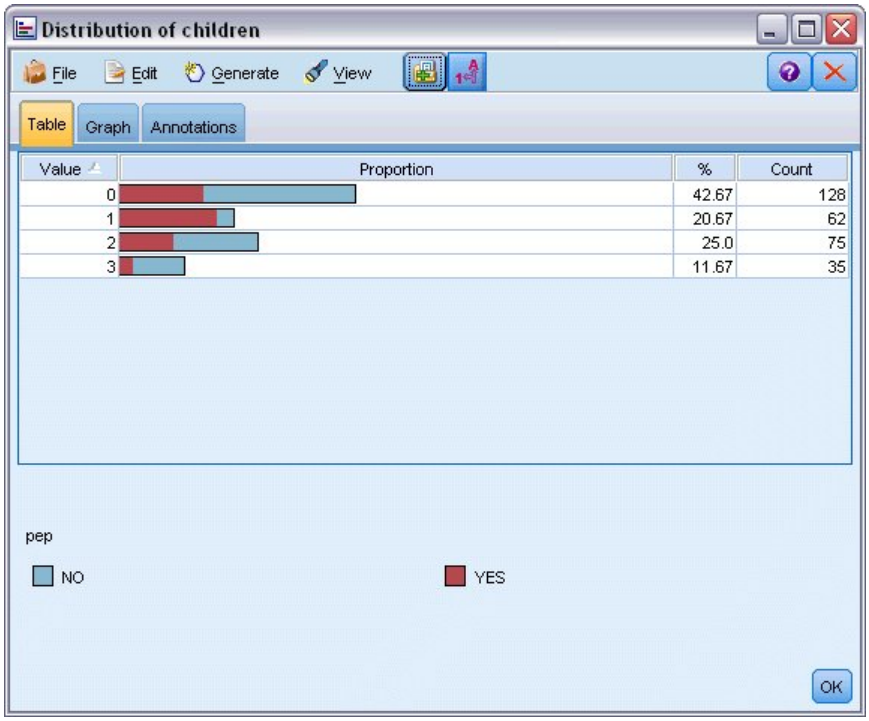


图 *30:* 分布表显示对市场活动有响应的人（有孩子或没孩子）数所占比例

在创建分布表和分布图并检查结果后，您可以使用菜单中的选项对值进行分组、复制值并生成一些用于数据准备的节点。 此外，您也可以复制或导出图形和图表信息以在其他应用程序（如 MS Word 或 MS PowerPoint）中使用。 有关更多信息，请参阅主题 第 226 页的『打印、保存、复制和导出图形』。

选择和复制条形图表中的值

1. 单击并按住鼠标按键，同时在行间拖动鼠标以选择值的集合。 也可以使用“编辑”菜单**全选**值。
2. 从“编辑”菜单，选择**复制表**或**复制表（包含 字段名称）**。
3. 粘贴到剪贴板或所需应用程序。

注意：不会直接复制图条。 而是复制表值。 也就是说，在由复制得到的表中不会显示出重叠的值。

从条形图表中将值分组

1. 通过按住 Ctrl 键并单击的方法选择要进行分组的值。
2. 在“编辑”菜单中选择**分组**。

注意：在对值分组或取消值分组时，将自动重新绘制“图形”选项卡上的图形以显示更改。

还可以进行下列操作：

- 在条形图列表中选择组名称并从“编辑”菜单中选择**取消分组**，取消值的分组。
- 在条形图列表中选择组名称并从“编辑”菜单中选择**编辑组**，对组进行编辑。 此时，将打开一个对话框，通过它可将值移入或移出该组。

“生成”菜单选项

可使用“生成”菜单上的选项选择数据子集、派生标志字段、重新对值进行分组、重新对值进行分类或平衡图形或表中的数据。 这些操作将生成一个数据准备节点，并将其放置在流画布中。 要使用生成的节点，请将其连接到现有流中。 有关更多信息，请参阅主题 第 214 页的『从图形中生成节点』。

直方图节点

直方图节点显示数字字段值的出现率。 在进行操作和建模之前，直方图常被用来检查数据。 与“分布”节点相同，“直方图”节点常常用来显示数据中的不平衡度。 也可以使用“图形板”节点生成直方图，此节点还提供了许多其他选项。 有关更多信息，请参阅主题 第 148 页的『可用的内置图形板可视化类型』。

注意：要显示符号字段值的出现情况，应该使用“分布”节点。

直方图选项卡

字段。 选择要为其显示值分布的数字字段。 在字段列表中，只会出现类型未被明确设置为符号（类别）的字段。

Overlay）。 选择一个符号字段以显示指定字段中值的类别。 选择重叠字段将使直方图变成堆积图，图中以各种颜色显示重叠字段的不同类别。 如果使用“直方图”节点，那么将存在三种类型的重叠：颜色、面板和动画。 有关更多信息，请参阅主题 第 142 页的『审美原则、重叠、面板和动画』。

直方图选项选项卡

自动 X 范围。 选择此选项将使用此轴上整个范围内的数据值。 取消选择此项，则将使用由您指定的 **最小值** 和 **最大值** 限定的值的精确子集。 您可以直接键入值或使用箭头。 缺省情况下，将选择自动范围以支持快速构建图形。

分箱。 请选择**按数字**或**按宽度**。

- 选择 **按数量** 可显示固定数量的图条，这些图条的宽度取决于指定分级的范围和数量。 指示 **分级数量** 选项的图形中要使用的分级数量。 使用箭头调整该数字。
- 选择 **按宽度** 可创建一个图条宽度固定的图形。 分级数量取决于指定的宽度和值的范围。 指示 **分级宽度** 选项中 图条的宽度。

按颜色标准化。 选择此选项可将所有条形调整为同一高度，以便重叠的值在每个条形中显示为占所有观测值的百分比。

显示正态曲线。 选择此选项可向图形添加正态曲线来显示数据的均值和方差。

每种颜色的独立带状区域。 选择此选项会在图形中将每个重叠的值显示为一个独立的带状区域。

直方图外观选项卡

可以在创建图形前指定外观选项。

标题。 输入要用作图形标题的文本。

子标题。 输入要用作图形子标题的文本。

文字说明。 输入要用作图形文字说明的文本。

X 标签。 接受自动生成的 *x* 轴（水平）标签，或者选择**定制**来指定标签。

Y 标签。 接受自动生成的 *y* 轴（垂直）标签，或者选择**定制**来指定标签。

显示网格线。 此选项缺省情况下处于选中状态，它显示散点图或图形背后的网格线，以便更轻松地确定区域和带状分界值点。 网格线通常用白色显示，但图形背景为白色时除外；此情况下，网格线用灰色显示。

使用直方图

直方图显示沿 *x* 轴分布的数字字段的值。 直方图和集合图的操作相似。 收集则显示与 其他字段值相关的 一个数字字段的值，而不是单个字段中值的出现率。

创建图形后，可以检查结果，定义沿 *x* 轴划分值的带状区域或定义区域。 也可以在图形内标记元素。 有关更多信息，请参阅主题 第 208 页的『探索图形』。

可以使用“生成”菜单上的选项来创建 Balance、Select 或 Derive 节点，而且这些节点使用图形中的数据或具体到条带区域、区域或标记元素内的数据。 此类图形常常用在操控类节点之前，可用来观察数据并通过在流中使用的图形生成平衡节点修正不平衡的问题。 您还可以生成“派生标志”节点以添加一个字段，将符合条件的记录标记出来；或生成“选择”节点以选择某个集合或值的范围内的所有记录。 上述操作将帮助您重点关注某个数据子集，以便进一步检查数据。 有关更多信息，请参阅主题 第 214 页的『从图形中生成节点』。

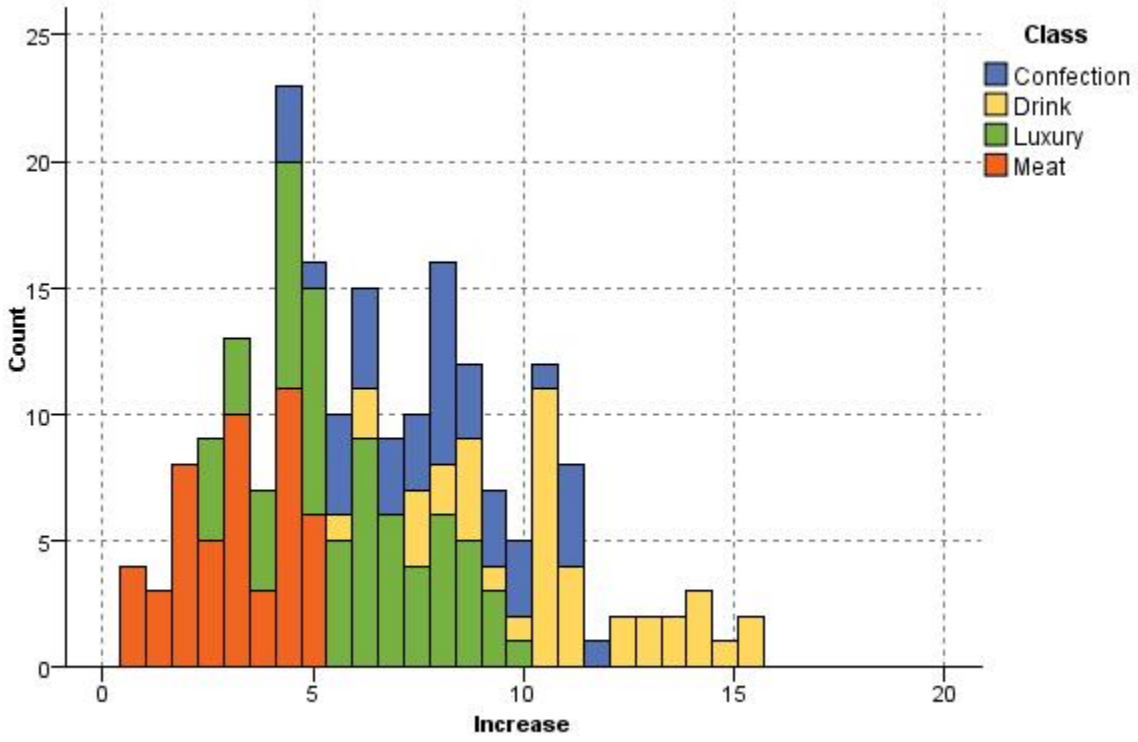


图 *31:* 直方图分类显示由促销活动带来的购买增长情况的分布

收集节点

收集与直方图基本相同，只有一个区别，即收集显示与其他字段值相关的一个数字字段的值，而不是单个字段中值的出现率。 要通过图示说明值不断变化的变量或字段时，可使用收集。 使用 3-D 图形表示时，还可以使用按分类显示分布的符号轴。 二维收集显示为使用重叠的堆积条形图。 有关更多信息，请参阅主题 第 142 页的『审美原则、重叠、面板和动画』。

收集散点图选项卡

收集。 选择一个字段，并依据指定的**超出**字段值的范围来收集和显示该字段的值。 只有字段类型未被定义为符号的字段会被列出。

范围。 选择一个字段，其值将用于显示**收集**中指定的字段。

依据。 创建 3-D 图形时，该选项将被启用，您可以通过它选择用于按类别显示收集字段的名义或标志字段。

操作。 选择集合图中每个图条代表的内容。 选项包括 **合计**， **平均值**， **最大值**， **最小值** 和 **标准差**。

Overlay）。 选择一个符号字段以显示选定字段中值的类别。 选择一个重叠字段，将收集进行转换并为不同颜色的各个类别分别创建多个图条。 此节点可使用三种类型的重叠：颜色、面板和动画。 有关更多信息，请参阅主题 第 142 页的『审美原则、重叠、面板和动画』。

收集选项选项卡

自动 X 范围。 选择此选项将使用此轴上整个范围内的数据值。 取消选择此项，则将使用由您指定的 **最小值** 和 **最大值** 限定的值的精确子集。 您可以直接键入值或使用箭头。 缺省情况下，将选择自动范围以支持快速构建图形。

分箱。 请选择**按数字**或**按宽度**。

- 选择 **按数量** 可显示固定数量的图条，这些图条的宽度取决于指定分级的范围和数量。 指示 **分级数量** 选项的图形中要使用的分级数量。 使用箭头调整该数字。
- 选择 **按宽度** 可创建一个图条宽度固定的图形。 分级数量取决于指定的宽度和值的范围。 指示 **分级宽度** 选项中 图条的宽度。

收集外观选项卡

可以在创建图形前指定外观选项。

标题。 输入要用作图形标题的文本。

子标题。 输入要用作图形子标题的文本。

文字说明。 输入要用作图形文字说明的文本。

上方标签。 接受自动生成的标签，或选择**定制**指定标签。

收集标签。 接受自动生成的标签，或选择**定制**指定标签。

按标签。 接受自动生成的标签，或选择**定制**指定标签。

显示网格线。 此选项缺省情况下处于选中状态，它显示散点图或图形背后的网格线，以便更轻松地确定区域和带状分界值点。 网格线通常用白色显示，但图形背景为白色时除外；此情况下，网格线用灰色显示。

以下示例显示三维图形上各个外观选项的位置。

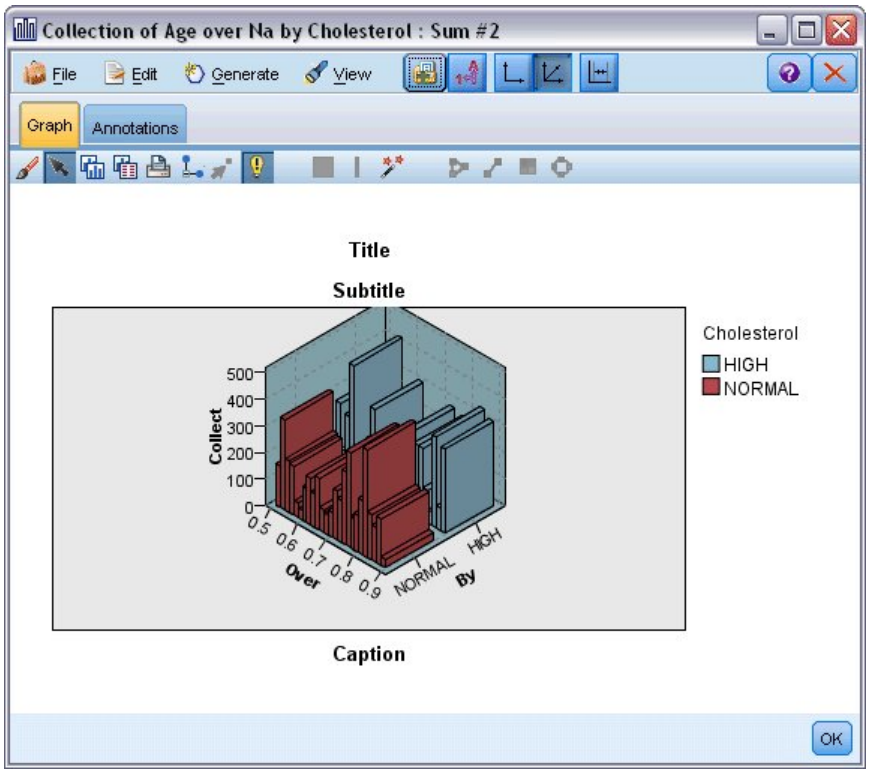


图 *32:* 三维集合图形上各个图形外观选项的位置

使用集合图

收集则显示与 其他字段值相关的 一个数字字段的值，而不是单个字段中值的出现率。 直方图和集合图的操作相似。 直方图显示沿 *x* 轴分布的数字字段的值。

创建图形后，可以检查结果，定义沿 *x* 轴划分值的带状区域或定义区域。 也可以在图形内标记元素。 有关更多信息，请参阅主题 第 208 页的『探索图形』。

可以使用“生成”菜单上的选项来创建 Balance、Select 或 Derive 节点，而且这些节点使用图形中的数据或具体到条带区域、区域或标记元素内的数据。 此类图形常常用在操控类节点之前，可用来观察数据并通过在流

中使用的图形生成平衡节点修正不平衡的问题。 您还可以生成“派生标志”节点以添加一个字段，将符合条件的记录标记出来；或生成“选择”节点以选择某个集合或值的范围内的所有记录。 上述操作将帮助您重点关注某个数据子集，以便进一步检查数据。 有关更多信息，请参阅主题 第 214 页的『从图形中生成节点』。

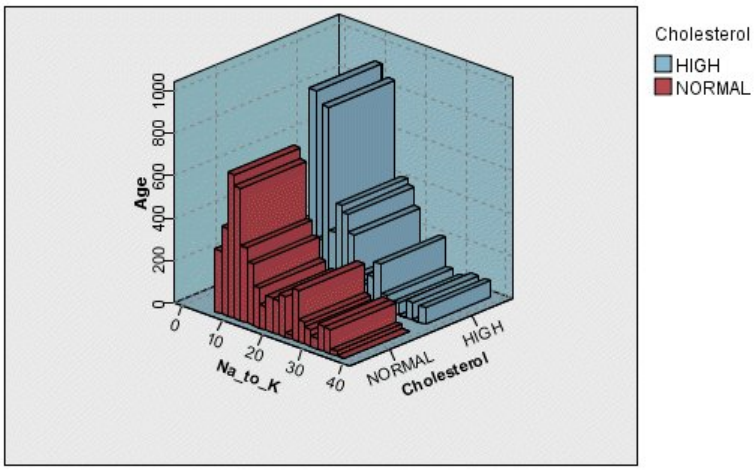


图 *33:* 三维集合图显示针对高、正常胆固醇水平，*Na_to_K* 相对于 *Age* 的合计

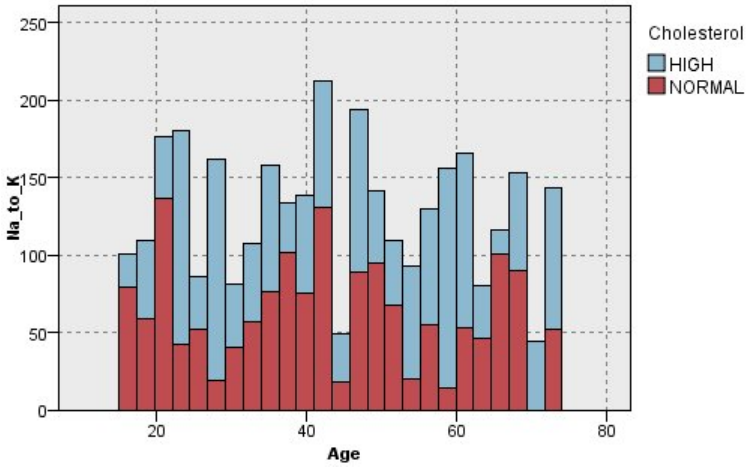


图 *34:* 未显示 *z* 轴，但将胆固醇作为重叠以颜色标出的集合图

网络节点

Web 节点显示两个或两个以上符号字段的值之间关系的紧密程度。 其图形使用不同类型的线条显示链接，说明链接强度。 例如，您可以使用网络节点研究通过电子贸易网站（或在传统零售店）所购买的不同商品之间的关系。

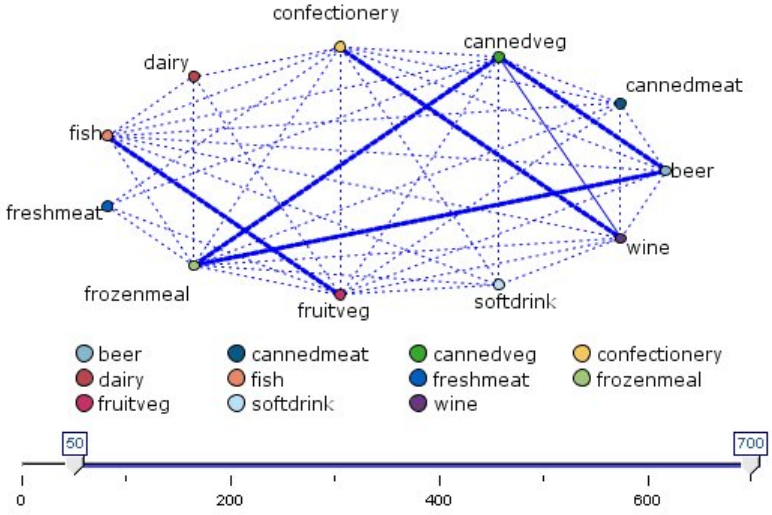


图 *35: Web* 图形显示购买杂货物品之间的关系

导向网络

导向 Web 节点与 Web 节点相同，二者都可以显示符号字段间关系的紧密程度。 但是，导向 Web 图形只显示从一个或多个源字段到一个结束字段之间的连接。 这些链接都是有方向性的，即它们都是单向的。

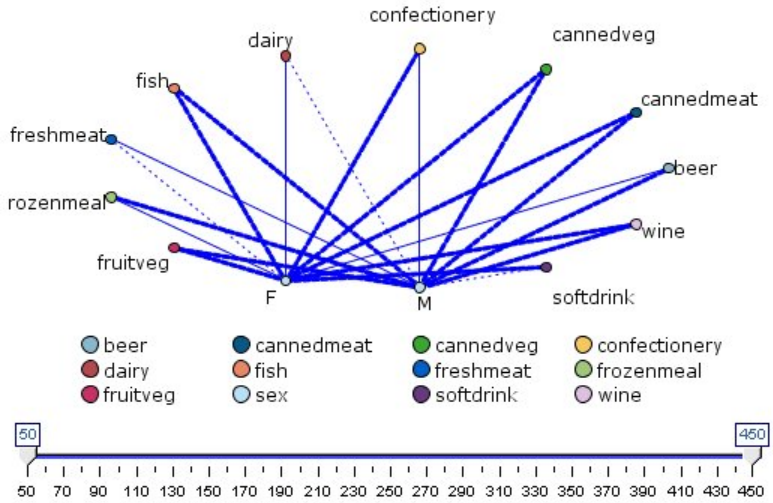


图 *36:* 定向 *Web* 图形显示购买杂货物品与性别之间的关系

与 Web 节点相同，其图形使用不同类型的线条显示连接来说明连接强度。 例如，您可以使用导向网络节点研究性别与对购买某正商品的倾向性之间的关系。

网络散点图选项卡

Web. 选择此选项将创建一个 Web 图形，显示所有指定字段间的关系紧密程度。

定向 Web。 选择此选项将创建一个具有方向性的 Web 图形，显示多个字段与某个字段的值（如性别或宗教信仰）之间的关系紧密程度。 如果选择此选项，则将激活“结束字段”并且其下面的字段控件也被重命名为“源字段”，以加以明确区分。

目标字段（仅限定向 Web）。 选择一个用于导向 Web 的标志或名义字段。 只有字段类型未被明确定义为数字的字段会被列出。

字段/源字段。 选择字段来创建 Web 图形。 只有字段类型未被明确定义为数字的字段会被列出。 使用“字段选择器”按钮选择多个字段或根据类型选择字段。

注意：对于定向 Web 来说，这一控件用来选择“源字段”。

仅显示 true 标志。 选择此选项将仅显示标志字段中值为真的标志。 此选项简化了网络图的显示，常用于正面数据值的出现意义异常重要的情况。

行值为。 从下拉列表中选择一个阈值类型。

- **绝对值**选项将根据带有成对值的记录数设置阈值。
- **总体百分比**选项将链接所代表的观测值数的绝对值显示为相对于 Web 图形全部对值的出现次数的比例。
- **较小字段/值的百分比**和**较大字段/值的百分比**说明要使用哪个字段/值来估计百分比。 例如，字段 *Drug* 中有 100 条记录值为 *drugY*，但字段 *BP* 中只有 10 条记录值为 *LOW*。 有七条记录同时具有值 *drugY* 和 *LOW*，因此，根据您用来参考的字段不同（较小： *BP* 或较大： *Drug*），百分比分别为 70% 或 7%。

注意：对于定向 Web 图来说，上述第三和第四个选项不可用。 作为替代，您可以选择**“目标”字段/值的百分比**和**“源”字段/值的百分比**。

强链接较重。 此选项缺省情况下处于选中状态，即查看字段间链接的标准方式。

弱链接较重。 选择此选项将使显示线条粗细所代表的含义与标准方式相反。 在检测欺诈行为或检查离群值的时候经常使用此选项。

网络选项选项卡

网络节点的“选项”选项卡包含一些用于定制输出图形的其他选项。

链接数。 以下选项用于控制在输出图形中显示的链接数。 有些选项，如 **以上弱链接** 和 **以上粗链接**，也可在输出窗口中使用。 同时，您也可以在最终图形中滑动控件调整显示的链接数量。

- **可显示的最大链接数**。 指定一个数字，说明要在输出图形中显示的最大链接数。 使用箭头调整该值。
- **仅显示以上链接**。 指定一个数字，说明要在 Web 中显示的连接必须达到的最小值。 使用箭头调整该值。
- **显示所有链接**。 无论最大或最小值是多少，都显示所有链接。 如果字段数量过多，选择此选项将延长处理时间。

如果记录很少，则废弃。 选择此选项将忽略受少量记录支持的连接。 通过在以下项中输入数字，设置此选项的阈值：**最小 记录数/行**。

如果记录很多，则废弃。 选择此选项将忽略受到较强支持的连接。 在以下项中输入数字：**最大 记录数/行**。

弱链接低于。 指定一个数字，用以区分弱连接（虚线）与常规连接（普通线条）的阈值。 所有低于该值的连接都被认为是弱连接。

以上强链接。 指定区分强连接（粗线）与常规连接（普通线条）的阈值。 所有高于该值的链接都被认为是强链接。

链接大小。 指定控制链接大小的选项：

- **链接大小连续变化**。 选择此选项将显示链接大小范围，从而反映由实际数据值产生的连接强度变化。
- **链接大小显示强/正常/弱类别**。 选择此选项将显示三种强度的连接 - 强、正常及弱。 三种类别的区分点可以在上面指定，也可以在最终图形中指定。

Web 显示。 选择 Web 显示的类型：

- **圆形布局**。 选择此选项将使用标准 Web 显示。
- **网络布局**。 选择此选项将使用一种算法，将最强的链接分在一起。 这样做的目的是以空间差别及加粗线条突出强链接。
- **定向布局**。 选择此选项将创建一个导向网络显示。此图使用“散点图”选项卡**目标字段**中的选择作为方向的集中点。
- **网格布局**。 选择此选项将创建一个以大小相同的网格形式显示的网络图。

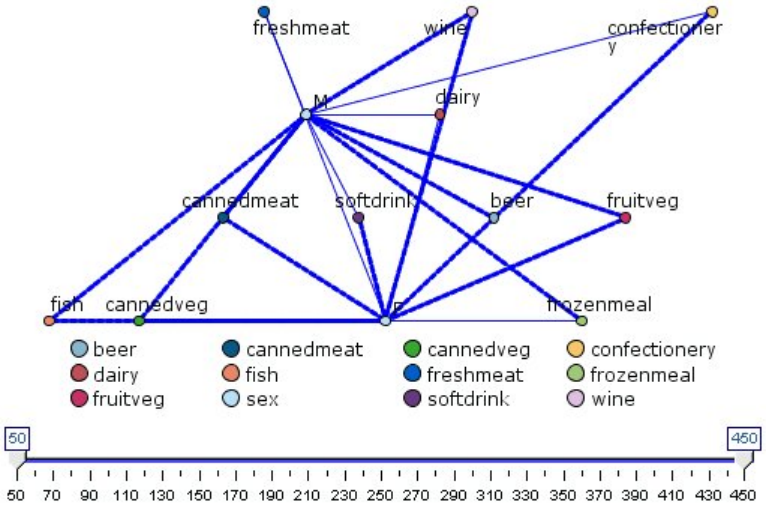


图 *37:* 显示从 *frozenmeal* 及 *cannedveg* 到其他杂货的强连接的网络图形

注: 当过滤所显示的链接（使用 Web 图形中的滑块或 Web 节点“选项”选项卡上的**仅显示上述链接**控件）时，最后可能会出现这样一种情况，即保持显示的所有链接都是一种值（换言之，全都是弱链接、中级链接或强链接，这取决于 Web 节点“选项”选项卡上的**以下弱链接**和**以上强链接**控件的定义）。 如果出现这种情况，所有这些链接在 Web 图形输出中都显示为中等宽度的线条。

网络外观选项卡

可以在创建图形前指定外观选项。

标题。 输入要用作图形标题的文本。

子标题。 输入要用作图形子标题的文本。

文字说明。 输入要用作图形文字说明的文本。

显示图注。 可以指定是否显示图例。 对于具有大量字段的散点图，隐藏图例将改善散点图的外观。

将标签用作节点。 可以将标签文本包括在每个节点中，而不是显示邻近的标签。 对于字段数量较少的散点图，此选项可以提高图表可读性。

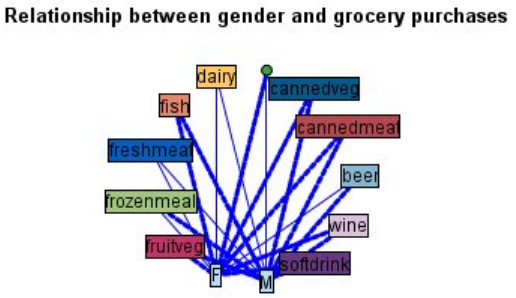


图 *38:* 将标签显示为节点的 *Web* 图形

使用网络图形

Web 节点用于显示两个或更多符号字段的值之间关系的紧密程度。 在图形中显示的链接以不同类型的线条表示，依次说明链接的强度不同。 例如，您可以使用 Web 节点来研究胆固醇水平、血压及可有效治疗病人疾患的药品之间的关系。

- 强链接以粗线条显示。 用以说明两个值之间关系紧密，应该进一步研究。
- 普通链接用普通粗细的线条显示。

- 弱链接以虚线显示。
- 如果在两个值之间没有显示线条，则说明两个值从不同时出现在同一记录中或这种同时出现的情况只发生在有限的记录中，即记录数量低于在“网络节点”对话框中指定的阈值。

在创建了 Web 节点后，有几个选项可用来调整图形显示并生成节点以备进一步分析。

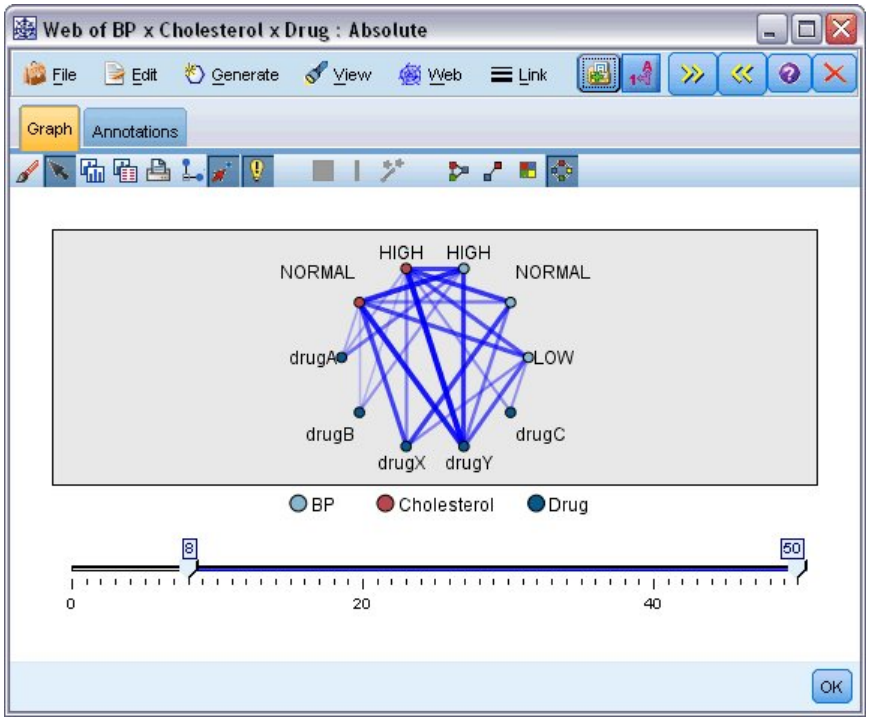


图 *39:* 网络图形说明一组紧密程度高的关系，如正常血压与 *DrugX* 及高胆固醇与 *DrugY*。

对于网络节点和导向网络节点，您都可以：

- 更改 Web 显示的布局。
- 隐藏一些点以简化显示。
- 更改控制线条样式的阈值。
- 突出显示值之间的线条，说明“选定”关系。
- 生成一个或多个“选定”记录的“选择”节点或与 Web 中的一种或多种关系关联的“派生标志”节点。

调整点

- 在某点上单击鼠标并将其拖动到新的位置来**移动**点。 将重新绘制 Web 以反映这个新的位置。
- 在 Web 中某点上单击鼠标右键并从上下文菜单中选择**隐藏**或**隐藏并重新规划**来**隐藏**点。 **隐藏**仅隐藏所选点及与其相关联的所有线条。 **隐藏并重新规划**将根据您所作的更改来重新绘制 Web。 所有手动移动都是未完成的。
- 通过选择图形窗口中“网络”菜单上的**全部显示**或**全部显示并重新计划**来**显示**所有隐藏的点。 选择 **全部显示并重新计划** 将重新绘制网络，即，将之前隐藏的点和它们的链接都包括进来。

选择或“突出显示”线

所选线以红色突出显示。

1. 要选择一条线，请左键单击该线。
2. 要选择多条线，进行以下操作之一：

- 使用光标在您希望选择线的点周围绘制圆形。
- 按下 Ctrl 键并左键单击您希望选择的单个线。

您可以单击图形背景，或从图形窗口中的 “Web” 菜单选择**清除选择**取消选择所有选定线。

使用其他布局查看 Web

在“网络”菜单中，选择**圆形布局**、**网状布局**、**定向布局**或**网格布局**更改图形布局。

打开或关闭链接滑块

在“视图”菜单中选择**链接滑块**。

选择或标记单个关系的记录

1. 对于感兴趣关系，可在表示该关系的线上单击鼠标右键。
2. 在上下文菜单中，选择 **生成链接的选择节点** 或 **生成链接的导出节点** 。

带有相应选项和指定条件的选择节点或导出节点将自动添加到流工作区中：

- “选择”节点选择指定关系中的所有记录。
- “派生”节点生成一个标志，为整个数据集中的记录一一标明所选关系是否存在，即标志值为真。 通过用下划线连接关系中的两个值来命名标志字段，例如 `LOW_drugC` 或 `drugC_LOW`。

选择或标记一组关系的记录

1. 选择 Web 显示中代表相关关系的线条。
2. 在图形窗口的“生成”菜单中选择**选择节点 ("与")**，**导出节点 ("或")**，或**导出节点 ("与")**和**导出节点 ("或")**。

- “或”节点将条件进行析取。 也就是说，只要在记录中存在所选关系中的任何一个，即可产生此节点。
- “与”节点将条件进行合取。 也就是说，只有当记录满足所有所选关系时，才可产生此节点。 如果所选关系中存在任何互斥的关系，则产生错误。

在选择完成之后，带有相应选项和指定条件的选择节点或导出节点将自动添加到流工作区中。

注: 当过滤所显示的链接（使用 Web 图形中的滑块或 Web 节点“选项”选项卡上的**仅显示上述链接**控件）时，最后可能会出现这样一种情况，即保持显示的所有链接都是一种值（换言之，全都是弱链接、中级链接或强链接，这取决于 Web 节点“选项”选项卡上的**以下弱链接**和**以上强链接**控件的定义）。 如果出现这种情况，所有这些链接在 Web 图形输出中都显示为中等宽度的线条。

调整网络阈值

在您创建了 Web 图形后，您可以通过工具栏滑块调整控制线条样式的阈值，以改变最小可见线条。 您还可以单击工具栏上的黄色的双箭头扩展 Web 图形窗口，以查看其它阈值选项。 然后单击 **控制** 选项卡查看其它选项。

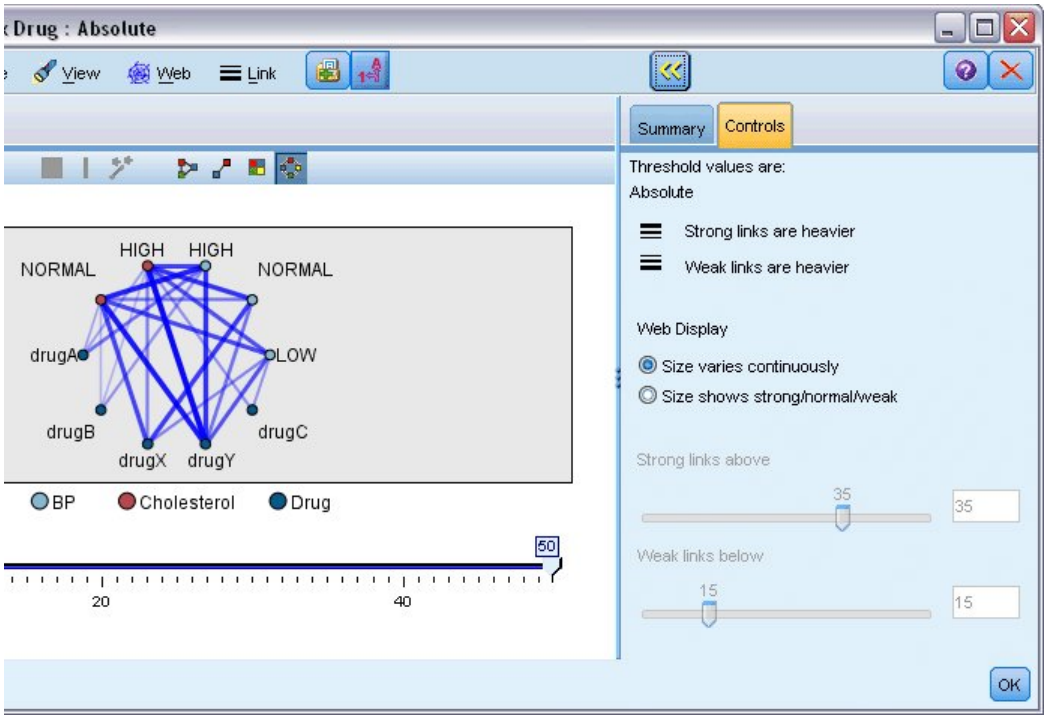


图 *40:* 扩展窗口中的选项主要针对显示和阈值

阈值为。 显示在 Web 节点对话框中创建节点时所选的阈值类型。

粗链接较重。 此选项缺省情况下处于选中状态，即查看字段间链接的标准方式。

弱链接较重。 选择此选项将使显示线条粗细所代表的含义与标准方式相反。 在检测欺诈行为或检查离群值的时候经常使用此选项。

Web 显示。 在输出图形中指定控制链接大小的选项：

- **大小连续变化**。 选择此选项将显示链接大小范围，从而反映由实际数据值产生的连接强度变化。
- **大小显示强/正常/弱**。 选择此选项将显示三种强度的连接 - 强、正常及弱。 三种类别的区分点可以在上面指定，也可以在最终图形中指定。

以上强链接。 指定区分强连接（粗线）与常规连接（普通线条）的阈值。 所有高于该值的链接都被认为是强链接。 使用滑块调整值或在字段中输入一个数字。

弱链接低于。 指定一个数字，用以区分弱连接（虚线）与常规连接（普通线条）的阈值。 所有低于该值的连接都被认为是弱连接。 使用滑块调整值或在字段中输入一个数字。

在您调整了 Web 的阈值之后，您可以通过 Web 图形工具栏上的 Web 菜单重新计划或重新绘制基于新阈值的 Web 显示。 在您确定了能使图形的意义最为明显的设置之后，您可以单击图形窗口“网络”菜单中的**更新父节点**来更新网络节点（也叫作父节点）中的原有设置。

创建 Web 汇总

您可以单击工具栏上的黄色双箭头按钮扩展 Web 图形窗口，以创建 Web 汇总文档，其中将列出粗、中等及弱链接。 然后单击 **汇总** 选项卡，查看每类链接的表。 可以通过每个表的切换按钮展开和折叠表。

要打印汇总，从网络图形窗口的菜单中选择：

文件 > 打印摘要

评估节点

“评估”节点为您提供了一个评估并比较预测模型，以选择最适合模型的便捷方法。 评估图表显示模型如何执行对特定结果的预测。 评估图表的工作原理是：根据预测值及预测的置信度排序记录、将记录分割为大小相

等的组（**分位数**）并按由高到低顺序为每个分位数绘制业务标准值。 在散点图中，将以单独的线条显示多个模型。

通过将具体值或值的范围定义为 **匹配**，处理结果。 通常，匹配表示相关的某类别（如向顾客销售）或某事件（如某项医疗诊断）成功执行。 您可以在对话框的“选项”选项卡上定义匹配标准，或使用以下描述的缺省匹配标准：

- **标志**输出字段是正向的，即匹配表现为 *true* 值。
- 对于**名义**输出字段，集合中的第一个值确定是否匹配。
- 对于**连续**输出字段，大于字段范围中点的值即为匹配。

有六种类型的评估图表，每一种类型针对不同的评估标准。

增益图

收益的定义是相对于全部匹配，发生于每个分位数中的匹配的百分比。 其计算方法为（分位数中的匹配数量/全部匹配数量）× 100%。

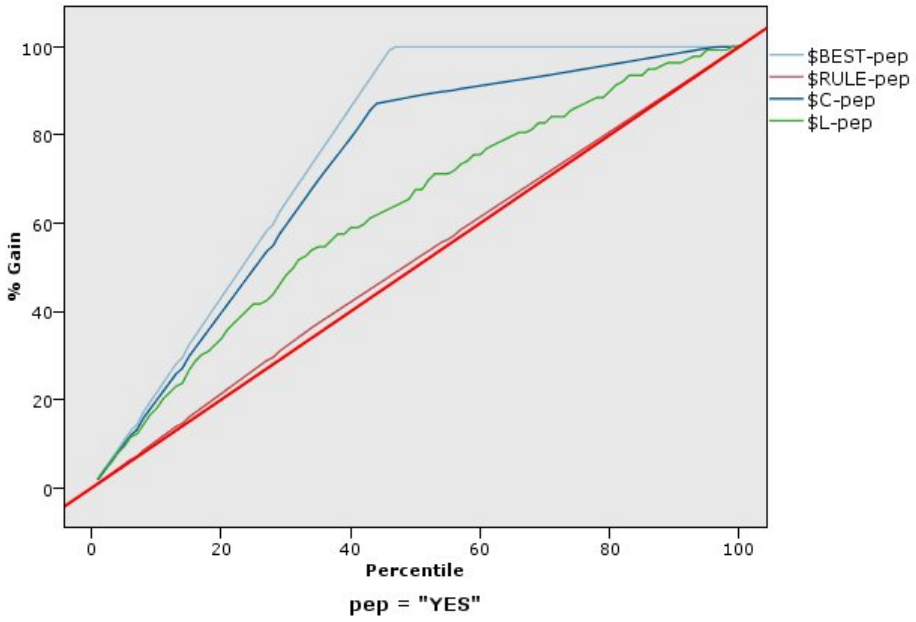


图 *41:* 显示带有基线、最佳线及业务规则的收益图（累积）

提升图

提升将每个分位数中匹配记录的百分比与在全部训练数据中匹配的百分比进行比较。 其计算方式为（在分位数中的匹配/在分位数中的记录）/（全部匹配/全部记录）。

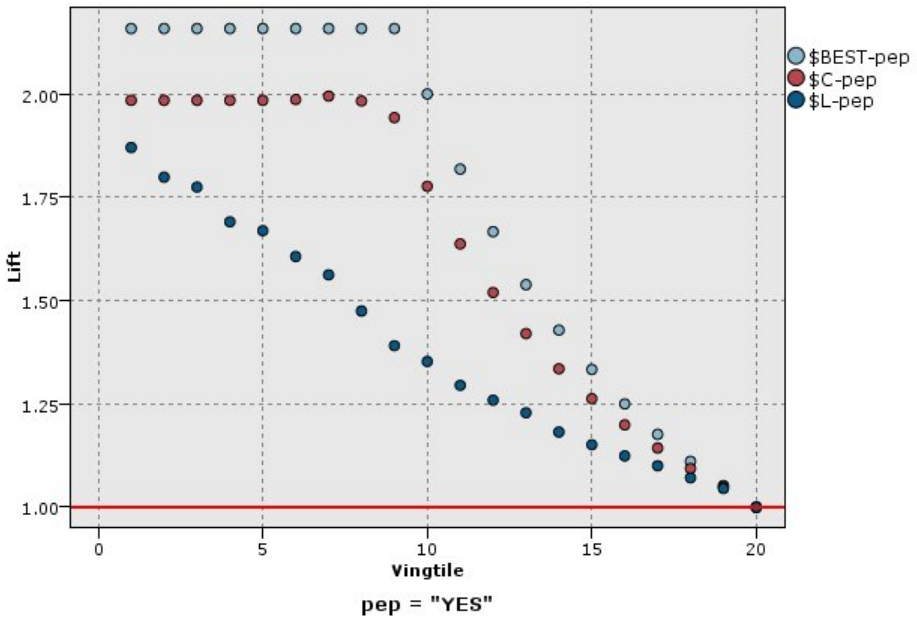


图 *42:* 使用点和最佳线的提升图（累积）

响应图

响应即分位数中，匹配记录的比例。 响应的计算方法为（分位数中的命中数/分位数中的记录数）× 100%。

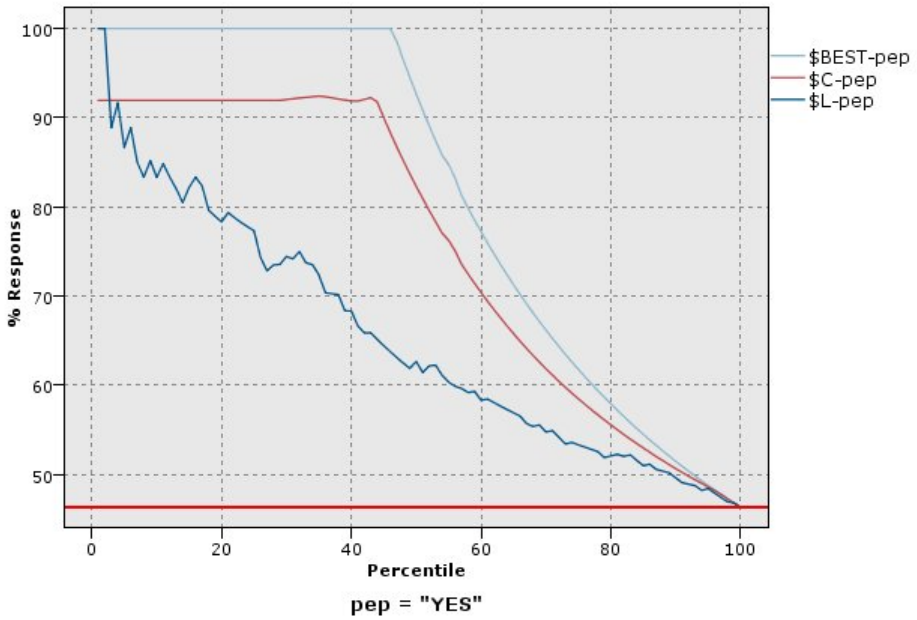


图 *43:* 具有最佳线的响应图（累积）

利润图

利润等于每个记录的 **收入** 减去该记录的 **成本** 。 也就是说，分位数的利润就是位于该分位数内的所有记录的利润总和。 这里假定收入仅应用于匹配项，但成本可应用于所有的记录。 利润及成本都可以是固定的，也可以由数据中的字段决定。 其计算方法为（分位数中所有记录收入的总和 - 分位数中所有记录成本的总合）。

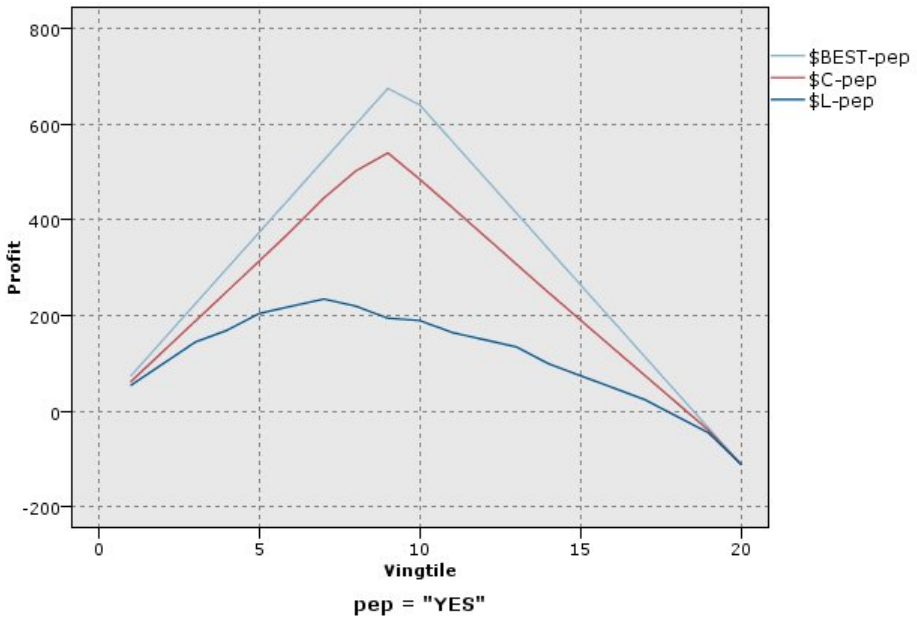


图 *44:* 具有最佳线的利润图（累积）

投资回报图

投资回报 (ROI) 也需要确定收入和成本，从这一点上来说，它与利润相同。 ROI 将分位数的成本和利润进行比较。 其计算方法为（分位数利润/分位数成本）× 100%。

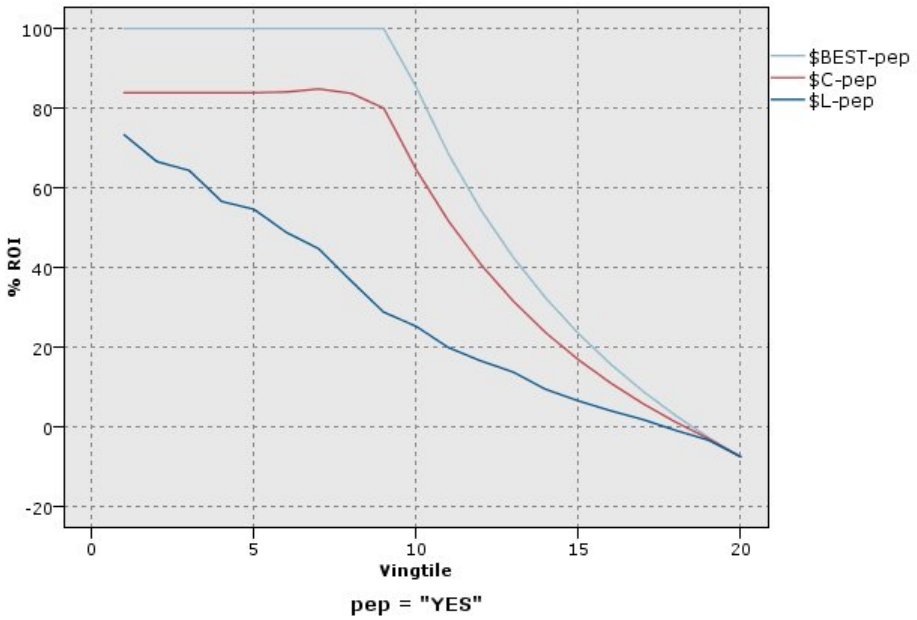


图 *45:* 具有最佳线的投资回报图（累积）

ROC 图表

只能将 ROC（受试者工作特征）与二元分类器配合使用。 ROC 可用于根据性能显示、组织和选择分类器。 ROC 图表绘制分类器的真阳性率（或灵敏度）与假阳性率的比率。 ROC 图表描述了收益（真阳性）与成本（假阳性）之间的相对平衡。 真阳性是一个匹配实例，并且分类为匹配项。 因此，真阳性率按照真阳性数/实际匹配的实例数进行计算。 假阳性是一个未匹配实例，并且分类为匹配项。 因此，假阳性率按照假阳性数/实际未匹配的实例数进行计算。

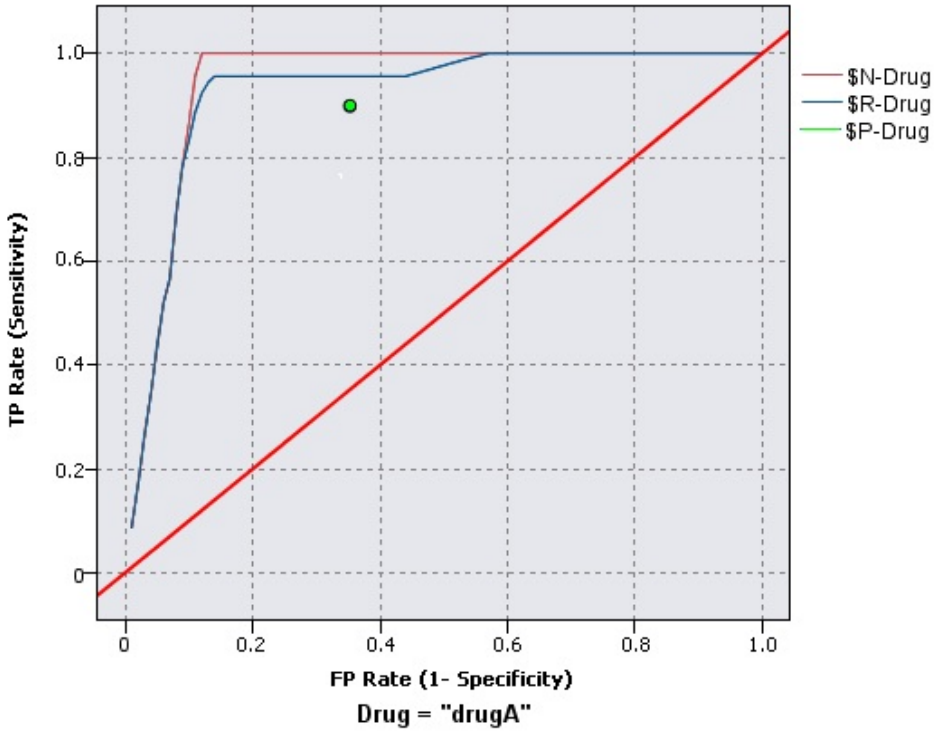


图 *46:* 具有最佳线的 *ROC* 图表

评估图表也可以累积，因此每个点等于相应分位数的值加上所有更高分位数的值。 累积图表通常能够更好的表现模型性能，而非累积图则更有利于指出模型中可能存在问题的地方。

注: “评估”节点不支持在字段名称中使用逗号。 如果存在包含逗号的字段名称，您必须移除逗号或者将字段名称括在引号内。

评估散点图选项卡

图表类型。 选择下列其中一种类型：**增益图**、**响应图**、**提升图**、**利润图**、**ROI**（投资收益率）或 **ROC**（受试者工作特征）。

累积散点图。 选择此选项将创建累积图表。 累积图表中绘制的值代表每个分位数与所有更高分位数的和。（**累积图**不适用于 ROC 图表。）

包含基线。 选择此选项将在散点图中包含基线，表示匹配值的完全随机分布（此时置信度并不相关）。（**包含基线**不适用于利润图及投资回报图。）

包含最佳线。 选择此选项将在散点图中包含最佳线，代表最佳置信度（即匹配项 = 观测值的 100%）。（**包含最佳线**不适用于 ROC 图表。）

对所有图表类型使用利润标准。 选择此选项可在计算评估度量时使用利润标准（成本、收入和权重）而不是正常匹配计数。 对于具有特定数字目标的模型（例如用于预测从客户那里获取的供应收入的模型），目标字段的值提供了一种比匹配项数更好的模型性能度量方式。 选择此选项将对增益图、响应图和提升图启用**成本**、**收入**和**权重**字段。 要对这三种图表类型使用利润条件，建议将**收入**设置为目标字段，将**成本**设置为 0.0，从而使利润等于收入，并且建议指定一个使所有记录都计为匹配项的用户定义“ture”匹配条件。 （**将利润条件用于所有图表类型**不适用于 ROC 图表。）

使用以下内容查找预测值/预测变量字段。 选择**模型输出字段元数据**以使用其元数据搜索图形中的预测字段，或者选择**字段名称格式**按名称进行搜索。

绘制评分字段。 选择此复选框可启用评分字段选择器。 然后选择一个或多个范围或连续型评分字段；即，不是严格预测模型但可用于根据匹配倾向程度对记录排序的字段。 “评估”节点可以将一个或多个评分字段的任意组合与一个或多个预测模型进行对比。 一个典型的示例是将几个 RFM 字段和最佳预测模型进行对比。

目标。 使用字段选择器选择目标字段。 选择任意实例化标志或具有两个或多个值的名义字段。

注意：此目标字段只适用于评分字段（预测模型会定义自己的目标字段），并且如果在“选项”选项卡上设置了定制匹配标准，那么将省略目标字段。

按分区分割。 如果要用分区字段将记录分割为训练、测试及验证样本，请选择此选项，以便为每个分区显示一个单独的评估图表。 有关更多信息，请参阅主题 第 131 页的『分区节点』。

注意：按分区分割时，将不对分区字段中具有空值的记录进行评估。 由于分区节点不生成空值，因此，如果使用分区节点，则这永远不会成为问题。

散点图。 从下拉列表中选择要在图表中绘制的分位数的大小。 选项包括 **四分位数** 、 **五分位数** 、 **十分位数** 、 **二十分位数** 、 **百分位数** 和 **千分位数** 。 （**散点图**不适用于 ROC 图表。）

样式。 选择**线**或**点**。

对于除 ROC 图表以外的所有图表类型，提供了使您可以指定成本、收入和权重的更多控件。

- **成本**。 指定与每个记录关联的成本。 您可以选择 **固定** 或 **可变** 成本。 对于固定成本，请指定成本值。 对于可变成本，请单击“字段选择器”按钮，将某个字段选择为成本字段。 （**成本**不适用于 ROC 图表。）
- **收入**。 指定与表示匹配项的每个记录关联的收入。 您可以选择 **固定** 或 **可变** 成本。 对于固定收入，请指定收入值。 对于可变收入，请单击“字段选择器”按钮，将某个字段选择为收入字段。 （**收入**不适用于 ROC 图表。）
- **权重**。 如果数据中的记录代表多个单元，那么可以使用频率权重来调整结果。 使用**固定**或**可变**加权，指定与每个记录关联的加权。 对于固定加权，请指定加权值（每个记录的单元数）。 对于可变加权，请单击“字段选择器”按钮，将某个字段选择为权重字段。 （**权重**不适用于 ROC 图表。）

“评估选项”选项卡

评估图表的“选项”选项卡提供了定义在图表中显示的匹配、评分标准及业务规则的灵活性。 您可以设置这些选项，以导出模型评估结果。

用户定义的匹配项。 选择此选项可指定一个用于指示匹配项的定制条件。 此选项更适合于定义相关结果，而不是从目标字段类型和值的顺序中推测结果。

- **条件**。 如果选择了上面的**用户定义的匹配项**，那么您必须指定一个 CLEM 表达式作为匹配条件。 例如，`@TARGET = "YES"` 是一个有效条件，指示目标字段的值 *Yes* 将在评估中计为命中。 指定的条件将用于所有目标字段。 要创建一个条件，请在字段中键入或使用表达式构建器生成条件表达式。 如果数据没有实例化，则您可以直接从表达式构建器重插入值。

用户定义的分数。 选择此选项后，用户可以指定一个在将观测值分配到分位数之前，对观测值评分的条件。缺省分数将根据预测值及置信度计算。 使用“表达式”字段创建一个自定义评分表达式。

- **表达式**。 指定用于评分的 CLEM 表达式。 例如，如果取值范围在 `0 –1` 之间的数字输出是按如下顺序排列的：较小的值好于较大的值，那么您可以定义一个匹配项 `@TARGET < 0.5`，且相应的评分为 `1 – @PREDICTED`。 评分表达式必须返回一个数字值。 要创建一个条件，请在字段中键入或使用表达式构建器生成条件表达式。

包含业务规则。 选择此选项后，您可指定满足相关标准的规则条件。 例如，您可能希望为 `mortgage = "Y" and income >= 33000` 的所有情况显示一条规则。 业务规则将在图表中绘制出来，在键字段中，将被标记为 *Rule* 。 （对于 ROC 图表，**包括业务规则**不受支持。）

- **条件**。 指定用于定义输出图表中业务规则的 CLEM 表达式。 请直接在字段中输入或使用表达式构建器生成条件表达式。 如果数据没有实例化，则您可以直接从表达式构建器重插入值。

将结果导出到文件。 选择此选项后，用户可将模型评估结果导出到分隔的文本文件中。 您可以读取此文件，对计算结果执行特定的分析。 为导出设置如下选项：

- **文件名**。 输入输出文件的文件名。 单击省略号按钮 **(...)** ，打开需要的文件夹。
- **分隔符**。 输入一个字符（如逗号或空格）作为字段分隔符。

包含字段名。 选择此选项可使字段名称在输出文件的第一行显示出来。

每条记录后新建一行。 选择此选项后，每条记录将另起一个新行。

评估外观选项卡

可以在创建图形前指定外观选项。

标题。 输入要用作图形标题的文本。

子标题。 输入要用作图形子标题的文本。

文本。 接受自动生成的文本标签，或选择**自定义**指定标签。

X 标签。 接受自动生成的 *x* 轴（水平）标签，或者选择**定制**来指定标签。

Y 标签。 接受自动生成的 *y* 轴（垂直）标签，或者选择**定制**来指定标签。

显示网格线。 此选项缺省情况下处于选中状态，它显示散点图或图形背后的网格线，以便更轻松地确定区域和带状分界值点。 网格线通常用白色显示，但图形背景为白色时除外；此情况下，网格线用灰色显示。

读取模型评估结果

评估图表的解读方法在某种程度上取决于图表类型，但是，有些特点是所有评估图表共有的。 对于累积图表而言，线位置越高（特别是当图表左侧线位置高时）表明模型越优秀。 在很多情况下，在比较多个模型时，线会发生交叉。因此，一个模型的线可能会在某处较高；但在图表另一处，另一个模型的线较高。 如果出现这种情况，您需要考虑要哪个部分的样本（由此确定 *x* 轴上点的位置），以确定选择哪个模型。

大多数非累积图表都极其相似。 优秀模型的非累积图应该是左侧较高，右侧较低。 （如果非累积图呈锯齿状，您可以减少分位数的数量，重新绘制并执行图形，由此获得较为平滑的图形。） 线在图表左侧偏低而在右侧偏高，可能意味着模型预测结果较差的区域。 一条在整个图形中平直的线条则说明此模型基本不能提供任何信息。

收益图。 累积收益图的线从左至右的走势通常是从 0% 到 100%。 对于良好的模型，收益图表向 100% 突增，然后趋于平稳。 无法提供有用信息的模型将呈对角线状，即从左下角到右上角（选择了 **包含基线** 后将显示类似图表）。

提升图。 累积效益图的线从左至右的走势通常为：起始于大于 1.0 的值，并渐渐下降，直到接近 1.0。 图表的右侧边缘表示整个数据集，因此累积分位数的匹配与数据中的匹配的比例为 1.0。 对于优秀模型的提升图，其线开始于图表左侧大于 1.0 的值，且在向右移动的过程中，始终保持在较高的水平；然后，在图表右侧，向 1.0 的方向迅速下降。 如果模型不能提供任何信息，那么其线在整个图形中将始终围绕在 1.0 左右。（如果选择了 **包含基线**，一条值为 1.0 的水平线将显示在图表中供您参考。）

响应图。 累积响应图通常与效益图极其类似，只在尺度标准方面有所区别。 通常，响应图开始于接近 100% 之处，并逐渐下降，最终将在延伸至图表右侧边缘时达到整体响应率（全部匹配/全部记录）。 对于优秀模型的响应图，其线开始于图表左侧接近或等于 100% 的值，且在向右移动的过程中，始终保持在较高的水平；然后，在图表右侧，向整体响应率的方向迅速下降。 如果模型不能提供任何信息，那么其线在整个图形中将始终围绕在整体响应率左右。 （如果选择了 **包含基线**，一条值相当于整体响应率的水平线将显示在图表中供您参考。）

利润图。 累积利润图线从左至右的走势代表随着所选样本数量的增加，利润总和的增长。 利润图通常开始于 0 附近，并在向右延伸的过程中，稳步增长直至在图表中部到达峰值或保持较高的值；随后，在向右侧边缘延伸的过程中，逐渐下降。 优秀模型的利润图将在图表中部某处显示定义良好的峰值。 而无法提供任何信息的模型，其线相对而言比较平直，也可能由于成本/收入结构的不同增加、降低或保持不变。

投资回报图。 累积投资回报 (ROI) 图通常与响应图及提升图类似，只有在尺度标准方面有所差别。 投资回报图通常开始于大于 0% 的值，并逐渐下降，直到达到整个数据集的整体 ROI（可能为负）。 对于优秀模型的投资回报图，其线开始于图表左侧大于 0% 的值，且在向右移动的过程中，始终保持在较高的水平；然后，在图表右侧，向整体 ROI 的方向迅速下降。 如果模型不能提供任何信息，则其线在整个图形中将始终围绕在整体 ROI 左右。

ROC 图表 ROC 曲线的形状通常为累积增益图。 该曲线开始于 (0,0) 坐标，结束于 (1,1) 坐标，方向为从左到右。 图表曲线朝 (0,1) 坐标急剧上升随后趋于平稳，这表示分类器较好。 将实例随机分类为匹配项或未匹配项的模型将呈对角线状，即从左下角到右上角（如果选择了**包含基线**，那么此对角线将显示在图表中）。 如果未提供模型的置信度字段，那么模型将绘制为单个点。 具有最优分类阈值的分类器位于最接近 (0,1) 坐标的位置或者图表的左上角。 此位置表示正确分类为匹配项的实例数较多，并且错误分类为匹配项的实例数较少。 对角线上方的点表示较好的分类结果。 对角线下方的点表示较差的分类结果，这些结果比对实例进行随机分类的结果更差。

使用评估图表

用鼠标检查评估图表的方法与在直方图或集合图中相同。 *x* 轴表示指定分位数（如二十分位数或十分位数）的模型分数。

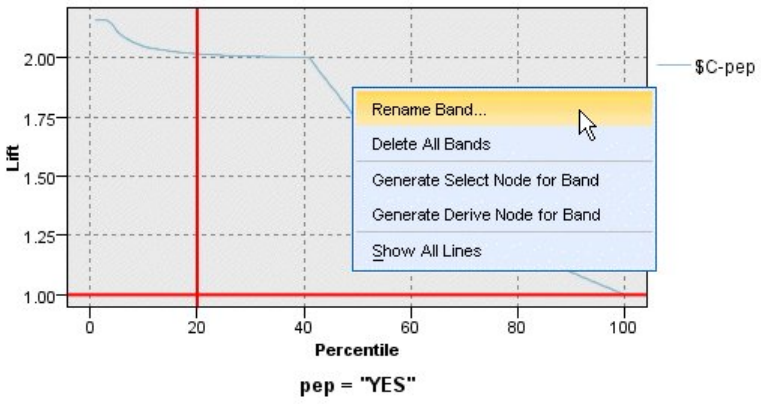


图 *47:* 处理评估图表

您可以像处理直方图那样，使用分割标志显示将轴自动分为等宽带状区域的选项，以将评估图表的 *x* 轴分为带状区域。 有关更多信息，请参阅主题 第 208 页的『探索图形』。 您可以选择“编辑”菜单的 **图形带状区域** 来手动编辑带状区域的边界。

在创建了评估图表、定义了带状区域并查看了结果之后，您可以使用“生成”菜单上的选项及上下文菜单根据图形中的选择自动创建节点。 有关更多信息，请参阅主题 第 214 页的『从图形中生成节点』。

从评估图表中生成节点时，将提示您选择图表中所有可用模型中的一个。

选择一个模型并单击 **确定** 在流工作区中生成新节点。

“地图可视化”节点

“地图可视化”节点可以接受多个输入连接，并在地图上将地理空间数据显示为一系列层。 每个层都是单个地理空间字段；例如，底层可能是国家或地区的地图，在其之上可能存在一个道路层、一个河流层和一个城镇层。

虽然大部分地理空间数据集通常包含单个地理空间字段，但是单个输入中存在多个地理空间字段时，您可以选择要显示的字段。 同一个输入连接中的两个字段不能同时显示；但是，您可以复制并粘贴传入连接并显示这两个连接中的不同字段。

地图可视化绘图选项卡

层

这个表显示有关地图节点的输入的信息。 各个层的顺序指示了执行该节点时各个层在地图预览和可视输出中的显示顺序。 表中的顶行是“顶层”，底行是“底层”；换而言之，每一层在地图上都显示在表中正好位于其下方的层前面。

注: 如果表中的层包含三维地理空间字段，那么将仅绘制 x 和 y 轴。 将忽略 z 轴。

Name
自动为每层创建名称，并且使用以下格式组成：`tag[source node:connected node]`。 缺省情况下，标记显示为数字，1 表示所连接的第一个输入，2 表示第二个输入，依此类推。 有需要时，请按**编辑层**按钮，以便在“**更改地图层选项**”对话框中更改标记。 例如，可以将标记更改为“道路”或“城市”以反映数据输入。

类型
显示已选择作为层的地理空间字段的测量类型图标。 如果输入数据包含多个具有地理空间测量类型的字段，那么缺省选择将使用以下排序顺序：

1. 点
2. 线串
3. 多边形
4. 多点
5. 多线串
6. 多多边形

注: 如果存在两个具有同一测量类型的字段，那么缺省情况下将选中第一个字段（按名称的字母顺序排列）。

symbol

注: 仅对于“点”和“多点”字段才会填写此列。

显示用于“点”或“多点”字段的符号。 有需要时，请按**编辑层**按钮，以便在“**更改地图层选项**”对话框中更改符号。

COLOR

显示已选择用于在地图上表示层的颜色。 有需要时，请按**编辑层**按钮，以便在“**更改地图层选项**”对话框中更改颜色。 根据测量类型不同，此颜色将应用于不同的项。

- 对于“点”或“多点”，此颜色将应用于层的符号。
- 对于“线串”和“多边形”，此颜色将应用于整个形状。 多边形始终具有黑色轮廓；此列中显示的颜色是用于填充形状的颜色。

预览

此窗格显示**层**表中当前选择的输入的预览。 此预览将考虑层顺序、符号、颜色以及任何其他与层相关联的显示设置，有可能时，在这些设置每次更改时更新显示。 如果您在流中的其他位置更改了详细信息，例如更改了要用作层的地理空间字段，或者修改了详细信息（例如相关联的汇总函数），那么可能需要单击**刷新数据**按钮以更新预览。

在运行流之前，请使用**预览**来设置显示设置。 为了避免因使用大型数据集而造成时间延迟，预览功能将对每个层进行采样并根据前 100 个记录来创建显示。

更改地图层

您可以使用“**更改地图层选项**”对话框来修改“地图可视化”节点的**图**选项卡上显示的任何层的各种详细信息。

输入详细信息

tag

缺省情况下，标记是数字；您可以将此数字替换为更有意义的标记，以帮助在地图上标识层。 例如，此标记可以是数据输入的名称，例如“城市”。

层字段

如果输入数据中存在多个地理空间字段，请使用此选项来选择要在地图上显示为层的字段。

缺省情况下，可供选择的层按以下顺序排列。

- 点
- 线串
- 多边形
- 多点
- 多线串
- 多多边形

显示设置

进行六边形分箱

注: 此选项只影响点字段和多点字段。

六边形（六角形）分箱根据临近点的 x 和 y 坐标将这些点组合成单个点以显示在地图上。 这个单点显示为六角形，但实际上呈示为多边形。

由于六角形呈示为多边形，因此所有开启了六边形分箱的点字段都将被视为多边形。 这意味着，如果在地图节点对话框中选择了**按类型排序**，那么所有应用了六边形分箱的点层都将呈示在多边形层上方，但呈示在线串层和点层下方。

如果将六边形分箱用于多点字段，那么会先通过对多点值进行分箱来计算中心点，将该字段转换为点字段。 中心点用于计算六边形分箱。

汇总

注: 仅当选中了**使用六边形分箱**复选框，并选中了**覆盖**时，此列才可用。

如果您为使用了六边形分箱功能的点层选中了**覆盖**字段，那么该字段中所有的值都必须针对该六边形内的所有点进行汇总。 请为任何要应用于地图的覆盖字段指定汇总函数。 可用的汇总函数视测量类型而定。

- 对于“实数”或“整数”存储，用于“连续”测量类型的汇总函数：
  - 总和
  - 平均值
  - 最小值
  - 最大值
  - 中位数
  - 第一个四分位数
  - 第三个四分位数
- 对于“时间”、“日期”或“时间戳记”存储，用于“连续”测量类型的汇总函数：
  - 平均值
  - 最小值
  - 最大值
- 用于“名义”或“分类”测量类型的汇总函数：
  - 方式
  - 最小值
  - 最大值
- 用于“标志”测量类型的汇总函数：
  - True（如果任何项为 true）
  - False（如果任何项为 false）

COLOR

使用此选项可以选择标准颜色（此颜色将应用于地理空间字段的所有特征）或覆盖字段（此字段根据数据中另一字段的值对各个特征进行着色）。

如果您选择了**标准**，那么可以从“**用户选项**”对话框的“显示”选项卡中**图表类别颜色顺序**窗格中显示的调色板中选择颜色。

如果您选择了**覆盖**，那么可以从选择作为**层字段**的地理空间字段所在的数据源中选择任意字段。

- 对于名义字段或分类覆盖字段，可以从中进行选择的调色板与针对**标准**颜色选项显示的调色板相同。

- 对于连续字段和有序覆盖字段，将显示另一个下拉列表，您可以从中选择颜色。 您选择颜色时，将通过根据该连续字段或有序字段中的值改变该颜色的饱和度来应用该覆盖。 最大的值将使用从下拉列表中选择的颜色，较小的值由相应的较低饱和度显示。

symbol

注: 仅对点测量类型和多点测量类型启用。

使用此选项可以选择是使用**标准**符号（此符号将应用于地理空间字段的所有记录），还是使用**覆盖**符号（这将根据数据中另一个字段的值来更改各个点的符号图标）。

如果您选择了**标准**，那么可以从下拉列表中选择其中一个缺省符号，用于在地图上表示点数据。

如果您选择了**覆盖**，那么可以从选择作为**层字段**的地理空间字段所在的数据源中选择任意名义字段、有序字段或分类字段。 对于覆盖字段中的每个值，都将在地图上显示不同的符号。

例如，您的数据可能包含表示商店位置的点字段以及可能是商店类型字段的覆盖项。 在此示例中，所有饮食店在地图上可能由十字符号标识，而所有电器店由方形符号标识。

大小

注: 仅对点、多点、线串和多线串测量类型启用。

使用此选项可以选择是使用**标准**大小（此大小将应用于地理空间字段的所有记录），还是使用**覆盖**大小（这将根据数据中另一个字段的值来更改符号图标大小或线宽）。

如果您选择了**标准**，那么可以选择像素宽度值。 可用的选项包括 1、2、3、4、5、10、20 或 30。

如果您选择了**覆盖**，那么可以从选择作为**层字段**的地理空间字段所在的数据源中选择任意字段。 根据所选字段的值不同，线宽或点大小也将有所变化。

透明度

使用此选项可以选择是使用**标准**透明度（此透明度将应用于地理空间字段的所有记录），还是使用**覆盖**透明度（这将根据数据中另一个字段的值来更改符号、线条或多边形的透明度）。

如果您选择了**标准**，那么可以从一组由 0%（不透明）开始并以 10% 递增到 100%（透明）的透明度级别中进行选择。

如果您选择了**覆盖**，那么可以从选择作为**层字段**的地理空间字段所在的数据源中选择任意字段。 在地图上，对于覆盖字段中的每个值，将显示不同的透明度级别。 此透明度将应用于您从颜色下拉列表中为点、线或多边形选择的颜色。

数据标号

注: 如果选中了**进行六边形分箱**复选框，那么此选项不可用。

使用此选项可以选择地图上要用作数据标签的字段。 例如，如果应用于多边形层，那么数据标签可能是名称字段，其中包含每个多边形的名称。 如果您选择了名称字段，那么这些名称将显示在地图上。

“地图可视化外观”选项卡

可以在创建图形前指定外观选项。

标题。 输入要用作图形标题的文本。

子标题。 输入要用作图形子标题的文本。

文字说明。 输入要用作图形文字说明的文本。

t-SNE 节点

t 分布随机邻域嵌入 (t-Distributed Stochastic Neighbor Embedding, t-SNE)© 是用于将高维数据可视化的工具。 其将数据点亲缘关系转换为可能性。 原始空间中的亲缘关系通过高斯联合概率表示，并且内嵌空间中的亲缘关系通过 Student t 分布表示。 这可使 t-SNE 对于本地结构特别敏感，而且与现有技术相比具有以下优点：[1]

- 在单个映射的多个尺度上揭示结构

- 揭示位于多个不同集群中的数据
- 降低在点在中心拥挤在一起的趋势

t-SNE 节点在 SPSS Modeler 中使用 Python 进行实现并且需要 `scikit-learn`© Python 库。 有关 t-SNE 和 scikit-learn 库的详细信息，请参阅：

- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE
- https://scikit-learn.org/stable/modules/manifold.html#t-sne

节点选用板上的 Python 选项卡包含此节点和其他 Python 节点。 “图形”选项卡上还提供 t-SNE 节点。

[1] 引用：

van der Maaten, L.J.P.; Hinton, G. "Visualizing High-Dimensional Data using t-SNE." Journal of Machine Learning Research. 9:2579-2605, 2008.

van der Maaten, L.J.P. "t-Distributed Stochastic Neighbor Embedding."

van der Maaten, L.J.P. "Accelerating t-SNE using Tree-Based Algorithms." Journal of Machine Learning Research. 15(Oct):3221-3245, 2014.

t-SNE 节点专家选项

根据想要针对 t-SNE 节点设置的选项，选择**简单**方式或**专家**方式。

可视化类型。 选择 **2D** 或 **3D** 以指定是将图像绘制为二维还是三维。

方法。 选择 **Barnes Hut** 或 **Exact**。 缺省情况下，梯度计算算法使用 Barnes-Hut 近似值，其运行速度必须大幅快于 Exact 方法。 Barnes-Hut 近似值允许将 t-SNE 技术应用于大型现实世界数据集。 Exact 算法在避免最近邻元素错误方面更好一些。

初始化。 对于嵌套初始化选择**随机**或 **PCA**。

目标字段。 选择目标字段以显示为输出图形上的颜色映射图。 如果此处未指定目标字段，那么图像将使用一种颜色。

优化

困惑度。 困惑度与其他各种学习算法中使用的最近邻元素数量相关。 通常，数据集越大，需要的困惑度越大。 请考虑选择 **5** 和 **50** 之间的值。 缺省值为 **30**，范围为 **2 - 9999999**。

早期夸大。 此设置控制原始空间中的自然聚类将在内嵌空间中的紧密程度以及两者之间的空间量。 缺省值为 **12**，范围为 **2 - 9999999**。

学习速率。 如果学习速率太高，那么数据可能看起来好像一个“球”，其中任意点与其最近的邻域大致等距。 如果学习速率太低，那么大多数点可能看起来好像压缩在一个具有极少离群值的密集云中。 如果成本函数陷入错误的局部最小值中，那么提高学习速率可能会有所帮助。 缺省值为 **200**，范围为 **0 - 9999999**。

最大迭代次数。 优化的最大迭代次数。 缺省值为 **1000**，范围为 **250 - 9999999**。

角度大小。 从一个点度量的远距离节点的角度大小。 输入 **0** 到 **1** 之间的值。 缺省值为 **0.5**。

随机种子(D)

设置随机种子值。 选择此信息并单击**生成**可以生成由随机数字生成器使用的种子。

优化停止条件

不含进度的最大迭代次数。 在停止优化之前要执行的不含进度的最大迭代次数，在 250 次包含早期夸大 (early exaggeration) 的初始迭代后将使用该迭代数。 请注意，每隔 50 次迭代后才会检查进度，因此该值舍入为 50 的下一个倍数。 缺省值为 **300**，范围为 **0 - 9999999**。

最小梯度标准值。 如果梯度标准值低于此最低阈值，那么优化将停止。 缺省值为 **1.0E-7**。

度量。 在计算特征数组中实例之间的距离时要使用的度量。 如果度量是字符串，那么它必须是 `scipy.spatial.distance.pdist` 针对其度量参数允许的选项之一，或是

pairwise.PAIRWISE_DISTANCE_FUNCTIONS 中列出的度量。 选择其中一个可用度量类型。 缺省值为 **euclidean**。

当记录数大于以下值时。 为大型数据集指定一种绘制方法。 可以指定数据集大小上限，或使用缺省的 2000 点。 如果选择 **分隔** 或 **抽样** 选项，则处理大数据集的性能将显著提高。 另外，您也可以选择 **使用所有数据**，但必须要注意，这一选项可能大幅降低软件的执行效率。

- **分级**。 选择此选项可对所包含记录数超过指定数字的数据集进行分级。 “分级”使图形在实际绘制前被分散在较小的网格中，并计算在每个单元格中将出现的连接数。 在最终图形中，每个网格中的分级矩心处将使用一个连接（该连接即代表分级中所有连接点位置的平均数）。
- **样本**。 选择此选项将随机抽取指定记录数的数据样本。

下表显示 SPSS Modeler t-SNE 节点对话框的“专家”选项卡上的设置与 Python t-SNE 库参数之间的关系。

表 *37:* 映射到 *Python* 库参数的节点属性

SPSS Modeler 设置	**脚本名称（属性名称）**	**Python t-SNE 参数**
方式	mode_type	
可视化类型	n_components	n_components
Method	method	method
嵌套初始化	init	init
目标	target_field	target_field
困惑度	perplexity	perplexity
早期夸大	early_exaggeration	early_exaggeration
学习速率	learning_rate	learning_rate
最大迭代次数	n_iter	n_iter
角度大小	angle	angle
设置随机种子	enable_random_seed	
随机种子(D)	random_seed	random_state
不含进度的最大迭代次数	n_iter_without_progress	n_iter_without_progress
最小梯度标准值	min_grad_norm	min_grad_norm
使用多个困惑度执行 t-SNE	isGridSearch	

t-SNE 节点输出选项

在**输出**选项卡上指定 t-SNE 节点输出的选项。

输出名称。 指定在节点运行时生成的输出的名称。 如果选择**自动**，那么将自动设置输出的名称。

输出到屏幕。 选择此选项以在新窗口中生成并显示输出。 这还会将输出添加到输出管理器。

输出到文件。 选择此选项可将输出保存到文件。 执行此操作将启用**文件名**和**文件类型**字段。 如果要使用其他字段创建绘图以进行比较，或者要使用输出文件的输出作为分类或回归模型中的预测变量，那么 t-SNE 节点需要此输出文件的访问权。 t-SNE 模型创建 x、y（和 z）坐标字段的结果文件，使用“固定文件”源节点可非常轻松地对其进行访问。

访问和绘制 t-SNE 数据

如果您使用**输出到文件**选项将 t-SNE 输出保存到文件，那么可以使用其他字段创建绘图以进行比较，或者使用输出作为分类或回归模型中的预测变量。 t-SNE 模型创建 x、y（和 z）坐标字段的结果文件，使用“固定文件”源节点可非常轻松地对其进行访问。 此部分提供示例信息。

1. 在 t-SNE 节点对话框中，打开**输出**选项卡。

2. 选择**输出到文件**并输入文件名。 使用缺省 HTML 文件类型。 运行模型时，这将在输出位置中生成三个输出文件：
   - 文本文件 (`result_xxxxxx.txt`)
   - HTML 文件（所指定的文件名）
   - PNG 文件 (`tsne_chart_yyyyyy.png`)

   文本文件将包含所需的数据，但是由于技术原因，它可能为标准或科学格式。 如果它是科学格式 (1.11111111e+01)，那么需要创建可识别该格式的新流：

当文本文件为科学数字格式时访问 t-SNE 绘图数据

1. 创建新流（**文件** > **新建流**）。
2. 转至**工具** > **流属性** > **选项**，选择**数字格式**，然后针对数字显示格式选择**科学 (#.###E+##)**。
3. 向工作区中添加“固定文件”源节点，并在“文件”选项卡上使用以下设置：
   - 跳过标题行：1
   - 记录长度：54
   - tSNE_x 开始：3，长度：16
   - tSNE_y 开始：20，长度：16
   - tSNE_z 开始：36，长度：16
4. 在“类型”选项卡上，数字应识别为“实数”。 单击“读取值”，您应该看到与以下类似的字段值：

   表 *38:* 示例字段值

字段	测量	值
tSNE_x	连续	[-7.07176703,7.14338837]
tSNE_y	连续	[-9.2188112,8.89647667]
tSNE_x	连续	[-9.95892882,9.95742482]

5. 向流中添加“选择”节点，从而可以删除文件中读取为空值的下列两个底部文本行：

```
*****************
Perform t-SNE (total time 9.5s)
```

   在“选择”节点的“设置”选项卡上，针对模式选择**废弃**，并使用条件 `@NULL(tSNE_x)` 来删除行。
6. 将“类型”节点和“平面文件”导出节点添加到流中以创建 Var。 将复制并粘贴回原始流中的文件源节点。

当文本文件为标准数字格式时访问 t-SNE 绘图数据

1. 创建新流（**文件** > **新建流**）。
2. 向工作区中添加“固定文件”源节点。 以下三个节点全都是访问 t-SNE 节点时所需的节点。

   图 *48: Stream for accessing t-SNE plot data in standard numeric format*
3. 在“固定文件”源节点的“文件”选项卡上使用以下设置：
   - 跳过标题行：1
   - 记录长度：29
   - tSNE_x 开始：3，长度：12

- tSNE_y 开始：16，长度：12

4. 在“过滤器”选项卡上，可以将 field1 和 field2 重命名为 tsneX 和 tsneY。
5. 添加“合并”节点以使用**顺序**合并方法将其连接到流。
6. 现在，可以使用“绘图”节点来绘制 tsneX 与 tsneY，并且使用受调查的字段将其着色。

t-SNE 模型块

t-SNE 模型块包含 t-SNE 模型捕获的所有信息。 以下选项卡可用。

图形

图形选项卡显示 t-SNE 节点的图表输出。 Pyplot 分布图表显示低纬度结果。 如果未在 t-SNE 节点的专家选项卡上选择**使用多个困惑度执行 t-SNE** 选项，那么仅包含一个图形，而不是使用不同困惑度的六个图形。

文本输出

文本输出选项卡显示 t-SNE 算法的结果。 如果在 t-SNE 节点的专家选项卡上选择 **2D** 可视化类型，那么此处的结果是两个维度中的点值。 如果选择 **3D**，那么结果是三个维度中的点值。

E-Plot (Beta) 节点

E-Plot (Beta) 节点显示数字字段之间的关系。 E-Plot (Beta) 节点与绘图节点类似，但是其选项不同，并且其使用新图形功能。 使用该节点可运用 SPSS Modeler 中的新图形功能。

E-Plot (Beta) 节点提供散点图、折线图和条形图来说明数字字段之间的关系。 此节点中的新图形界面直观且现代，可定制程度非常高，并且数据图表为交互式。 有关更多信息，请参阅 第 206 页的『使用 E-Plot 图』。

E-Plot (Beta) 节点“绘图”选项卡

散点图对照 *X* 字段的值，显示 *Y* 的值。 通常而言，这两个字段分别对应于一个自变量和一个因变量。

X 字段。 从列表中选择显示在水平 *x* 轴上的字段。

Y 字段。 从列表中选择显示在垂直 *y* 轴上的字段。

Overlay）。 有不同的方法用于说明数据值的类别。 例如，您可能使用 *maincrop* 字段作为颜色重叠，以指示补贴申请人种植的主要作物的 *estincome* 和 *claimvalue* 值。 选择输出中表示颜色映射、尺寸映射和形状映射的字段。 此外，选择要在交互式输出中包含的任何其他有关字段。 有关更多信息，请参阅主题 第 142 页的『审美原则、重叠、面板和动画』。

一旦为 e-plot 设置了选项，即可通过单击**运行**直接从对话框中运行绘图。 但是，您可能想要使用“选项”选项卡进行其他指定。

E-Plot (Beta) 节点“选项”选项卡

要绘制的最大记录数。 为大型数据集指定一种绘制方法。 可以指定数据集大小上限，或使用缺省的 2000 条记录。 如果选择**抽样**选项，那么处理大数据集时的性能将会提高。 “抽样”选项根据文本字段中输入的记录数对数据进行随机抽样。 另外，您也可以选择 **使用所有数据** ，但必须要注意，这一选项可能大幅降低软件的执行效率。

E-Plot (Beta)“外观”选项卡

如果需要，可以在图形创建之前指定标题和子标题。 这些选项也可以在图形创建之后进行指定或更改。

标题。 输入要用作图形标题的文本。

子标题。 输入要用作图形子标题的文本。

使用 E-Plot 图

E-Plot (Beta) 节点提供散点图、折线图和条形图来说明数字字段之间的关系。 此 Beta 节点中引入的新图形界面包含许多新功能和改进功能。

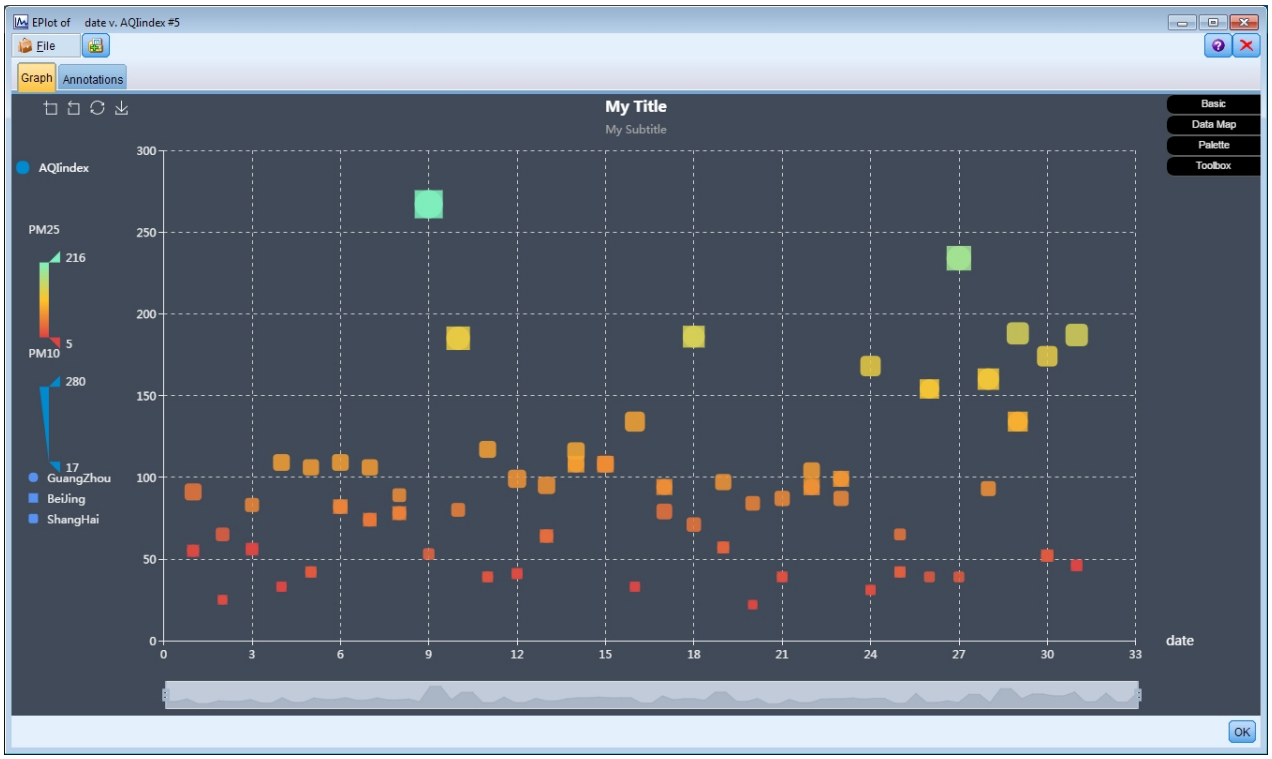


图 *49: E-Plot (Beta) scatterplot graph*

在“图形”选项卡上，左上角提供一个工具栏，用于对图表的特定部分进行放大，重新缩小，返回到初始完整视图，或者保存图表以供外部使用：

图 *50: Toolbar*

在窗口的底部，可以使用滑块对图表的特定部分进行放大。 移动左右两侧的小矩形控件可进行缩放。 要使用此滑块，必须首先在“工具箱”选项区域中将其开启。

图 *51: Zoom slider*

窗口的左侧提供用于更改所显示值的范围的控件。 要使用这些控件，必须首先在“数据图”选项区域中指定选项。 在以下示例中，为颜色映射图选择了名为 PM25 的字段，为尺寸图选择了名为 PM10 的字段，为形状图选择了名为 City 的字段。 可以将鼠标悬停在垂直彩色条形上方以突出显示图形的对应区域，或者滑动上三角形和下三角形。

图 *52: Range controls*

在窗口的右侧，提供了一组可用于实时与数据交互和更改图表外观的可扩展选项：

图 *53: Expandable options*

“基本”选项

图 *54: Basic options*

选择深色或浅色主题，指定标题和子标题，选择图表类型（散点图、折线图或条形图），然后选择 Y 轴上显示的序列。 如果选择**折线**图，那么将仅显示 Y 轴上的字段，并且在颜色映射图和尺寸图的“数据图”选项中将仅提供 Y 轴上的字段。 如果选择**条形**图，那么在“数据图”选项中将仅提供颜色映射图选项。 对于序列，此处将提供在 E-Plot 节点的“绘图”选项卡上选定的所有**有关**字段。

“数据图”选项

图 *55: Data map options*

为**颜色映射图**选择连续字段或分类字段。 如果选择连续字段，那么将显示从绿色到红色的所有颜色。 值越低，其颜色越接近于红色；值越高，其颜色将越接近于绿色。 如果选择了分类字段，那么将根据定义的调色板来显示字段颜色。
大小映射仅支持连续字段。 图表上的值越低，其绘图大小就越小。
形状图仅支持分类字段。 图上显示的形状由分类字段定义，该分类字段将可视化拆分为不同形状的元素（每个类别对应一个元素）。

“调色板”选项

图 *56: Palette options*

如果要定制用于标题和序列的颜色，请使用“调色板”。 从下拉列表中选择标题或序列，单击**编辑预定义颜色**，然后单击**更多**以选择颜色。 或者，可以使用 RGB 或十六进制字段来指定确切颜色。

“工具箱”选项

图 *57: Toolbox options*

使用“工具箱”选项可开启或关闭缩放滑块，设置网格线属性，以及开启或关闭鼠标跟踪。 当鼠标悬停在图表上方时，鼠标跟踪会显示确切坐标位置。

探索图形

使用编辑模式可以编辑图形的布局和外观，使用探索模式可以分析地探索图形表示的数据和值。 探索的主要目的是分析数据并使用带状区域、区域和标记标识值，以生成“选择”节点、“派生”节点或“平衡”节点。 要选择此方式，请从菜单中选择**查看 > 探索方式**（或者单击工具栏图标）。

有些图形可以使用所有探索工具，而有些图形只能使用一个探索工具。 探索模式包括：

- 定义和编辑带状区域，这些区域用于分割尺度 *x* 轴上的值。 有关更多信息，请参阅主题 第 209 页的『使用带状区域』。
- 定义和编辑区域，这些区域用于标识矩形区域中的一组值。 有关更多信息，请参阅主题 第 212 页的『使用区域』。

- 对元素进行标记或取消标记，以手动选择可用于生成“选择”节点或“派生”节点的值。 有关更多信息，请参阅主题 第 214 页的『使用标记后的元素』。
- 使用由带状区域、区域、标记元素和网络链接标识的值生成可在流中使用的节点。 有关更多信息，请参阅主题 第 214 页的『从图形中生成节点』。

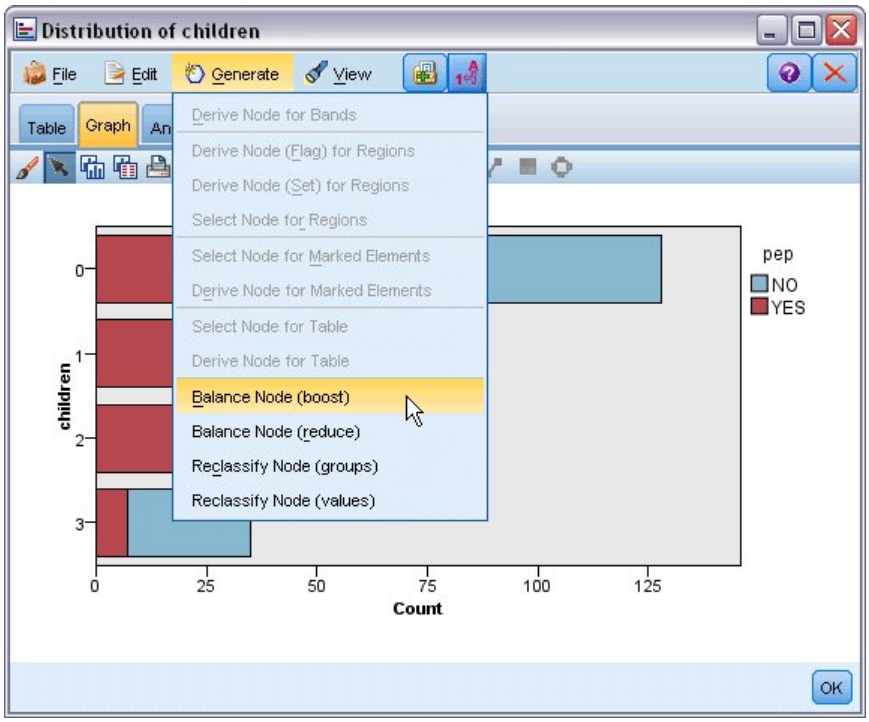


图 *58:* 显示了生成菜单的图

使用带状区域

对于在 *x* 轴上出现尺度字段的任一图形，通过绘制垂直带状线可分割 *x* 轴上的值范围。 如果图形中包含多个面板，那么在一个面板上绘制的带状区域线也会显示在其他面板上。

不是所有图形都可以使用带状区域。 可以使用带状区域的部分图形包括：直方图、条形图以及分布、绘图（折线图、散点图、时间图等）、 集合和评估图表。 在带有面板的图形中，带状区域会显示在所有面板中。 而某些情况下，SPLOM 中显示的为水平带状区域线，这是因为绘制字段/变量带状所在的轴已翻转。

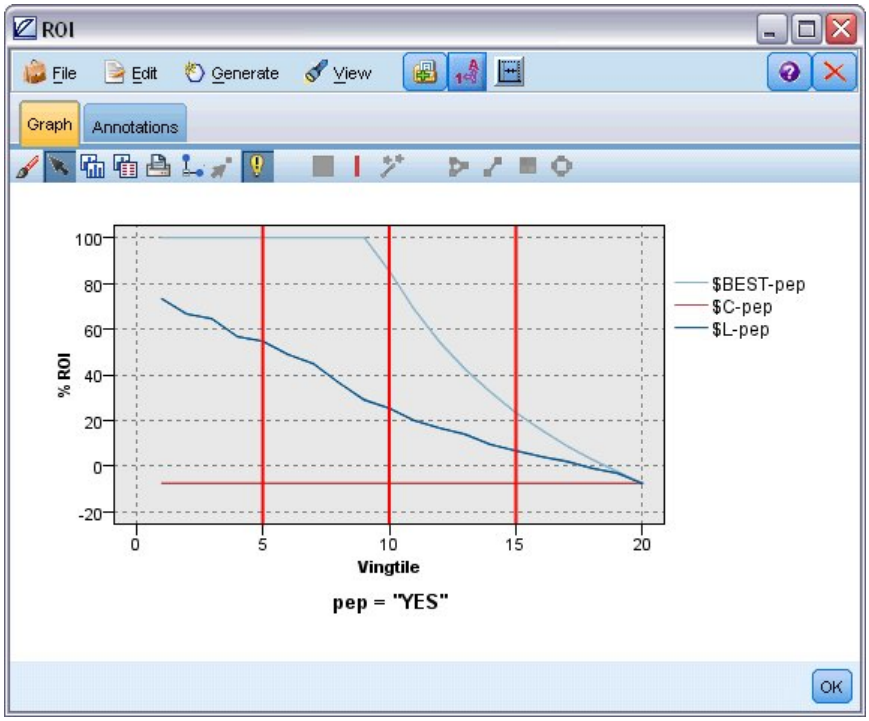


图 *59:* 包含三个带状区域的图形

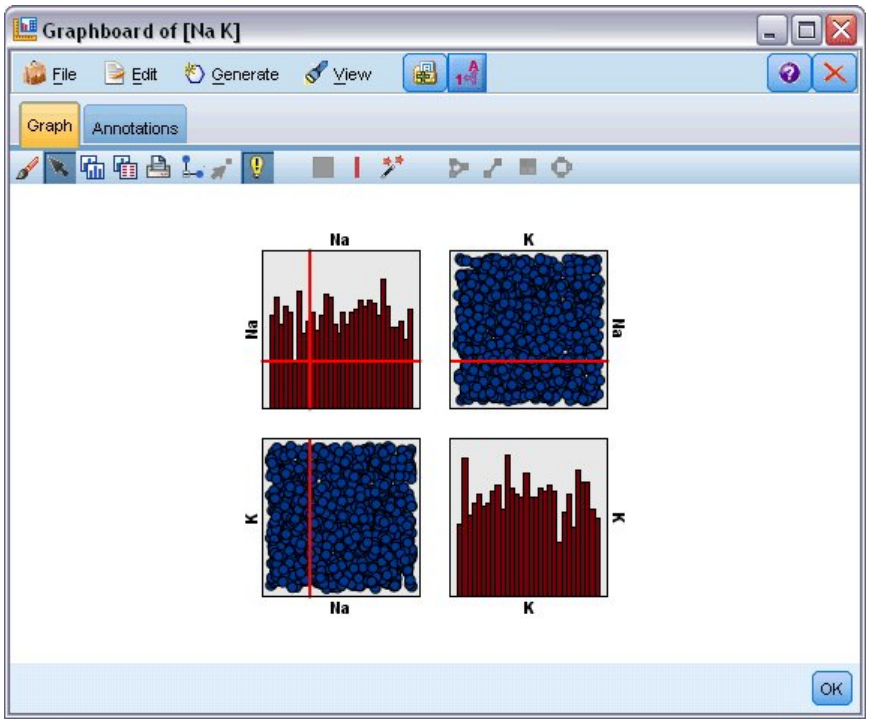


图 *60:* 包含带状区域的 *SPLOM*

定义带状区域

在不包含带状区域的图形中，添加带状线可将图形分割为两个带状区域。 带状区域线值表示从左至右查看图形时第二个带状区域的起点，也称为下限。 同样，在包含两个带状区域的图形中，添加带状区域线可将其中一个带状区域分割为两部分，最终形成三个带状区域。 缺省情况下，带状区域的名称为 带状区域 *N*，其中 *N* 相当于沿 *x* 轴从左到右带状区域的序号。

定义带状区域后，通过拖放带状区域可在 *x* 轴上重新对其进行定位。 通过在带状区域内单击右键可以查看更多的快捷方式，这些快捷方式用于重命名、删除或生成指定带状区域的节点。

要定义带状区域：

1. 验证您是否处于探索模式。 从菜单中选择**视图 > 探索模式**。
2. 在探索模式工具栏上，单击“绘制带状区域”按钮。

图 *61: “绘制带状区域”工具栏按钮*

3. 在接受带状区域的图形中，单击定义带状区域线所在的 *x* 轴值点。

注意：或者，单击**将图形分割为带状区域**工具栏图标，然后输入需要的相等带状区域量，然后单击**分割**。

图 *62:* 分割图标可扩展工具栏，提供用于分割为带状区域的选项

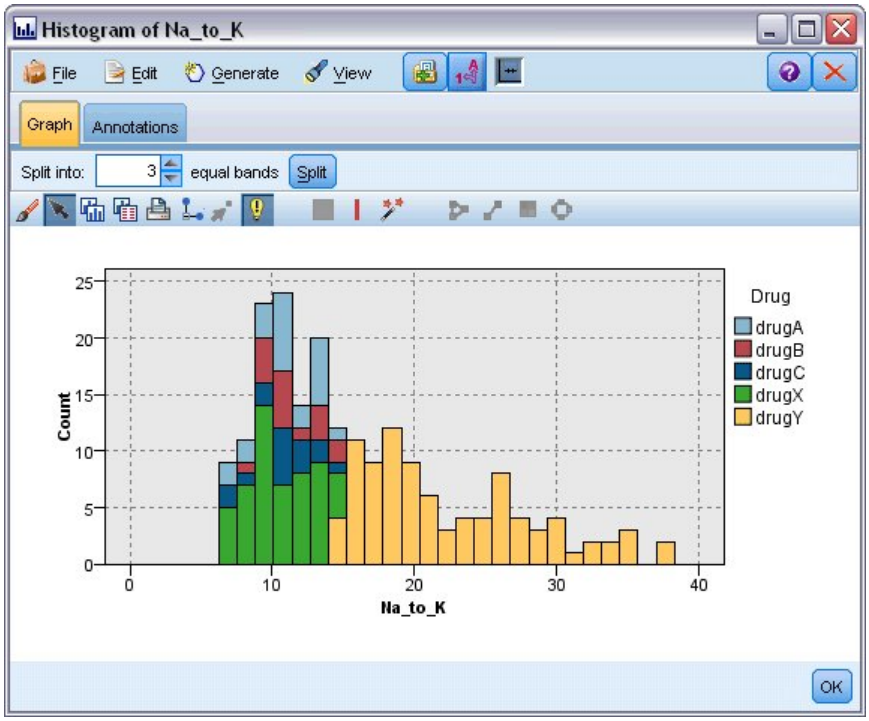


图 *63:* 创建启用带状区域的相等带状区域工具栏

编辑、重命名和删除带状区域

可以在“编辑图形带状区域”对话框中或通过图形自身的上下文菜单编辑现有带状区域的属性。

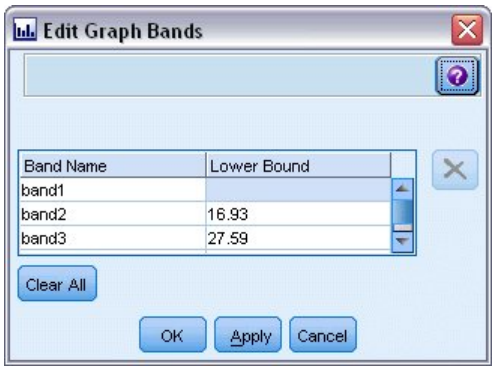


图 *64: “编辑图形带状区域”对话框*

要编辑带状区域：

1. 验证您是否处于探索模式。 从菜单中选择**视图 > 探索模式**。
2. 在探索模式工具栏上，单击“绘制带状区域”按钮。
3. 从菜单中，选择**编辑 > 图形带状区域**。 此时将打开“编辑图形带状区域”对话框。
4. 如果图形中有多个字段（例如 SPLOM 图形），可以在下拉列表中选择所需字段。
5. 通过输入名称和下限添加新的带状区域。 按 Enter 键另起一行。
6. 通过调整**下限**值编辑带状区域的边界。
7. 通过输入新的带状区域名称重命名带状区域。
8. 通过选择表中的线并单击“删除”按钮删除带状区域。
9. 单击 **确定** 应用更改并关闭此对话框。

注意：另外，通过右键单击带状区域的线并从上下文菜单中选择需要的选项，可直接删除和重命名图形中的带状区域。

使用区域

在包含两个尺度（或范围）轴的任一图形中，通过绘制区域可在绘制好的矩形区域（称为区域）中对值进行分组。 **区域** 为图形中的一部分，由 *X* 和 *Y* 的最小值及最大值决定。 如果图形中包含多个面板，那么在一个面板上绘制的区域也会显示在其他面板上。

不是所有图形都可以使用区域。 其中一些接受区域的图形包括：绘图（折线图、散点图、气泡图、时间图等）、 SPLOM 和集合。 这些区域是在 X、Y 空间中绘制的，因此无法定义在 1D 散点图、3D 散点图或动画散点图中。 在包含面板的图形中，区域会显示在所有面板中。 在散点图矩阵图 (SPLOM) 中，相应的区域会显示在相应的上部散点图中，而不是显示在对角散点图中，因为后者只显示一个尺度字段。

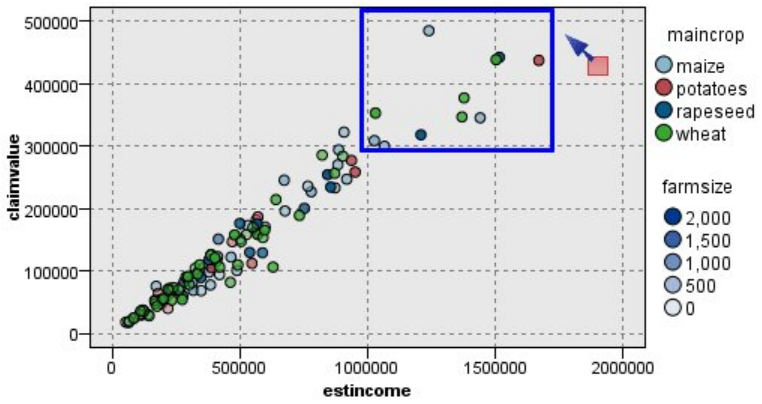


图 *65:* 定义具有高索赔值的区域

定义区域

无论在哪个位置上定义区域，都要对值进行分组。 缺省情况下，每个新区域都称为 *Region<N>*，其中 *N* 对应于已创建的区域数。

定义区域后，通过右键单击区域线可获得一些基本的快捷方式。 另外，通过在区域内（而不是在线上）单击右键可以查看许多其他的快捷方式，这些快捷方式用于重命名、删除或生成指定区域的节点。

您可以根据记录与某个特定区域或多个区域中的某一个区域的包含关系选择记录的子集。 通过生成“派生”节点，并根据记录与区域的包含关系生成标志值，以此为记录添加区域信息。 有关更多信息，请参阅主题 第 214 页的『从图形中生成节点』。

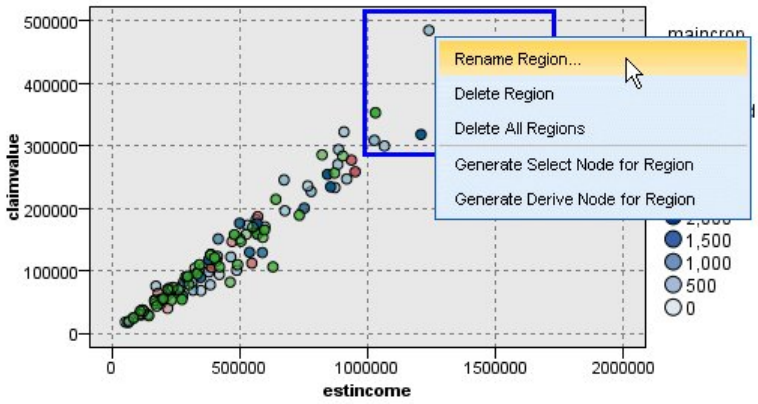


图 *66:* 探索具有高索赔值的区域

要定义区域：

1. 验证您是否处于探索模式。 从菜单中选择**视图 > 探索模式**。
2. 在探索模式工具栏上，单击“绘制区域”按钮。

图 *67:* “绘制区域”工具栏按钮

3. 在可使用区域的图形中，通过单击并同时拖放鼠标可绘制矩形区域。

编辑、重命名和删除区域

可以在“编辑图形带状区域”对话框中或通过图形自身的上下文菜单编辑现有带状区域的属性。

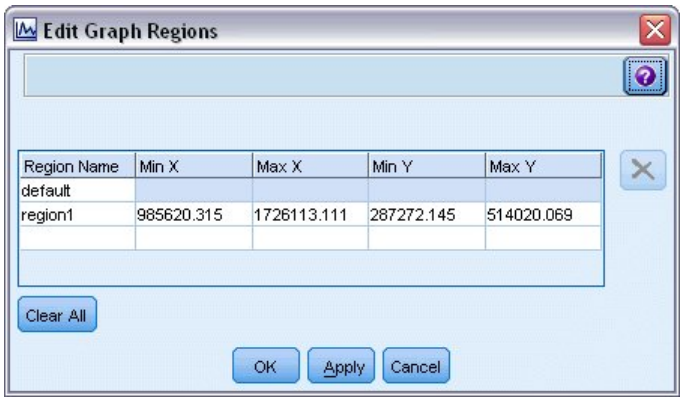


图 *68:* 指定定义区域的属性

要编辑区域：

1. 验证您是否处于探索模式。 从菜单中选择**视图 > 探索模式**。
2. 在探索模式工具栏上，单击“绘制区域”按钮。
3. 从菜单中选择**编辑 > 图形区域**。 此时将打开“编辑图形区域”对话框。
4. 如果图形中包含多个字段（例如 SPLOM 图形），必须在字段 *A* 列和字段 *B* 列中定义该区域的字段。
5. 通过输入名称，选择字段名称（如果适用）并定义各个字段的上限和下限在新行上添加新区域。 按 Enter 键另起一行。
6. 通过调整 *A* 和 *B* 的**最小值**和**最大值**编辑现有区域边界。
7. 通过更改表中区域的名称重命名区域。
8. 通过选择表中的线并单击“删除”按钮删除区域。
9. 单击 **确定** 应用更改并关闭此对话框。

注意：另外，通过右键单击区域的线并从上下文菜单中选择需要的选项，可直接删除和重命名图形中的区域。

使用标记后的元素

您可以对任一图形中的元素（例如条形、饼块和点）进行标记。 除时间散点图、多散点图和评估图形之外的图形中不能标记线、区域和表面，因为在这些图形中线表示字段。 标记元素时，通常要突出显示该元素表示的所有数据。 同一元素出现在多个位置的任一图形（例如 SPLOM）中，标记会与涂抹同步进行。 您可以对图形中、甚至带状区域和区域中的元素进行标记。 标记元素并返回编辑模式时，都会显示标记。

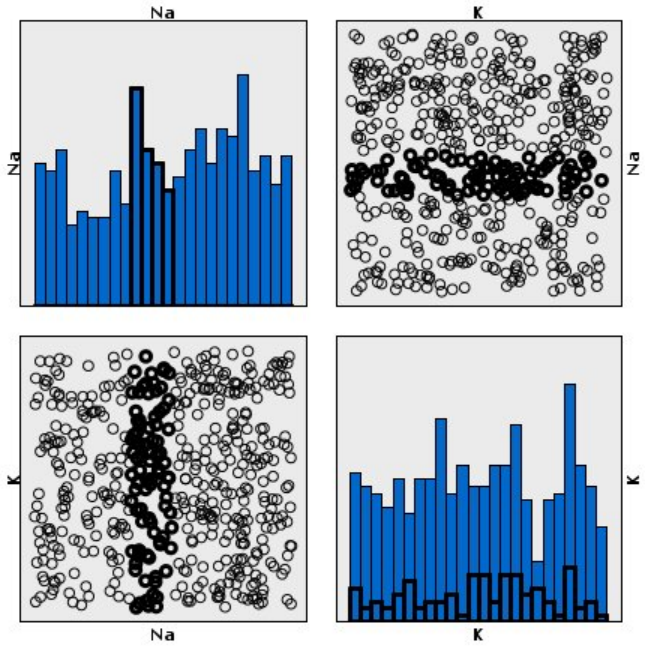


图 *69:* 对 *SPLOM* 中的元素进行标记

通过单击图形中的元素，可以对元素进行标记和取消标记。 首次单击元素进行标记时，该元素的边框显示为深色，表示已标记。 如果再次单击该元素，该边框会消失，表示该元素不再进行标记。 要对多个元素进行标记，可以按住 Ctrl 键并单击元素，也可以使用“魔棒”将鼠标拖放到需要标记的所有元素周围。 请记住，如果在没有按住 Ctrl 键的情况下单击其他区域或元素，则会清除之前标记的所有元素。

可以生成图中的标记元素的“选择”节点和“派生”节点。 有关更多信息，请参阅主题 第 214 页的『从图形中生成节点』。

要标记元素：

1. 验证您是否处于探索模式。 从菜单中选择**视图 > 探索模式**。
2. 在探索模式工具栏上，单击“标记元素”按钮。
3. 单击所需元素，或单击并拖动鼠标，在包含多个元素的区域周围画一条线。

从图形中生成节点

IBM SPSS Modeler 图形提供的最强大功能之一是可从图形中生成节点或从图形中所选内容生成节点。 例如，可在时间散点图中根据选择和数据区域生成“派生”节点和“选择”节点，有效地将数据划分为多个“子集”。 例如，可使用此强大功能来识别和排除离群值。

在绘制带状区域时，也可以生成“派生”节点。 在包含两个尺度轴的图形中，可以从在图形中绘制的区域中生成“派生”节点或“选择”节点。 在包含标记元素的图形中，可以从这些元素中生成“派生”节点和“选择”节点，有些情况下可以生成“过滤”节点。 可以为显示计数分布的任一图形启用平衡节点的生成。

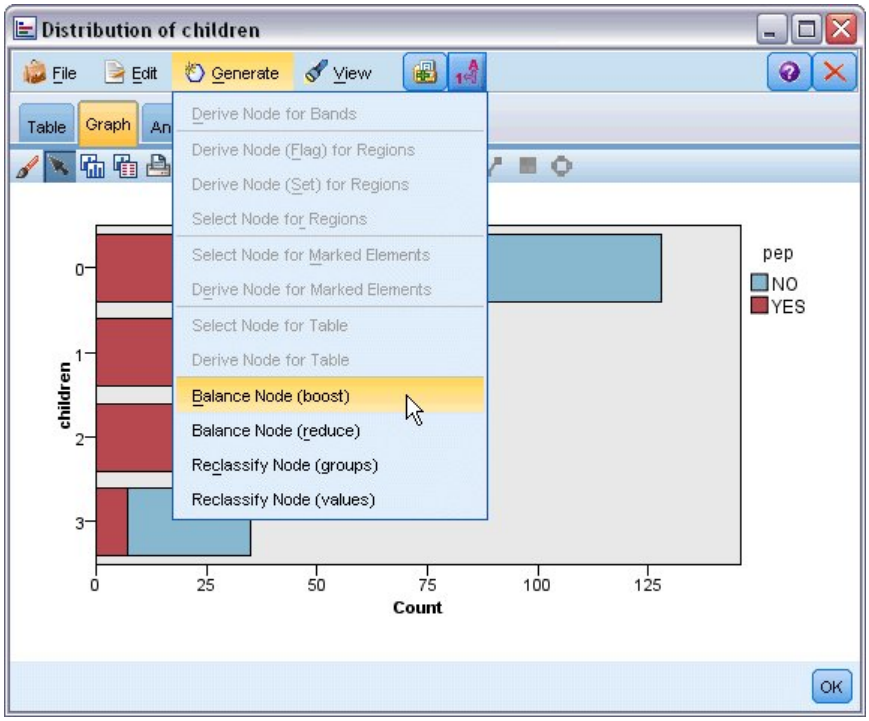


图 *70:* 显示了生成菜单的图

生成节点时，该节点直接放置在流画布中，以便将其连接到现有的流。 下列节点可从图形中生成：选择、派生、平衡、过滤和重新分类。

选择节点

通过生成选择节点，可以检验用于为下游处理检验区域中记录的包含关系或区域外所有记录的排除关系，或检验与之相反的情况。

- **适用于带状区域**。 可以生成用于在该带状区域中包含或排除记录的“选择”节点。 **仅适用于带状区域的“选择”节点**只能通过上下文菜单访问，因为您需要选择要在“选择”节点中使用的带状区域。
- **适用于区域**。 可以生成用于包括或排除该区域中的记录的“选择”节点。
- **适用于标记元素**。 您可以生成“选择”节点捕获与标记元素或 Web 图形链接相对应的记录。

导出节点

“派生”节点可从区域、带状区域和标记元素中生成。 所有图形都可生成“派生”节点。 在评估图表中，会出现用于选择模型的对话框。 在网络图形中， **导出节点（“与”）** 和 **导出节点（“或”）** 都有可能生成。

- **适用于带状区域**。 可以生成“派生”节点，它通过将“编辑带状区域”对话框中所列的带状区域名称用作类别名称，为轴上标记的每个间隔生成类别。
- **适用于区域**。 可以生成导出节点（ **导出为标志** ），该节点用于创建一个名为 *in_region* 的标志字段，其中所含的标志 *T* 表示记录位于任一区域内， *F* 表示记录在所有区域外。 您也可以生成导出节点（ **导出为集** ），该节点将为各个区域生成值集、为各个记录生成名为 区域 的新字段，并采用其值作为记录所在区域的名称。 而不在任何区域内的记录将与缺省区域同名。 值名称成为“编辑区域”对话框中所列的区域名称。
- **适用于标记元素**。 可以生成计算标志的导出节点，该标志为用于所有标记元素的 *True* 和用于所有其他记录的 *False* 。

平衡节点

生成的平衡节点用于纠正数据中的不平衡情况，例如减少常用值的频率（使用 **平衡节点（减少）** 菜单选项）或推进非常用值的出现次数（使用 **平衡节点（推进）** 菜单选项）。 为显示计数分布情况的任一图形启用平衡节点的生成，这些图形包括直方图、点图、收集图、计数条形图、计数饼图和多散点图。

过滤节点

通过生成“过滤”节点可根据图形中所标的线或节点重命名字段或对它们进行过滤。 在评估图表中，最佳拟合线不会生成过滤节点。

重新分类节点

通过生成“重新分类”节点可对值进行重新编码。 此选项适用于分布图。 可以为**组**生成“重新分类”节点，以便根据所显示字段的特定值在某个组（在**表**选项卡上，通过按住 Ctrl 键并同时单击来选择组）中的包括关系对这些值进行重新编码。 也可以为 **值** 生成重新分类节点，以将数据重新编码到许多值的现有集，例如将数据重新分类到值的标准集，以合并多个公司的财务数据进行分析。

注: 如果已经预定义了值，那么可以将它们作为平面文件读取到 IBM SPSS Modeler 中，并使用分布图显示所有值。 然后从图表中直接生成一个此字段的“重新分类”（值）节点。 如果执行该操作，则所有目标值都会出现在重新分类节点的 新值 列（下拉列表）中。

设置“重新分类”节点的选项时，表将启用从旧集合值到指定的新值的全新映射：

- **原始值**。 此列列出选择字段的现有值。
- **新值**。 使用此列可输入新的类别值或从下拉列表中选择类别值。 使用分布图中的值自动生成“重新分类”节点时，这些值将包括在该下拉列表中。 这样，您可以将现有值快速映射至已知值集合。 例如，医疗保健组织有时会根据网络或语言环境对诊断进行不同分组。 经过合并或采集，所有各方都需要采用一致方式对新的或现有数据进行重新分类。 可以将值的主列表读入 IBM SPSS Modeler，对 *Diagnosis* 字段运行条形图，然后直接从该图生成字段的重新分类（值）节点，而无需手动键入冗长列表中的每个目标值。 此过程将使所有目标 Diagnosis 值显示在“新值”下拉列表中。

有关“重新分类”节点的更多信息，请参阅第 121 页的『为重新分类节点设置选项』。

从图形中生成节点

可以使用图形输出窗口中的“生成”菜单生成节点。 生成的节点将出现在流画布中。 要使用此节点，请将其连接到现有流中。

要从图形生成节点：

1. 验证您是否处于探索模式。 从菜单中选择**视图 > 探索模式**。
2. 在探索模式工具栏上，单击“区域”按钮。
3. 定义生成节点所需的带状区域、区域或任何标记元素。
4. 从“生成”菜单中选择要生成的节点类型。 只能使用现有的类型。

注意：或者，也可以通过右键单击并从上下文菜单中选择需要的生成选项来直接从图形中生成节点。

编辑可视化

探索模式允许您以分析的方式探索可视化代表的数据和值，而编辑模式允许您更改可视化的布局和外观。 例如，您可以更改字体和颜色，以匹配您组织的样式指南。 要选择此方式，请从菜单中选择 **查看 > 编辑方式**（或单击工具栏图标）。

在编辑模式中，有几种工具栏会影响可视化布局的不同方面。 如果您发现存在任何不用的内容，可以隐藏以增加显示图形的对话框中的空间量。 要选择或取消选择工具栏，请在“视图”菜单上单击相关的工具栏名称。

注：要将进一步详细信息添加到您的可视化，可以应用标题、脚注和轴标签。 有关更多信息，请参阅第 224 页的『添加标题和脚注』主题。

在 **编辑模式** 中有几个编辑直观表示的选项。 您可以执行以下操作：

- 编辑文本并设置格式。
- 更改边框和图形元素的填充颜色、透明度和模式。
- 更改边框和线的颜色和划线。
- 旋转并更改点元素的形状和宽高比。
- 更改图形元素的大小（如条和点）。
- 通过使用边距和填充调整项目周围的空间。
- 指定数字格式。

- 更改轴和刻度设置。
- 排序、排除和拼并分类轴上的类别。
- 设置面板方向。
- 应用坐标系转换。
- 更改统计、图形元素类型和冲突修饰符。
- 更改图注的位置。
- 应用可视化样式表。

以下主题描述了如何执行这些不同任务。 同时建议您阅读编辑图形的一般规则。

如何切换到编辑模式

从菜单中选择：

视图 > 编辑方式

编辑可视化的一般规则

编辑模式

在编辑模式中进行所有编辑。 要启用编辑模式，请从菜单中选择：

查看 > 编辑方式

选择(S)

编辑的可用选项取决于选择情况。 根据选择的内容启用不同的工具栏和属性调色板选项。 只有启用的项目才能应用到当前选择中。 例如，如果选择了一个轴，则将在“属性”调色板中可用“刻度”、“主刻度标记”和“次刻度标记”选项卡。

以下是有关在可视化中选择项目的一些提示：

- 单击某项将其选中。
- 使用单击选择一个图形元素（如散点图中的点或条形图中的条）。 在初始选择后，再次单击将选择缩小为图形元素组或单个图形元素。
- 按 Esc 取消所有选择。

调色板

当在可视化中选择了一个项目时，各种调色板会更新以反映选择。 调色板包含用于编辑选择的控件。 调色板可以是工具栏或带有多个控件和选项卡的面板。 调色板可以隐藏，以确保显示必要调色板用于编辑。 查看“视图”菜单了解当前显示的调色板。

您可以通过单击并拖动工具栏调色板中或其他调色板左侧的空白空间更改调色板位置。 可视反馈让您知道您可以将调色板放在哪里。 对于非工具栏调色板，您还可以单击关闭按钮隐藏调色板，或单击取消放置按钮以在单独窗口中显示调色板。 单击帮助按钮以显示特定调色板的帮助。

自动设置

一些设置提供 **-自动-** 选项。 这表示应用自动值。 使用哪些自动设置取决于特定可视化和数据值。 您可以输入一个值以覆盖自动设置。 如果您想恢复自动设置，请删除当前值并按 Enter。 此设置将再次显示 **-auto-**。

删除/隐藏项目

您可以删除/隐藏可视化中的各种项目。 例如，您可以隐藏图注或轴标签。 要删除一个项目，请选择并按“删除”键。 如果项目不允许删除，则不会发生任何操作。 如果您意外删除了一个项目，请按 Ctrl+Z 取消删除。

状态

某些工具栏可反映当前选定内容的状态，其他工具栏则不会反映。 属性选用板始终反映此状态。 如果某个工具栏不 反映状态，则会在描述此工具栏的主题中予以说明。

编辑和设置文本格式

可以在适当的位置编辑文本并更改整个文本块的格式。 注意，不能编辑直接链接到数据值的文本。 例如，您无法编辑一个标记标签，因为标签的内容是从基本数据中派生出来的。 然而，您可以设置直观表示中任何文本的格式。

如何编辑所在位置的文本

1. 双击该文本块。 此操作可选中所有文本。 此时禁用所有工具栏，因为您在编辑文本时无法更改可视化的任何其他部分。
2. 输入新内容以替换现有文本。 也可以再次单击此文本以显示光标。 将光标置于所需位置并输入其他文本。

如何设置文本格式

1. 选择包含文本的边框。 不要双击此文本。
2. 使用字体工具栏设置文本格式。 如果工具栏未启用，确保只选择包含文本的边框。 如果选择了文本本身，工具栏将禁用。

可以对字体进行以下更改：

- 颜色
- 族（例如，Arial 或 Verdana）
- 字号（单位为 pt，除非指定了其他单位，例如 pc）
- 权重
- 相对于文本框进行调整

格式应用于一个边框中的所有文本。 您无法更改任何特定文本区中的单个字母或单词的格式。

更改颜色、模式、划线和透明度

直观表示中的许多不同项目都有一个填充和边框。 最明显的示例是条形图中的条形。 条的颜色是填充颜色。 这些条形四周还会有黑色的实线边框。

在有填充颜色的可视化中有其他不太明显的项目。 如果填充颜色是透明的，那么您可能不知道这里进行了填充。 例如，考虑轴标签中的文本。 看起来好像此文本是“悬浮”文本，但是实际上出现在拥有透明填充颜色的边框中。 您可以通过选择轴标签查看边框。

直观表示中的任何边框可以拥有一个填充和边框样式，包括整个直观表示周围的边框。 另外，任何填充都有可以调整的相关不透明度/透明度级别。

如何更改颜色、模式、划线和透明度

1. 选定要对其进行格式化的项目。 例如，选定条形图中的条形或包含文本的框。 如果直观表示被分类变量或字段分割，那么您还可以选择对应于单个类别的组。 这样您可以更改应用于该组的缺省外观。 例如，可以更改堆积条形图中其中一个堆积组的颜色。
2. 要更改填充颜色、边框颜色或填充模式，请使用颜色工具栏。

   注：此工具栏不反映当前选择的状态。

   要更改颜色或填充，您可以单击按钮以选择显示的选项或单击下拉箭头以选择另一个选项。 对于颜色，注意有一个看起来像白色的颜色，其间有一条红色对角线穿过。 这是一种透明颜色。 例如，您可以用此隐藏直方图中的条边框。

   - 第一个按钮可控制填充颜色。 如果颜色与连续字段或有序字段相关联，那么此按钮将更改与数据中最高值相关联的颜色的填充颜色。 您可以使用属性选用板上的“颜色”选项卡来更改与最低值和缺失数据相关联的颜色。 随着基本数据值的增大，元素的颜色将逐步由“低”颜色更改为“高”颜色。
   - 第二个按钮控制边框颜色。
   - 第三个按钮控制填充模式。 填充模式使用边框颜色。 因此，只有存在可见边框颜色的情况下填充模式才可见。

- 第四个控件是控制填充颜色和模式的不透明度的滑动条和文本框。 更低的百分比意味着更少的不透明度和更高的透明度。100% 表示完全不透明（不透明度）。

3. 要更改边框或线的划线，请使用线工具栏。

   注：此工具栏不反映当前选择的状态。

   和其他工具栏一样，可以单击按钮选择已显示的选项，或单击下拉箭头选择其他选项。

旋转并更改点元素的形状和宽高比

可以旋转点元素，向其应用其他预定义的形状，或更改其高宽比（宽度相对高度的比率）。

如何修改点元素

1. 选择点元素。 不能旋转并更改单个点元素的形状和高宽比。
2. 可使用符号工具栏修改点。
   - 第一个按钮允许您更改点的形状。 单击下拉箭头并选择一个预定义的形状。
   - 第二个按钮允许您将点旋转到一个特定的罗盘位置。 单击下拉箭头然后将指针拖动到所需位置。
   - 第三个按钮允许您更改宽高比。 单击下拉箭头然后单击并拖动弹出的矩形。 矩形的形状代表宽高比。

更改图形元素的大小

您可以更改直观表示中图形元素的大小。 这些图形元素包括条形、线和点等等。 如果由变量或字段确定图形元素的大小，则指定的大小为 最小 。

更改图形元素大小的方式

1. 选定要更改其大小的图形元素。
2. 使用滑块来更改大小。

指定边距和填充

如果可视化中的边框周围或内部空间太大或太小，您可以更改其边距和填充设置。 **边距**是边框及其周围其他项目之间的空间量。 **填充**指位于框的边框和框的内容 之间的空白量。

如何指定边距和填充

1. 选择您想指定边距和填充的边框。 这可以是文本框、图注周围的边框，或者甚至是显示图形元素（如条或点）的数据框。
2. 使用“属性”调色板上的“页边距”选项卡指定设置。 所有大小单位都是像素，除非您指定了一个不同的单位（如 cm 或 in）。

设置数字格式

您可以指定连续轴或显示数字的数据值标签上刻度标记标签的数字格式。 例如，您可以指定刻度标记标签中显示的数字以千显示。

如何指定数字格式

1. 选择连续轴刻度标记标签或数据值标签（如果包含数字）。
2. 单击“属性”选用板上的**格式**选项卡。
3. 选择所需的数字格式设置选项：

   前缀。 一个显示在数字开头的字符。 例如，如果数字是以美元为单位的薪金，请输入美元符号 ($)。

   后缀。 一个显示在数字末尾的字符。 例如，如果数字是百分比，请输入百分比符号 (%)。

   最短 整数位数。 小数表示法的整数部分中显示的最小位数。 如果实际值不包含最小位数，该值的整数部分将用零填充。

最大值 整数位数。 小数表示法的整数部分中显示的最大位数。 如果实际值超过最大位数，该值的整数部分将用星号替换。

最短 十进制数字。 小数或科学表示法的小数部分中显示的最小位数。 如果实际值不包含最小位数，该值的小数部分将用零填充。

最大值 十进制数字。 小数或科学表示法的小数部分中显示的最大位数。 如果实际值超过最大位数，小数将四舍五入为适当的位数。

科学表示法。 是否以科学记数法显示数字。 科学表示法对于非常大或非常小的数字是有用的。**-auto-** 允许应用程序确定科学表示法何时适用。

比例。 刻度因子，原始值除以的数字。 当数字很大，而您又不想使标签为容纳该数字而延长太多时，可使用刻度因子。 如果更改刻度标记标签的数字格式，请务必编辑轴标题来说明如何理解数字。 例如，假定刻度轴显示工资，标签为 30000、50000 和 70000。 可以输入刻度因子 1000 以显示 30、50 和 70。 然后，应该编辑刻度轴标题以包含文本（以千为单位）。

-ve 圆括号。 圆括号是否应该显示在负值周围。

分组。 是否在数字组之间显示字符。 您计算机的当前语言环境确定使用哪个字符用于数字组。

更改轴和刻度设置

有几个选项可用于修改轴和尺度。

如何更改轴和刻度设置

1. 选择轴的任何部分（例如，轴标签或刻度标记标签）。
2. 使用属性选项板上的“尺度”、“主要核对项”和“次要核对项”选项卡更改轴和尺度的设置。

“刻度”选项卡

注：对于数据已预先汇总的图形（例如，直方图），不会显示“刻度”选项卡。

类型。 指定刻度为线性还是转换的。 尺度变换可帮助您了解数据或对统计推论进行必要的假设。 在散点图中，如果自变量和因变量之间或独立字段和相关字段之间的关系是非线性的，那么可以使用变换尺度。 刻度转换还可以用于使偏斜的直方图看上去更对称一点，以便与正态分布类似。 注意：您转换的只是数据显示的刻度；而不是转换实际数据。

- **线性**。 指定线性非变换的刻度。
- **log.** 指定以 10 为底数的对数转换的刻度。 为了容纳零和负值，此转换使用一个修改版本的对数函数。 此“安全对数”函数定义为 `sign(x) * log(1 + abs(x))`。 因此 `safeLog(-99)` 等于：

  ```
  sign(-99) * log(1 + abs(-99)) = -1 * log(1 + 99) = -1 * 2 = -2
  ```
- **幂**。 指定指数为 0.5 的幂变换刻度。 为适用于负数，此变换使用幂函数的修改版。 此“安全幂”函数定义为 `sign(x) * pow(abs(x), 0.5)`。 因此 `safePower(-100)` 等于：

  ```
  sign(-100) * pow(abs(-100), 0.5) = -1* pow(100, 0.5) = -1 * 10 = -10
  ```

最小值/最大值/适当低值/适当高值。 指定刻度的范围。 选择**适当低值**和**适当高值**允许应用程序选择一个基于数据的适当刻度。 最小值和最大值之所以成为“略值”是因为它们通常分别是大于最大数据值或小于最小数据值的整数。 例如，如果数据范围是从 4 到 92，刻度的适当低值和高值可以是 0 和 100，而非实际数据的最小值和最大值。 请谨慎，不要设置一个太小并隐藏重要项目的范围。 同时注意，如果选择了**包括零**选项，您无法设置一个显式的最小值和最大值。

低边距/高边距。 在轴的低和/或高端创建边距。 边距垂直于选定的轴显示。 此单位为像素，除非指定了其他单位（例如厘米或英寸）。 例如，如果您将垂直轴的**高边距**设置为 5，则 5 px 的水平轴沿数据边框的顶部运行。

反向。 指定刻度是否反转。

包括零。 指示刻度应包括 0。 此选项通常用于条形图以确保条形从 0 处开始，而不是从接近最小条形高度的值开始。 如果选择了此选项，则禁用**最小值**和**最大值**，因为您无法为刻度范围设置一个定制的最小值和最大值。

主刻度标记/次刻度标记选项卡

刻度标记或**刻度线**是出现在轴上的线。 这些表示特定区间或类别的值。 **主刻度标记**是带有标签的刻度线。 这些记号也比其他勾号标记长。 **次刻度标记**是出现在主刻度线之间的刻度线。 某些选项只可用于某种记号类型，但大多数选项对于主要核对项和次要核对项都可用。

显示刻度标记。 指定在图形上是否显示主刻度标记或次刻度标记。

显示网格线。 指定是否在主刻度标记或次刻度标记处显示网格线。 **网格线**是从一个轴到另一个轴穿过整个图形的线。

位置。 指定刻度标记相对于轴的位置。

长度。 指定刻度线的长度。 此单位为像素，除非指定了其他单位（例如厘米或英寸）。

Base。 仅适用于主刻度标记。 指定第一个主刻度标记出现位置对应的值。

Delta 仅适用于主刻度标记。 指定主刻度标记之间的差值。 即，主刻度标记将在每第 *n* 个值处出现，其中 *n* 是 delta 值。

分区。 仅适用于次刻度标记。 指定主刻度标记之间的次刻度标记分区数。 次刻度标记的数量比分区的数量少一。 例如，假定在 0 和 100 处有主刻度标记。 如果您输入 2 作为次刻度标记分区的数目，那么在 50 处将有一个次刻度标记，用于划分 0 到 100 的范围并创建两个分区。

编辑类别

可以以多种方式编辑分类轴上的类别：

- 更改显示类别的排序顺序。
- 排除特定类别。
- 添加数据集中不出现的类别。
- 将小类别拼并/合并为一个类别。

如何更改类别的排序顺序

1. 选择分类轴。 类别选用板显示轴上的类别。

   注意：如果选用板不可见，请确保已将其启用。 从 IBM SPSS Modeler 的“视图”菜单中选择**类别**。
2. 在“类别”调色板中，从下拉列表中选择一个排序选项。

   定制。 根据类别在选用板中的显示顺序排列类别。 使用方向按钮可以将类别移动至列表顶部、上移、下移或者列表底部。

   数据。 根据类别在数据集中显示的顺序排列类别。

   名称。 使用调色板中显示的名称按字母顺序对类别进行排序。 这可能是值或标签，取决于是否选择了由工具栏按钮来显示值和标签。

   值。 使用选用板括号中显示的值按基本数据值对类别进行排序。 只有带有元数据的数据源（如 IBM SPSS Statistics 数据文件）才支持此选项。

   统计。 根据每个类别的计算统计对类别进行排序。 统计的示例包括计数、百分比和平均值。 只有图形中使用统计时此选项才可用。

如何添加类别

缺省情况下，只可使用数据集中出现的类别。 如果需要，您可以将类别添加到可视化中。

1. 选择分类轴。 类别选用板显示轴上的类别。

   注意：如果选用板不可见，请确保已将其启用。 从 IBM SPSS Modeler 的“视图”菜单中选择**类别**。
2. 在“类别”选用板中，单击添加类别按钮：

   图 *71:* 添加类别按钮
3. 在“添加新类别”对话框中，输入类别名称。

4. 单击**确定**。

如何排除特定类别

1. 选择分类轴。 类别选用板显示轴上的类别。

   注意：如果选用板不可见，请确保已将其启用。 从 IBM SPSS Modeler 的“视图”菜单中选择**类别**。
2. 在“类别”选用板中，选择“包括”列表中的类别名称，然后单击 X 按钮。 要将类别移回，请选择“排除”列表中的类别名称，然后单击指向右方列表的箭头。

如何拼并/合并小类别

您可以将没有必要单独显示的小类别进行合并。 例如，在一个具有许多类别的饼图中，可以折叠百分比不足 10 的类别。 折叠仅可用于可加的统计值。 例如，您无法一起添加平均值，因为平均值不可加。 因此，使用均值合并/拼并类别不可用。

1. 选择分类轴。 类别选用板显示轴上的类别。

   注意：如果选用板不可见，请确保已将其启用。 从 IBM SPSS Modeler 的“视图”菜单中选择**类别**。
2. 在类别选用板中，选择 **折叠** 并指定百分比。 任何总计百分比小于指定数量的类别将合并为一个类别。 百分比基于图表中显示的统计量。 拼并只对基于计数的和求和（总和）统计可用。

更改方向面板

如果您在直观表示中使用面板，您可以更改其方向。

如何更改面板方向

1. 选择可视化的任何部分。
2. 单击“属性”调色板上的**面板**选项卡。
3. 从 **布局** 中选择选项：

表。 像表一样布局面板，其中行或列分配到每个单个值。

转置。 将面板布局成类似表的样子，也交换原来的行和列。 此选项的作用和转置图形本身不一样。 注意，选择此选项时， *x* 轴和 *y* 轴不会改变。

列表。 像列表一样布局面板，其中每个单元格代表值的组合。 不再对列和行分配单个值。 此选项使面板能够在必要时换行。

转换坐标系

许多可视化显示在平滑的直角坐标系中。 如果需要，您可以转换坐标系。 例如，您可以将一个极转换应用到坐标系，添加斜交阴影效果，并变换轴。 如果已应用于当前可视化，您还可以取消任何这些转换。 例如，在极坐标系中绘制饼图。 您可以撤销极转换，并在直角坐标系中将饼图显示为单个堆积条形图。

如何转换坐标系

1. 选择您想转换的坐标系。 您可以通过选择单个图形周围的边框选择坐标系。
2. 单击“属性”选用板上的**坐标**选项卡。
3. 选择想要应用到该坐标系的转换。 您还可以取消选择转换以取消该转换。

   转置。 更改轴的方向称为**转置**。 这类似于对调二维可视化中的水平和垂直轴。

   极。 极转换以与图形中心之间的特定角度和距离绘制图形元素。 饼图是一维可视化，具有以特定角度绘制单个条形的极转换。 雷达图表是二维可视化，具有以与图形中心之间的特定角度和距离绘制图形元素的极转换。 三维可视化还会包括一个附加深度维度。

   斜交。 斜交转换向图形元素添加三维效果。 此转换向图形元素添加深度，但是深度纯粹是为了装饰之用。 不受特定数据值的影响。

   相同比率。 应用相同比率会指定每个刻度上的相同距离代表数据值中的相同差异。 例如，两个刻度上的 2cm 代表 1000 的差异。

预转换缩进距离 %。 如果转换后轴被切割，则在应用转换前您可能想要向图形添加缩进距离。 在将任何转换应用到坐标系前，缩进距离会按某个百分比缩小维度。 您可以按该顺序控制较低的 *x*、较高的 *x*，较低的 *y* 和较高的 *y* 维度。

后转换缩进距离 %。 如果您要更改图形的宽高比，那么在应用转换后您可以向图形添加缩进距离。 在任何转换应用到坐标系后，缩进距离按某个百分比缩小维度。 即使没有转换应用到图形，也可应用这些缩进距离。 您可以按该顺序控制较低的 *x*、较高的 *x*，较低的 *y* 和较高的 *y* 维度。

更改统计值和图形元素

您可以将一个图形元素转换为另一个类型，更改用于绘制图形元素的统计量，或指定确定在图形元素重叠时如何操作的冲突修饰符。

如何转换图形元素

1. 选择要转换的图形元素。
2. 单击“属性”调色板上的**元素**选项卡。
3. 从“类型”列表中选择一个新的图形元素类型。

表 *39:* 图形元素类型

图形元素类型	**描述**
点状图	这是标识特定数据点的标记。 点元素用于散点图以及其他相关可视化中。
时间间隔	在特定数据值处绘制矩形，并填充原点到另一个数据值之间的空隙。 间隔元素用于条形图和直方图中。
行	连接数据值的线。
路径	按数据值在数据集中出现的顺序进行连接的线。
区域	连接数据元素的线，其中线和原点间的面被填充。
多边形	多边形构成闭合数据区域。 多边形元素可以用于分箱化散点图或地图中。
模式	由箱图组成的元素，其中细线和标记表示离群值。 模式元素用于箱图。

如何更改统计量

1. 选择要更改统计值的图形元素。
2. 单击“属性”调色板上的**元素**选项卡。

如何指定冲突修饰符

冲突修饰符确定在图形元素重叠时如何操作。

1. 选择您想指定冲突修饰符的图形元素。
2. 单击“属性”调色板上的**元素**选项卡。
3. 从“修饰符”下拉列表中，选择冲突修饰符。**-auto-** 允许应用程序确定适用于图形元素类型和统计信息的冲突修饰符。

   重叠。 当拥有相同的值时，可在彼此顶部绘制图形元素。

   堆积。 堆积在数据值相同时通常会被叠加的图形元素。

   回避。 将图形元素移至出现在同一值处的其他图形元素旁边，而非进行叠加。 图形元素将被对称放置。 即，图形元素移至中心位置的另一边。 回避与聚类非常类似。

   堆叠。 将图形元素移至出现在同一值处的其他图形元素旁边，而非进行叠加。 图形元素将被不对称放置。 即，图形元素堆叠在彼此顶部，其中底部的图形元素置于刻度上的特定值处。

抖动（正态）。 使用正态分布随机更改相同数据值处的图形元素的位置。

抖动（均匀）。 使用正态分布随机更改同一数据值处的图形元素的位置。

更改图例的位置

如果图像包括图注，图注通常显示在图形右侧。 如果需要可以更改此位置。

如何更改图注位置

1. 选择图注。
2. 单击“属性”调色板上的**图注**选项卡。
3. 选择位置。

复制直观表示和直观表示数据

“一般”选项板包括复制直观表示及其数据的按钮。

图 *72:* “复制可视化”按钮

复制直观表示。 此操作会将可视化作为图像复制到剪贴板中。 有多个图像格式可用。 将图像粘贴到其他应用程序时，您可以选择“选择性粘贴”选项来选择其中一个可用的图像格式进行粘贴。

图 *73:* “复制可视化数据”按钮

复制可视化数据。 此操作复制用于绘制直观表示的基本数据。 数据将作为纯文本或 HTML 格式文本复制到剪贴板中。 当您将数据粘贴到另一个应用程序中时，您可以选择“选择性粘贴”选项以选择一个格式用于粘贴。

图形板编辑器键盘快捷键

表 *40:* 键盘快捷键

快捷键	**Function**
Ctrl+空格键	在“探索”和“编辑”模式之间进行切换
删除	删除一个可视化项目
Ctrl+Z 键	撤销
Ctrl+Y	重做
F2	显示概要用于选择图形中的项目

添加标题和脚注

对于所有图形类型，均可以添加标题、脚注或轴标签，以帮助确定图形的显示内容。

向图形添加标题

1. 从菜单中选择**编辑 > 添加图形标题**。 图形上方将显示包含 **<TITLE>** 的文本框。
2. 确保处于编辑模式。 从菜单中选择**查看 > 编辑方式**。
3. 双击 **<TITLE>** 文本。

4. 输入需要的标题并按“返回”按钮。

向图形添加脚注

1. 从菜单中选择**编辑 > 添加图形脚注**。 图形下方将显示包含 **<FOOTNOTE>** 的文本框。
2. 确保处于编辑模式。 从菜单中选择**查看 > 编辑方式**。
3. 双击 **<FOOTNOTE>** 文本。
4. 输入需要的标题并按“返回”按钮。

使用图形样式表

基本图形显示信息（比如颜色、字体、符号和线宽）都可用样式表进行控制。 IBM SPSS Modeler 提供了一个缺省样式表；不过可以根据需要修改该样式表。 例如，可以在图形中使用自己所需的共用颜色方案。 有关更多信息，请参阅主题 第 216 页的『编辑可视化』。

在图形节点中，可使用编辑模式来更改图形外观的样式。 然后，可以使用**编辑 > 样式**菜单将更改保存为样式表，以应用于随后从当前图形节点生成的所有图形，或者作为使用 IBM SPSS Modeler 生成的所有图形的新缺省样式表。

“编辑”菜单的**样式**选项上有五个可用的样式表选项：

- **切换样式表**。 这会显示不同的已存储的样式表列表，您可以选择以更改您的图形的外观。 有关更多信息，请参阅主题 第 225 页的『应用样式表』。
- **在节点中存储样式**。 此选项将保存对选定图形样式的修改，以便将来在当前流中通过同一图形节点创建图形时可以应用该样式。
- **将样式存储为缺省值**。 此选项将保存对选定图形样式的修改，以便将来在任何流中通过任何图形节点创建图形时可以应用该样式。 选中此选项后，可使用**应用缺省样式**来更改要使用相同样式的任何其他现有图形。
- **应用缺省样式**。 此选项将把选定图形的样式更改为当前保存的缺省样式。
- **应用原始样式**。 此选项将把图形样式还原为提供的原始缺省样式。

应用样式表

您可以应用一个指定可视化样式属性的可视化样式表。 例如，样式表可以定义字体、虚线、颜色等其他选项。 样式表还提供了一定的编辑快捷方式，供您手动执行。 但是，注意样式表限于样式更改。 其他更改，如图注或刻度范围的位置，不存储在样式表中。

如何应用样式表

1. 从菜单中选择：

   编辑 > 样式 > 切换样式表
2. 使用切换样式表对话框来选择一个样式表。
3. 单击**应用**将样式表应用到可视化中，而不关闭对话框。 单击**确定**应用样式表并关闭该对话框。

切换/选择样式表对话框

对话框顶部的表格列出当前可用的所有直观表示样式表。 一些样式表已预先安装，而其他可能创建在 IBM SPSS 可视化设计器 （独立的产品）中。

对话框底部显示带有样本数据的实例可视化。 选择一个样式表将其样式应用到示例直观表示。 这些示例可帮助您确定样式表将如何影响您的实际直观表示。

对话框还提供以下选项。

现有样式。 缺省情况下，样式表可以覆盖可视化中的所有样式。 您可以更改此行为。

- **覆盖所有样式**。 当应用样式表时，覆盖可视化中的所有样式，包括在当前编辑会话过程中在可视化中修改的那些样式。
- **保持修改样式**。 当应用样式表时，只覆盖那些在当前编辑会话过程中在可视化中未修改的样式。 保持在当前编辑会话过程中修改的样式。

管理。 管理计算机上的可视化模板，及样式表与地图。 您可以导入、导出、重命名和删除本地计算机上的可视化模板，样式表与地图。 有关更多信息，请参阅第 164 页的『管理模板、样式表和地图文件』主题。

位置。 更改可视化模板，样式表与地图的存储位置。 当前位置列出在按钮右侧。 有关更多信息，请参阅第 163 页的『设置模板、样式表和地图位置』主题。

打印、保存、复制和导出图形

每个图形都具有若干选项，这些选项可用于保存或打印图形或将图形导出为另外一个格式。 这些选项的大部分均可在“文件”菜单找到。 另外，还可以从“编辑”菜单选择复制图形、其中的数据或 Microsoft Office 图形对象以便在另一个应用程序中使用。

打印

要打印图形，请使用 **打印** 菜单项或按钮。 打印之前，可以使用**页面设置**和**打印预览**来设置打印选项并预览输出。

保存图形

要将图形保存到 IBM SPSS Modeler 输出文件 (`*.cou`)，请从菜单中选择**文件 > 保存**或**文件 > 另存为**。

或者

要将图形保存到存储库中，请从菜单中选择**文件 > 存储输出**。

复制图形

要复制图形以在其他应用程序（比如 MS Word 或 MS PowerPoint）中使用，请从菜单中选择**编辑 > 复制图形**。

复制数据

要复制数据以在其他应用程序（比如 MS Excel 或 MS Word）中使用，请从菜单中选择**编辑 > 复制数据**。 缺省情况下，数据采用 HTML 格式。 粘贴时，使用另一个应用程序中的**选择性粘贴**来查看其他格式选项。

复制 Microsoft Office 图形对象

您可以将图形复制为 Microsoft Office 图形对象，并在 Excel 或 PowerPoint 之类的 Microsoft Office 应用程序中使用。 要复制图形，请从菜单中选择**编辑 > 复制 Microsoft Office 图形对象**。 内容将复制到剪贴板，缺省情况下将使用二进制格式。 粘贴时，使用 Microsoft Office 应用程序中的**选择性粘贴**来指定其他格式选项。

请注意，某些内容可能不支持此功能，在此情况下，将禁用**复制 Microsoft Office 图形对象**菜单选项。 还需注意，图形在粘贴到 Office 应用程序后外观可能有所不同，但图形数据保持不变。

有六种类型的图形输出可以复制并粘贴到 Excel：简单条形图、堆积条形图、简单箱图、簇状箱图、简单散点图和分组散点图。 如果将“面板”和“动画”选项用于上述任何图形类型，**复制 Microsoft Office 图形对象**选项将在 SPSS Modeler 中处于禁用状态。 而对于其他设置，例如“可选外观”或“重叠”，该选项受到部分支持。 请参阅下表以了解详细信息：

表 *41:* 复制 *Microsoft* 图形对象支持

图形输出模板	Modeler 图形节点	Modeler 图形类型	基本设置	可选外观	重叠	Microsoft 图形对象复制支持	注释
简单条形图	图形板	条形图	是	否	N/A	是	
		计数条形图	是	否	N/A	是	
	分布	条形图	是	N/A	否	是	

表 *41:* 复制 *Microsoft* 图形对象支持 *(*继续*)*

图形输出模板	Modeler 图形节点	Modeler 图形类型	基本设置	可选外观	重叠	Microsoft 图形对象复制支持	注释
堆积条形图	图形板	条形图	是	是	N/A	是，但存在限制	仅对“可选外观”中的分类变量为 Yes。
		计数条形图	是	是	N/A	是，但存在限制	仅对“可选外观”中的分类变量为 Yes。
	分布	条形图	是	N/A	是	是	
箱图	图形板	箱图	是	否	N/A	是，但存在限制	仅在 Windows 上为 Yes。
		箱图	是	是	N/A	否	
簇状箱图	图形板	簇状箱图	是	否	N/A	是，但存在限制	仅在 Windows 上为 Yes。
		簇状箱图	是	是	N/A	否	
简单散点图	图形板	气泡图	是	否	N/A	是，但存在限制	仅对 X 和 Y 字段中的连续变量以及“大小”中的分类变量为 Yes。
		散点图	是	否	N/A	是，但存在限制	仅对 X 和 Y 字段中的连续变量为 Yes。
	统计图	点	是	N/A	否	是，但存在限制	仅对 X 和 Y 字段中的连续变量为 Yes。
分组散点图	图形板	气泡图	是	是	N/A	否	
		散点图	是	是	N/A	是，但存在限制	仅对 X 和 Y 字段中的连续变量以及“可选外观”中的分类变量为 Yes。
	统计图	点	是	N/A	是	是，但存在限制	仅对 X 和 Y 字段中的连续变量以及“重叠”选项中的分类变量为 Yes。

导出图形

导出图形选项使您能够以下列某种格式导出图形：位图 (.bmp)、JPEG (.jpg)、PNG (.png)、HTML (.html)、PDF (.pdf) 或 ViZml 文档 (.xml) 以在其他应用程序中使用。

注: 当选择 PDF 选项时，将作为高分辨率 PDF 文件导出图形，这些文件会裁剪为图形的大小。

要导出图形，请从菜单中选择**文件 > 导出图形**，然后选择格式。

导出表

使用**导出表**选项使您能够以下列某种格式导出表：制表符分隔 (.tab)、逗号分隔 (.csv) 或 HTML (.html)。

要导出表，请从菜单中选择**文件 > 导出表**，然后选择格式。

第 6 章 输出节点

输出节点概述

输出节点提供了用于获取数据和模型的相关信息的方法。 还提供了以各种格式导出数据以与其他软件工具相互作用的机制。

下列输出节点可用：

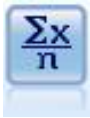
“表”节点以表格式显示数据，这些数据还可以写入到文件中。 每当您需要检查数据值或者采用可轻松阅读的格式导出这些数据值时，此节点非常有用。

“矩阵”节点创建一个表，用于显示字段之间的关系。 最常用于显示两个符号字段之间的关系，但也可以显示标志字段或数字字段之间的关系。

“分析”节点评估预测模型生成准确预测的能力。 “分析”节点执行一个或多个模型块的预测值和实际值之间的各种比较。 它们还可以将预测模型进行相互比较。

“数据审核”节点提供有关数据的全面概览，包括每个字段的汇总统计、直方图和分布以及有关离群值、缺失值和极值的信息。 结果显示在易于读取的矩阵中，可进行排序并且可用于生成完整大小的图形和数据准备节点。

“变换”节点允许您先选择并直观预览变换的结果，然后再将其应用于所选字段。



“统计”节点提供有关数字字段的基本汇总信息。 它计算各个字段的汇总统计以及字段间的相关性。

“平均值”节点在独立组之间或相关字段对之间进行平均值比较，以检验是否存在显著差别。 例如，您可以比较开展促销 前后的平均收入，或者将来自未接受促销客户的收入与接受促销客户的收入进行比较。

“报告”节点可创建格式化报告，其中包含固定文本、数据及得自数据的其他表达式。 您可以使用文本模板指定报告的格式，以定义固定文本和数据输出构造。 通过在模板中使用 HTML 标记以及在“输出”选项卡上设置选项，可以提供定制文本格式。 您可以通过在模板中使用 CLEM 表达式来包含数据值和其他条件输出。

设置全局量节点扫描数据并计算可在 CLEM 表达式中使用的汇总值。 例如，您可以使用此节点来计算名为年龄的字段的统计信息，然后通过插入函数 `@GLOBAL_MEAN(age)`，在 CLEM 表达式中使用年龄的总体平均值。

“模拟拟合”节点检查每个字段中数据的统计分布，并生成（或更新）“模拟生成”节点，同时将最佳拟合分布分配给每个字段。 然后，可以使用“模拟生成”节点来生成模拟数据。

“模拟评估”节点对指定的预测目标字段进行评估，并显示有关该目标字段的分布和相关性信息。

管理输出

输出管理器可显示在 IBM SPSS Modeler 会话期间生成的图表、图形和表格。 您始终可以通过在管理器中双击输出来将它重新打开，而不必重新运行相应的流或节点。

查看输出管理器

打开“查看”菜单，然后选择**管理器**。 单击**输出**选项卡。

从输出管理器中，可以：

- 显示现有输出对象，如直方图、评估图和表格。
- 重命名输出对象。
- 将输出对象保存到磁盘或 IBM SPSS Collaboration and Deployment Services Repository （如果可用）。
- 将输出文件添加到当前项目中。
- 从当前会话中删除未保存的输出对象。
- 打开已保存的输出对象或从 IBM SPSS Collaboration and Deployment Services Repository 进行检索（如果可用）。

要访问这些选项，请在“输出”选项卡上的任意位置单击右键。

查看输出

屏幕上的输出显示在输出浏览器窗口中。 输出浏览器窗口具有它自己的菜单集，使用这些菜单，可以打印或保存输出，也可以将输出导出为其他格式。 请注意，具体选项会因输出类型的不同而不同。

打印、保存和导出数据。 下面提供了详细信息：

- 要打印输出，请使用**打印**菜单选项或按钮。 打印之前，可以使用**页面设置**和**打印预览**来设置打印选项并预览输出。
- 要将输出保存到 IBM SPSS Modeler 输出文件 (.cou)，请从“文件”菜单中选择**保存**或**另存为**。
- 要以另一种格式（如文本或 HTML）保存输出，请从“文件”菜单中选择**导出**。 有关更多信息，请参阅主题第 232 页的『导出输出』。

  请注意，仅当输出包含可以通过该方式正确导出的数据时，才能选择这些格式。 例如，决策树的内容可以导出为文本，但 K-Means 模型的内容作为文本没有意义。
- 要将输出保存在共享存储库中使得其他用户可使用 IBM SPSS Collaboration and Deployment Services Deployment Portal 查看，从“文件”菜单中选择 **发布到网络**。 请注意，该选项需要 IBM SPSS 协作和部署服务 的单独许可证。

选择单元格和列。 “编辑”菜单包含用于选择、取消选择和复制单元格和列且适合于当前输出类型的各种选项。 有关更多信息，请参阅第 232 页的『选择单元格和列』。

生成新节点。 使用“生成”菜单，可以根据输出浏览器的内容生成新节点。 这些选项因输出类型以及当前在输出中选择的项的不同而有所差异。 有关特定输出类型的节点生成选项的详细信息，请参见该输出的文档。

发布到 Web

“发布到网络”功能可以将特定类型的流输出发布到形成 IBM SPSS 协作和部署服务 基础的中央共享 IBM SPSS Collaboration and Deployment Services Repository。 如果您使用此选项，要查看该输出的其他用户可使用因特网访问和 IBM SPSS 协作和部署服务 帐户进行查看，而无需安装 IBM SPSS Modeler。

下表列出了支持“发布到网络”功能的 IBM SPSS Modeler 节点。 这些节点的输出以输出对象格式 (*.cou*) 存储到 IBM SPSS Collaboration and Deployment Services Repository 中，并可直接在 IBM SPSS Collaboration and Deployment Services Deployment Portal 中查看。

只有在用户机器上安装了相关应用程序（例如 IBM SPSS Modeler， 用于流对象）时，才能查看其他类型的输出。

表 *42:* 支持发布到网络的节点

节点类型	**节点**
图形	all
输出	表
	矩阵
	数据审核
	变换
	平均值
	分析
	统计信息
	报告 (HTML)
IBM SPSS Statistics	Statistics 输出

发布输出到网络

要发布输出到网络：

1. 在 IBM SPSS Modeler 流中，执行表中列出的某个节点。 这会在新窗口中创建输出对象（例如，表、矩阵或报告对象）。
2. 从输出对象窗口中，选择：

   文件 > **发布到 Web**

   注意：如果想导出简单 HTML 文件用于标准 Web 浏览器，请从“文件”菜单选择**导出**并选择 **HTML**。
3. 连接到 IBM SPSS Collaboration and Deployment Services Repository。

   在连接成功后，将显示“存储库：存储”对话框，其中提供了多种存储选项。
4. 在您选择所需的存储选项后，单击**存储**。

在网络上查看发布的输出

使用此功能需要设置有 IBM SPSS 协作和部署服务 帐户。 如果您为要查看的对象类型安装有相关应用程序（例如 IBM SPSS Modeler 或 IBM SPSS Statistics），则输出将显示在应用程序本身而不是浏览器中。

要在网络上查看发布的输出：

1. 将浏览器指向 *http://<repos_host>:<repos_port>/peb*

   其中 *repos_host* 和 *repos_port* 为 IBM SPSS 协作和部署服务 主机的主机名和端口号。
2. 输入您的 IBM SPSS 协作和部署服务 帐户的详细登录信息。
3. 单击**内容存储库**。

4. 导航到或搜索您要查看的对象。
5. 单击对象名称。 对于某些对象类型，例如图形，在浏览器中呈现对象可能需要一些时间。

在 HTML 浏览器中查看输出

从线性、Logistic 和主成分分析/因子模型块的“高级”选项卡中，可以用单独的浏览器（如 Internet Explorer）查看所显示的信息。 信息以 HTML 形式输出，使您能够将其保存并在其他位置（例如，在企业内部网或因特网站点上）并进行复用。

要在浏览器中显示信息，请单击模型块“高级”选项卡对话框左上角的模型图标下的启动按钮。

导出输出

在输出浏览器窗口中，可以选择将输出导出为另一格式（如文本或 HTML）。 导出格式因输出类型的不同而不同，但一般类似于在用于生成输出的节点中选择**保存到文件**时可以使用的文件类型选项。

注: 仅当输出包含可以通过该方式正确导出的数据时，才能选择这些格式。 例如，决策树的内容可以导出为文本，但 K-Means 模型的内容作为文本没有意义。

导出输出

1. 在输出浏览器中，打开“文件”菜单并选择**导出**。 然后，选择要创建的文件类型：
   - **制表符分隔 (*.tab)**。 此选项生成包含数据值的格式文本文件。 此样式经常用于生成可导入到其他应用程序中的信息的纯文本表示形式。 此选项适用于“表格”节点、“矩阵”节点和“平均值”节点。
   - **逗号分隔 (*.dat)**。 此选项生成包含数据值的逗号分隔的文本文件。 此样式经常用于快速生成可导入到电子表格或其他数据分析应用程序中的数据文件。 此选项适用于“表格”节点、“矩阵”节点和“平均值”节点。
   - **转置制表符分隔 (*.tab)**。 此选项与“制表符分隔”选项相同，但是数据进行了转置，以便行表示字段，列表示记录。
   - **转置逗号分隔 (*.dat)**。 此选项与“逗号分隔”选项相同，但是数据进行了转置，以便行表示字段，列表示记录。
   - **HTML (*.html)**。 此选项将 HTML 格式的输出写至文件中。

选择单元格和列

许多节点（包括“表格”节点、“矩阵”节点和“平均值”节点）生成表格输出。 可通过相似方式查看并操纵这些输出表，包括选择单元格、将表格的全部或部分复制到剪贴板、根据当前选择生成新节点，以及保存和打印表格。

选择单元格。 请单击单元格以将其选中。 要选择矩形单元格区域，请单击所需区域的一个角，拖动鼠标至该区域的另一个角，然后松开鼠标按钮。 要选择一整列，请单击列标题。 要选择多列，请按住 Shift 键或 Ctrl 键并单击列标题。

进行新选择时，旧选择会被清除。 通过在按住 Ctrl 键的同时进行选择，可以将新选择添加到任何现有选择中，而不是清除旧选择。 可以使用此方法选择多个非连续的表格区域。 “编辑”菜单还包含**全选**和**清除选择**选项。

对列重新排序。 使用“表格”节点和“平均值”节点的输出浏览器，可以移动表格中的列，方法是：单击列标题，并将它拖至所需位置。 一次只能移动一列。

表节点

表节点可以创建能够列出数据中的值的表。 该表中包含了流中的所有字段和所有值，从而可以方便检查数据值或以易于读取的格式进行导出。 此外，您还可以突出显示满足特定条件的记录。

注: 除非您在使用小数据集，否则建议您选择数据的子集传递到“表”节点。 当记录数超过显示结构可以包含的大小（例如：1 亿行）时，“表”节点无法准确显示。

表节点的“设置”选项卡

突出显示满足以下条件的记录： 通过输入适用于要突出显示的记录的 CLEM 表达式，可以突出显示表格中的记录。 只有在选中**输出到屏幕**时，此选项才会启用。

表节点的“格式”选项卡

“格式”选项卡包含用于按字段指定格式的选项。 “类型”节点共享此选项卡。 有关更多信息，请参阅主题 第 111 页的『字段格式设置选项卡』。

输出节点的“输出”选项卡

对于生成表样式输出的节点，使用“输出”选项卡可指定结果的格式和位置。

输出名称。 指定执行节点时使用的输出名称。 **自动** 根据生成输出的节点选择名称。 （可选）可以选择**定制**以指定其他名称。

输出到屏幕（缺省选项）。 创建要在线查看的输出对象。 当执行输出节点时，该输出对象将显示在管理器窗口的“输出”选项卡上。

输出到文件。 执行节点时将输出保存到文件。 如果选择此选项，请输入文件名（或导航到某目录，并使用文件选择器按钮指定文件名）并选择文件类型。 请注意，有些文件类型可能不适用于某些特定类型的输出。

：

来自输出节点的输出数据根据以下规则进行编码：

- 在执行输出节点时，流编码值（在“流选项”选项卡上设置）将设置为输出。
- 在生成输出之后，即使流编码更改，其编码也不会更改。
- 导出输出节点输出时，将使用当前定义的编码来保存输出文件。 在创建输出后，即使更改流编码，也不会影响生成的输出。

请注意这些规则的以下例外：

- 所有 HTML 导出均采用 UTF-8 格式编码。
- 来自“扩展”输出节点的输出由定制用户脚本生成。 因此，编码由脚本控制。

以下选项可用于将输出保存到文件：

- **数据（制表符分隔）(*.tab)**。 此选项生成包含数据值的格式文本文件。 此样式经常用于生成可导入到其他应用程序中的信息的纯文本表示形式。 此选项适用于“表格”节点、“矩阵”节点和“平均值”节点。
- **数据（逗号分隔）(*.dat)**。 此选项生成包含数据值的逗号分隔的文本文件。 此样式经常用于快速生成可导入到电子表格或其他数据分析应用程序中的数据文件。 此选项适用于“表格”节点、“矩阵”节点和“平均值”节点。
- **HTML (*.html)**。 此选项将 HTML 格式的输出写至文件中。 对于来自“表格”节点、“矩阵”节点或“均值”节点的表格输出，一组 HTML 文件包含一个内容面板，该面板在 HTML 表格中列出了字段名称和数据。 如果表格中的行数超过**每页行数**规范，则可将表格拆分为多个 HTML 文件。 在这种情况下，该内容面板包含指向所有表格页的链接，并提供导航表格的方法。 对于非表格输出，会创建一个包含节点结果的 HTML 文件。

  注： 如果 HTML 输出只包含第一页的格式，请选择**分页输出**并调整**每页行数**指定以在单一页上包括所有输出。 或者，如果节点（如报告节点）的输出模板包含自定义的 HTML 标记，则一定要指定**自定义**作为格式类型。
- **文本文件 (*.txt)**。 此选项生成包含输出的文本文件。 此样式经常用于生成可导入到其他应用程序（如文字处理器或演示软件）中的输出。 此选项对于某些节点不适用。
- **输出对象 (*.cou)**。 对于以这种格式保存的输出对象，可在 IBM SPSS Modeler 中打开并查看、添加到项目，以及使用 IBM SPSS Collaboration and Deployment Services Repository 发布和跟踪。

输出视图。 对于“均值”节点，可以指定缺省情况下是显示简单输出还是高级输出。 请注意，也可以在浏览生成的输出时在这些视图间切换。 有关更多信息，请参阅主题 第 248 页的『“平均值”节点输出浏览器』。

格式。 对于“报告”节点，可以选择是自动设置输出格式，还是使用模板中包括的 HTML 设置输出格式。 选择**自定义**可允许使用模板中的 HTML 格式。

标题。 对于报告节点，可以指定将显示在报告输出顶部的可选标题文本。

突出显示插入的文本。 对于报告节点，选择此选项可在报告模板中突出显示由 CLEM 表达式生成的文本。 有关更多信息，请参阅主题 第 249 页的『报告节点的模板选项卡』。 建议不要在使用**定制**格式时时有此选项。

每页行数。 对于报告节点，指定在输出报告的**自动**格式设置期间要在每页上包括的行数。

转置数据。 此选项在导出前转置数据，以便行表示字段，列表示记录。

注: 对于大型表格，上述选项的效用可能会较差，尤其是在使用远程服务器时。 在这种情况下，使用文件输出节点可提供更好的性能。 有关更多信息，请参阅主题 第 277 页的『平面文件导出节点』。

表格浏览器

表格浏览器显示表格数据，并且可以用于执行标准操作，包括选择和复制单元格、重新为列排序，以及保存和打印表格。 有关更多信息，请参阅主题 第 232 页的『选择单元格和列』。 这些操作与预览节点中数据的操作相同。

导出表数据。 要从表格浏览器导出数据，请选择：

文件 > 导出

有关更多信息，请参阅主题 第 232 页的『导出输出』。

数据以系统缺省编码格式导出，编码格式可在 Windows 控制面板中指定，如果以分布式模式运行，则在服务器计算机上指定。

搜索表。 主工具栏上的搜索按钮（使用双筒望远镜图标）可激活搜索工具栏，从而使您可以在表格中搜索特定值。 您可以在表格中向前或向后搜索，可以指定区分大小写的搜索（**Aa** 按钮），并且可以使用中断搜索按钮中断正在进行的搜索。

生成新节点。 “生成”菜单包含节点生成操作。

- **选择节点（“记录”）**。 生成选择节点，该节点对表格中任何选中的单元格的记录进行选择。
- **选择（“和”）**。 生成选择节点，该节点选择包含了表格中所有 选中值的记录。
- **选择（“或”）**。 生成选择节点，该节点选择包含了表格中任何 选中值的记录。
- **派生（“记录”）**。 生成用于新建标志字段的“派生”节点。 标志字段包含 *T*（表示表格中任何选中的单元格的记录）和 *F*（表示其余记录）。
- **派生（“和”）**。 生成用于新建标志字段的“派生”节点。 标志字段包含 *T*（表示包含了表格中所有 选中值的记录）和 *F*（表示其余记录）。
- **派生（“或”）**。 生成用于新建标志字段的“派生”节点。 标志字段包含 *T*（表示包含了表格中任何 选中值的记录）和 *F*（表示其余记录）。

“矩阵”节点

使用“矩阵”节点可以创建显示字段间关系的表格。 它最常用于显示两个分类字段（标志、名义或有序）之间的关系，但也可用于显示连续（数值范围）字段之间的关系。

矩阵节点的设置选项卡

“设置”选项卡用于为矩阵结构指定选项。

字段。 从下列选项中选择字段选择类型：

- **选定字段**。 使用此选项可以为矩阵的行和列分别选择一个分类字段。 矩阵的行和列由所选分类字段的值列表定义。 矩阵的单元格包含了如下所列的汇总统计量中选中的部分。
- **所有标志（true 值）**。 此选项请求一个满足如下条件的矩阵：数据中的每个标志字段都包含一个行和一个列。 矩阵的单元格包含每个标志组合的双正数的计数。 换言之，对于与所购面包对应的行和与所购干酪对应的列，该行和列交叉处的单元格包含所购面包和所购干酪均为真值的记录的个数。

- **所有数字**。 此选项请求一个满足如下条件的矩阵：每个数字字段都包含一个行和一个列。 矩阵的单元格表示相应字段对的交叉乘积的总和。 换言之，对于矩阵中的每个单元格，会将该格内每条记录的行字段值和列字段值相乘，然后对格内所有记录的乘积值求和。

包含缺失值。 在行和列输出中包括用户缺失值（空白值）和系统缺失值 ($null$)。 例如，如果已将值 *N/A* 定义为所选列字段的用户缺失值，则表格中将包括标签为 *N/A* 的单独列（假设此值实际出现在数据中），就像任何其他类别一样。 如果取消选择此选项，则无论 *N/A* 列的出现频率高低，都会将它排除。

注意：用于包括缺失值的选项仅在选定字段为交叉列表形式时适用。 空白值映射到 $null$，并且在模式为**所选**且内容设置为**函数**时从函数字段的合计中排除，在模式设置为**所有数字**时从所有数字字段的合计中排除。

单元格内容。 如果已经选择上面的**所选**字段，则可以指定要在矩阵的单元格中使用的统计量。 选择基于计数的统计量，或选择重叠字段以根据行和列字段的值为汇总数字字段的值。

- **交叉列表**。 单元格值是用于统计多少记录具有对应值组合的计数和/或百分比。 您可以使用“外观”选项卡上的选项指定所需的交叉列表汇总。 全局卡方值也会和显著性一起显示。 有关更多信息，请参阅主题 第 236 页的『矩阵节点输出浏览器』。
- **函数**。 如果选择汇总函数，那么单元格值是所选重叠字段值的函数（对于具有适当的行值和列值的情况）。 例如，如果行字段为地区，列字段为产品，重叠字段为收入，则位于东北行和小器具列中的单元格将包含东北地区出售小器具所获得的收入的总和（或平均值、最小值或最大值）。 缺省汇总函数是 **Mean**。 您可以选择其他函数来汇总该函数字段。 选项包括 **Mean**、**Sum**、**SDev**（标准差）、**Max**（最大值）和 **Min**（最小值）。

矩阵节点的外观选项卡

使用“外观”选项卡，可以控制矩阵的排序和突出显示选项，以及为交叉列表矩阵提供的统计量。

行和列。 控制矩阵中行和列标题的排序。 缺省值为**不排序**。 选择**升序**或**降序**可按指定的方向为行和列标题排序。

Overlay）。 用于突出显示矩阵中的极值。 值根据单元格计数（对于交叉列表矩阵）或计算出的值（对于函数矩阵）突出显示。

- **突出显示顶部**。 您可以请求突出显示矩阵中大小排在前几位的值（以红色显示）。 指定要突出显示的值个数。
- **突出显示底部**。 您也可以请求突出显示矩阵中大小排在后几位的值（以绿色显示）。 指定要突出显示的值个数。

注意：对于这两个突出显示选项，结的存在可能导致突出显示的值比请求突出显示的值多。 例如，如果您有一个矩阵，而该矩阵在单元格中包括六个零，当您请求**突出显示后 5 个值**时，则将突出显示全部六个零。

交叉列表单元格内容。 对于交叉列表，可以指定包含在交叉列表的矩阵中的汇总统计量。 当在“设置”选项卡上选择了**所有数字**或**函数**选项时，这些选项不可用。

- **计数**。 单元格包括其行值具有对应列值的记录数。 这只是缺省单元格内容。
- **期望值**。 单元格中记录数的期望值（假设行和列之间没有关系）。 期望值基于以下公式：

$$p(row\ value) * p(column\ value) * total\ number\ of\ records$$

- **残差**。 观测值和期望值之间的差值。
- **行的百分比**。 其行值具有对应列值的所有记录的百分比。 行百分比的和为 100。
- **列的百分比**。 其列值具有对应行值的所有记录的百分比。 列百分比的和为 100。
- **占总数的百分比**。 具有行值和列值组合的所有记录的百分比。 整个矩阵的百分比和为 100。
- **包括行和列总计**。 将行和列添加到行和列合计的矩阵。
- **应用设置**。 （仅限输出浏览器）使您可以更改“矩阵”节点输出的外观，且不必关闭并重新打开输出浏览器。 在输出浏览器的此选项卡上进行更改，单击此按钮，然后选择“矩阵”选项卡以查看更改的效果。

矩阵节点输出浏览器

矩阵浏览器显示交叉列表数据，并可用于针对矩阵执行操作，包括选择单元格、将矩阵全部或部分复制到剪贴板、根据矩阵选择生成新节点，以及保存和打印矩阵。 矩阵浏览器还可用于显示某些特定模型的输出，如 Oracle 的朴素贝叶斯模型。

“文件”和“编辑”菜单提供用于打印、保存和导出输出以及用于选择和复制数据的常用选项。 有关更多信息，请参阅主题 第 230 页的『查看输出』。

卡方。 对于含两个类别字段的交叉列表，还会在该表格的下面显示全局 Pearson 卡方。 此检验指出两个字段之间不存在关系的概率（根据的是不存在关系时观测计数和期望计数之间的差值）。 例如，如果客户满意率和商店位置之间没有关系，那么您将预期所有商店的客户满意率相似。 但是，如果某些特定商店报告的客户满意率总是比其他商店高，那么您可能会怀疑这并非巧合。 差值越大，则单纯由随机抽样误差导致此结果的概率就越小。

- 卡方检验指出两个字段之间不存在关系的概率，如果两个字段之间不存在关系，那么观测频率和期望频率之间的任何差值完全归咎于随机效应。 如果此概率非常小（通常小于 5%），那么说明两个字段之间存在显著的联系。
- 如果只有一列或一行（单因素卡方检验），那么自由度等于单元格数减一。 对于双因素卡方，自由度等于行数和列数均减一之后二者的乘积。
- 如果任何期望单元格频率小于五，那么解释卡方统计量时请务必小心。
- 卡方检验只适用于含两个字段的交叉列表。 （当在“设置”选项卡上选择**所有标志**或**所有数字**时，此检验不会显示。）

生成菜单。 “生成”菜单包含节点生成操作。 这些操作只适用于交叉列表矩阵，并且必须至少在矩阵中选择了一个单元格。

- **"选择"节点**。 生成“选择”节点，该节点选择与矩阵中的任何选定单元格匹配的记录。
- **派生节点（标志）**。 生成用于新建标志字段的“派生”节点。 标志字段包含 *T*（表示与矩阵中的任何选定单元格匹配的记录）和 *F*（表示其余记录）。
- **派生节点（集合）**。 生成用于新建名义字段的“派生”节点。 对于矩阵中的每个连续的选定单元格集，名义字段都包含一个类别。

“分析”节点

通过使用“分析”节点，您可以对模型生成准确预测的能力进行评估。 “分析”节点将对一个或多个模型块的预测值和实际值（您的目标字段）进行各种比较。 “分析”节点也可用于将一些预测模型和其他预测模型进行比较。

执行“分析”节点时，对于已执行的流中的每个模型块，会自动将分析结果摘要添加至“摘要”选项卡上的“分析”部分。 详细分析结果显示在管理器窗口的“输出”选项卡中，或者可以直接写入文件。

注: 因为“分析”节点将预测值与实际值进行比较，所以只适用于需要目标字段的受监督模型。 对于非受监督模型（例如，聚类算法），没有实际结果可用作比较基础。

“分析”节点的“分析”选项卡

通过使用“分析”选项卡，您可以指定分析详细信息。

符合矩阵（用于符号或分类目标）。 显示分类目标（标志、名义或有序）的各个生成（预测）字段与其目标字段之间的匹配模式。 将显示一个表格，其中包含实际值所定义的行和预测值所定义的列，以及每个单元格中符合该模式的记录数。 这用于确定预测中的系统错误。 如果生成了多个与同一输出字段相关的字段，但这些字段由不同模型生成，那么将为这些字段相同和不相同的情况进行计数并显示总计值。 对于它们相同的情况，将显示另一组正确/错误统计量。

性能评估。 使用分类输出显示模型的性能评估统计量。 此统计量（针对输出字段的每个类别报告）是以位为单位对模型（用于预测属于该类别的记录）的平均信息内容的测量。 考虑到分类问题的难度，因此，罕见类别的准确性预测会比常见类别的准确性预测获得更高的性能评估。 对于某个类别，如果模型效果比随机猜测差，那么该类别的性能评估指数将为 0。

评估度量（AUC 和 Gini，仅限二元分类器）。 对于二元分类器，此选项将报告 AUC（曲线下面积）和 Gini 系数评估度量。 将对每个二元模型共同计算这两个评估度量。 将在表中的分析输出浏览器中报告这些度量的值。

AUC 评估度量按照 ROC（受试者工作特征）曲线下方的面积进行计算，它是分类器预期性能的标量表示。 AUC 始终介于 0 到 1 之间，数字越大表示分类器越好。 坐标 (0,0) 与 (1,1) 之间的对角线 ROC 曲线表示随机分类器，并且其 AUC 为 0.5。 因此，实际分类器的 AUC 不会小于 0.5。

有时，Gini 系数评估度量用作 AUC 的替代评估度量，并且这两个度量密切相关。 Gini 系数计算为 ROC 曲线与对角线之间的面积的两倍，或者按照 Gini = 2AUC - 1 进行计算。 Gini 系数始终介于 0 到 1 之间，数字越大表示分类器越好。 对于 ROC 曲线在对角线下方的不可能事件，Gini 系数为负。

置信度数字（如果可用）。 对于生成置信度字段的模型，此选项将报告关于置信度值及其与预测的关系的统计。 此选项有两项设置：

- **以下项的阈值：** 报告准确性达到指定百分比的置信度级别。
- **提高准确性**。 报告准确性提高指定系数的置信度级别。 例如，如果总准确性为 90%，而此选项设置为 2.0，那么所报告的值将是准确性为 95% 时所需的置信度。

使用以下内容查找预测值/预测变量字段。 确定预测字段与原始目标字段匹配的方式。

- **模型输出字段元数据**。 基于模型字段信息使预测字段与目标相匹配，即便在重命名预测字段的情况下也可以进行匹配。 通过使用“类型”节点，还可以从“值”对话框访问任何预测字段的模型字段信息。 有关更多信息，请参阅主题 第 108 页的『使用值对话框』。
- **字段名称格式**。 根据命名约定匹配字段。 例如，C5.0 模型块为目标 *response* 生成的预测值必须位于字段 *$C-response* 中。

按分区分隔。 如果使用分区字段将记录分割为训练样本、检验样本和验证样本，那么选择此选项可单独为每个分区显示结果。 有关更多信息，请参阅主题 第 131 页的『分区节点』。

注：按分区分隔时，将从分析中排除分区字段中具有空值的记录。 由于分区节点不生成空值，因此，如果使用分区节点，则这永远不会成为问题。

用户定义的分析。 您可以指定要在评估模型时使用的分析计算。 使用 CLEM 表达式可指定应为每条记录计算的内容以及如何将记录级别的评分合并到总评分中。 使用函数 **@TARGET** 和 **@PREDICTED** 可分别引用目标（实际输出）值和预测值。

- **If**。 指定需要根据某一条件使用不同计算时的条件表达式。
- **Then**。 指定条件为 true 时的计算。
- **Else**。 指定条件为 false 时的计算。
- **使用**。 选择用于根据单个评分计算总评分的统计。

按字段细分分析。 显示可用于分解分析的分类字段。 除整体分析外，将为每个分解字段的每个类别报告单独的分析。

分析输出浏览器

分析输出浏览器显示了“分析”节点的执行结果。 “文件”菜单中提供了常用的保存、导出和打印选项。 有关更多信息，请参阅主题 第 230 页的『查看输出』。

首次浏览分析输出时，结果会展开。 要在查看结果后将其隐藏，请使用项目左侧的扩展器控件将要隐藏的特定结果折叠，或单击**全部折叠**按钮以折叠所有结果。 要在折叠结果后再次对其进行查看，请使用项目左侧的展开器控件显示结果，或单击**全部展开**按钮以显示所有结果。

输出字段的结果。 对于具有所生成的模型创建的相应预测字段的每个输出字段，分析输出都包含一个相应部分。

正在比较。 在输出字段部分中，这是与该输出字段关联的每个预测字段的子部分。 对于分类输出字段，此部分的顶级位置包含一张表，其中显示正确预测和不正确预测的数目和百分比以及流中的记录总数。 对于数字输出字段，此部分显示以下信息：

- **最小误差**。 显示最小误差（观测值和预测值之间的差值）。
- **最大误差**。 显示最大误差。

- **平均误差**。 显示所有记录的误差的平均值（均数）。 这指示模型中是否有系统**偏差**（过高估计的趋势强于过低估计的趋势，或相反）。
- **平均绝对误差**。 显示所有记录的误差绝对值的平均值。 指出误差的平均量级（不考虑方向）。
- **标准差**。 显示误差的标准差。
- **线性相关性**。 显示预测值和实际值之间的线性相关。 此统计量介于 –1.0 和 1.0 之间。 值接近于 +1.0 表示强正相关，因此，高预测值与高实际值相关，而低预测值与低实际值相关。 值接近于 –1.0 表示强负相关，因此，高预测值与低实际值相关，而低预测值与高实际值相关。 值接近于 0.0 表示弱相关，因此，预测值或多或少地独立于实际值。 注：此处的空白条目表示由于实际或预测值为常量，因此在该案例中无法计算线性相关。
- **出现次数**。 显示分析中使用的记录数。

符合矩阵。 对于分类输出字段，如果在分析选项中请求了符合矩阵，那么此处会显示一个包含该矩阵的子部分。 行表示实际观测值，而列表示预测值。 表中的单元格表示每个预测值和实际值组合的记录数。

性能评估。 对于分类输出字段，如果在分析选项中请求了性能评估统计量，那么此处会显示性能评估结果。每个输出类别均与其性能评估统计量一起列出。

置信度值报告。 对于分类输出字段，如果在分析选项中请求了置信度值，那么此处会显示这些值。 对于模型置信度值，将报告下列统计量：

- **范围**。 显示流数据中记录的置信度值的范围（最小值和最大值）。
- **平均值正确**。 显示正确分类的记录的平均置信度。
- **平均值不正确**。 显示未正确分类的记录的平均置信度。
- **始终正确高于**。 显示预测始终正确的置信度阈值的下限，并显示符合此条件的案例所占百分比。
- **始终错误低于**。 显示预测始终错误的置信度阈值的上限，并显示符合此条件的案例所占百分比。
- **X% 准确性高于**。 显示准确度为 *X%* 时的置信度级别。 *X* 是分析选项中为**阈值**指定的近似值。 对于某些模型和数据集，无法选择提供精确阈值（在选项中指定的）的置信度值（通常是具有接近阈值的相同置信度值的类似观测值的聚类所致）。 所报告的阈值是最接近于指定准确性条件的值，该值可以通过单个置信度值阈值获取。
- **X 倍正确高于**。 显示比整个数据集的准确性好 *X* 倍时相应的置信度值。 *X* 是在分析选项中为**改进准确性**指定的值。

之间的协议。 如果流中包括两个或两个以上对同一输出字段进行预测的已生成模型，那么您还将看到与模型所生成的预测之间的一**致性**相关的统计量。 这包括预测一致的记录数和百分比（对于分类输出字段）或误差汇总统计量的数目和百分比（对于连续输出字段）。 对于分类字段，它包括分析与模型一致（生成相同预测值）的记录子集的实际值相比的预测值。

评估度量。 对于二元分类器，如果您已请求分析选项中的评估度量，那么 AUC 和 Gini 系数评估度量的值将显示在表的此部分中。 对于每个二元分类器模型，表都包含与之对应的一行。 评估度量表针对每个输出字段而不是每个模型显示。

“数据审核”节点

通过“数据审核”节点，可对您放置到 IBM SPSS Modeler 中并且以易于读取的矩阵（可对该矩阵进行排序并使用它生成正常大小的图形和各种数据准备节点）形式显示的数据有个初步的全面了解。

- “审核”选项卡显示具有汇总统计量、直方图和分布图的报告，它们有助于获得对数据的初步了解。 该报告在字段名之前还显示存储图标。
- 审核报告中的“质量”选项卡显示有关离群值、极值和缺失值的信息，并提供用于处理这些值的工具。

使用“数据审核”节点

Data Audit 节点可直接附加到源节点，或附加到已实例化的 Type 节点的下游。 您也可以根据结果生成多个数据准备节点。 例如，可以生成过滤节点（该节点将具有过多缺失值的字段排除，不在建模中使用），并生成为任何或所有保留字段填补缺失值的超节点。 这就是审核的真正作用所在，使您不仅可以评估数据的当前状态，还可以根据评估执行操作。

筛选数据或数据采样。 初始审核在处理大数据时特别有效，因此可以在初始探索期间使用“样本”节点选择部分记录，以此缩短处理时间。 在分析的探索阶段，也可以将 Data Audit 节点与 Feature Selection 节点和 Anomaly Detection 节点等节点组合使用。

数据审核节点的设置选项卡

使用“设置”选项卡，可指定用于审核的基本参数。

缺省值。 您可以只是将节点附加到流中，然后单击**运行**以根据缺省设置生成所有字段的审核报告，如下所示：

- 如果没有 Type 节点设置，那么报告中将包括所有字段。
- 如果有“类型”设置（无论它们是否已实例化），则显示中包括所有输入、目标和双向字段。 如果有一个目标字段，请使用它作为“重叠”字段。 如果指定了多个目标字段，则不指定缺省重叠。

使用定制字段。 选择此选项可手动选择字段。 使用右侧的字段选择器按钮可单独选择字段或按类型选择字段。

重叠字段。 重叠字段用于绘制审核报告中显示的缩略图图形。 如果是连续（数值范围）字段，则还计算二元统计量（协方差和相关系数）。 如果单个目标字段根据“类型”节点设置显示，则使用它作为缺省重叠字段，如上所述。 或者，您可以选择**使用自定义字段**以指定重叠。

显示。 用于指定输出中是否显示图形以及选择缺省显示的统计量。

- **图形**。 显示每个选定字段的图形；根据数据的情况显示为分布（条形）图、直方图或散点图。 图形在初始报告中显示为缩略图，但也可生成标准大小的图形和图形节点。 有关更多信息，请参阅主题 第 240 页的『数据审核输出浏览器』。
- **基本/高级统计量**。 指定缺省显示在输出中的统计量的级别。 当此设置确定初始显示时，无论此设置是什么，所有统计量均显示在输出中。 有关更多信息，请参阅主题 第 240 页的『显示统计量』。

中位数和众数。 计算报告中所有字段的中位数和众数。 请注意，对于大型数据集，由于这些统计量的计算时间比其他统计量长，因此它们可能会延长处理时间。 在仅含中位数的情况下，报告值有时基于含 2000 条记录的样本（而不是完整数据集）。 为了防止超过内存限制，此取样是逐个字段地进行的。 当抽样生效时，输出中的结果将标记为（样本中位数而不只是中位数）。 对于除中位数之外的所有统计量，始终使用完整数据集进行计算。

空字段或无类型字段。 当无类型字段与实例化数据配合使用时，这些字段不会包括在审核报告中。 要包括无类型字段（包括空字段），请在位于上游的所有类型节点中选择**清除所有值**。 这可确保不实例化数据，从而导致在报告中包括所有字段。 例如，如果要获取所有字段的完整列表或生成将排除空字段的 Filter 节点，那么这非常有用。 有关更多信息，请参阅主题 第 242 页的『过滤含缺失数据的字段』。

数据审核的质量选项卡

Data Audit 节点中的“质量”选项卡提供用于处理缺失值、离群集和极值的选项。

缺失值

- **具有有效值的记录计数**。 选择此选项可为每个评估字段显示含有效值的记录数。 请注意，null（未定义的）值、空白值、空白和空字符串总是被视为无效值。
- **具有无效值的记录的细目计数**。 选择此选项可为每个字段显示含每类无效值的记录数。

离群值和极值

离群值和极值的检测方法。 支持两种方法：

平均值的标准差。 根据与平均值的标准差的个数检测离群值和极值。 例如，如果您具有一个包含平均值 100 和标准差 10 的字段，那么可以指定 3.0 来指出应将任何低于 70 或高于 130 的值视为离群值。

四分位距。 根据四分位距（即中间两个四分位数的间距，介于 25% 百分位数和 75% 百分位数之间）检测离群值和极值。 例如，根据缺省设置 1.5，离群值的阈值下限将为 Q1 - 1.5 * IQR，阈值上限将为 Q3 + 1.5*IQR。 请注意，使用此选项可能会降低大型数据集的性能。

数据审核输出浏览器

数据审核浏览器是用于获取数据概述的强大工具。 “审核”选项卡显示所有字段的缩略图、存储图标以及统计量，而“质量”选项卡显示有关离群值、极值和缺失值的信息。 根据初始图形和汇总统计量，您可以决定重新为数字字段编码、派生新字段，或重新为名义字段的值分类。 或者，您可能需要使用更加高级的可视化功能来进一步进行探索。 通过使用“生成”菜单创建任意数目的节点（这些节点可用于变换或显示数据），可以直接从审核报告浏览器执行此操作。

- 通过单击列标题为列排序，或使用拖放来重新为列排序。 并且支持大多数标准输出操作。 有关更多信息，请参阅主题 第 230 页的『查看输出』。
- 通过双击“测量”列或“唯一”列中的字段查看字段的值和范围。
- 使用工具栏或“编辑”菜单显示或隐藏值标签，或选择要显示的统计量。 有关更多信息，请参阅主题 第 240 页的『显示统计量』。
- 检验字段名左侧的存储图标。 存储格式描述数据在某个字段中的存储方式。 例如，值为 1 和 0 的字段存储整型数据。 这点与测量级别明显不同，测量级别描述的是数据的使用方法，而且不影响存储。 有关更多信息，请参阅主题 第 6 页的『设置字段存储类型和格式』。

查看和生成图形

如果未选择重叠，那么“审核”选项卡显示条形图（对于名义字段或标志字段）或直方图（连续字段）。

对于名义或标志字段重叠，将根据重叠的值为图形着色。

对于连续字段重叠，会生成二维散点图，而不是一维条形图和直方图。 在这种情况下，*x* 轴映射到重叠字段，从而使您可以在沿着表向下读取时，看到的所有 *x* 轴上相同尺度。

- 对于标志或名义字段，将鼠标指针停留在条形图上可在工具提示中显示基础值或标签。
- 对于标志或名义字段，使用工具栏可将缩略图的方向从水平切换为垂直。
- 要从任何缩略图生成标准大小的图形，请双击缩略图，或选择缩略图，然后从“生成”菜单中选择**图形输出**。 注意：如果缩略图基于抽样数据，那么在原始数据流仍然打开时，生成的图形将包含所有观测值。

  如果创建输出的“数据审核”节点已连接到流，则只能生成图形。
- 要生成匹配的图形节点，请在“审核”选项卡上选择一个或多个字段，然后从“生成”菜单中选择**图形节点**。 最终结点会被添加到流工作区中，并且可在每次运行流时用于重新创建图形。
- 如果重叠集具有 100 个以上的值，则会发出警告，并且不会包括重叠。

显示统计量

使用“显示统计量”对话框，可以选择显示在“审核”选项卡上的统计量。 初始设置是在 Data Audit 节点中指定的。 有关更多信息，请参阅主题 第 239 页的『数据审核节点的设置选项卡』。

Minimum. 数值变量的最小值。

Maximum. 数值变量的最大值。

Sum. 所有带有非缺失值的观测值的值的合计或总计。

范围. 数字变量的最大值与最小值的差值就是用最大值减最小值后得出的值。

平均值. 集中趋势的测量。 算术平均值，等于总和除以观测值数。

均值标准误差 *(Standard Error of Mean)*. 对取自相同分布的样本之间的平均值可能有多大差异的测量。 用于粗略将观测到的均数与假设值对比（即，如果差异与标准误差的比率小于 -2 或大于 +2，则可以得出此均数与假设值不同的结论）。

标准差. 围绕平均值的离差的测量，等于方差的平方根。 以和原始变量相同的单位度量标准差。

偏差. 对围绕平均值的离差的测量，值等于与平均值的差的平方和除以个案数减一。 方差按单元计量，即变量自身单元数的平方。

偏度 *(Skewness)*. 分布不对称性的测量。 正态分布是一种对称性分布，其偏度值为 0。 具有显著性正偏度的分布右侧尾部较长。 具有显著负偏态的分布具有向左延伸的长尾。 提示：取大于其标准误差两倍的偏度值指示离开对称的距离。

偏度标准误差 *(Standard Error of Skewness)*. 偏度与其标准误差的比率可用作正态性检验 (即，如果该比率小于 -2 或大于 + 2 ，那么可以拒绝正态性)。 偏度正值越大表示长尾向右越长；负极值表示向左的长尾。

峰度 *(Kurtosis)*. 存在离群值的程度的测量。 对于正态分布，峰度统计量的值为零。 正峰度值表示数据呈现比正态分布更极端的离群值。 负峰度值表示数据呈现比正态分布极端程度较低的离群值。

峰度标准误差 *(Standard Error of Kurtosis)*. 峰度与其标准误差的比率可用作正态性检验 (即，如果比率小于 -2 或大于 + 2 ，那么可以拒绝正态性)。 峰度较大的正值表示该分布的尾部比正态分布的尾部长；峰度的负值表示较短的尾部（与箱形均匀分布的尾部变得相似）。

UNIQUE. 同步评估所有效应，同时为任意类型的所有其他效应调整每一个效应。

有效。有效观测值既不包含系统缺失值，也不包含定义为用户缺失的值。 请注意，null（未定义的）值、空白值、空白和空字符串总是被视为无效值。

中位数. 大于或小于一半观测值的值，即 50th 个百分位。 如果有偶数个观测值，则中位数为它们以升序或降序排列时两个中间观测值的平均值。 中位数是集中趋势的一种测量，对离群值不敏感（与平均值不同，平均值会受部分极高或极低值的影响）。

众数 *(Mode)*. 最常出现的值。 如果多个值共享最大出现频率，则每个值都是一个众数。

请注意，为了提高性能，缺省情况下不会显示中位数和众数，但是您可以在 Data Audit 节点的“设置”选项卡上将其选中。 有关更多信息，请参阅主题 第 239 页的『数据审核节点的设置选项卡』。

重叠的统计量

如果连续（数值范围）重叠字段正在使用，则下列统计量也可用：

协方差 *(Covariance)*. 两个变量间关联性的非标准化测量值，等于叉积偏差除以 N-1。

数据审核浏览器的质量选项卡

“数据审核”浏览器中的“质量”选项卡显示数据质量分析的结果，并且可用于指定离群值、极值和缺失值的处理。

填补缺失值

审核报告列出每个字段完整记录的百分比以及有效值、空值和空白值的数目。 您可以根据情况选择填补特定字段的缺失值，然后生成超节点以应用这些变换。

1. 在**填补缺失值**列中，指定要填补的值的类型（如果有）。 您可以选择填补空白值和/或空值，或指定用于选择待填补值的定制条件或表达式。

   IBM SPSS Modeler 可识别的缺失值类型有以下几种：

   - **Null 或系统缺失值**。 这两种类型是数据库或源文件中留空、并且尚未在源节点或类型节点中专门定义为“缺失”的非字符串值。 系统缺失值会显示为 **$null$**。 请注意，空字符串在 IBM SPSS Modeler 中不会被视为 Null，但它们可能会被某些数据库视为 Null。
   - **空字符串和空白**。 空字符串值和空白（带有不可见字符的字符串）不被视为空值。 对于大多数用途，空字符串都视为相当于空白。 例如，如果您选择在源节点或类型节点中将空白视为空白值的选项，则此设置也应用于空字符串。
   - **空白或用户定义的缺失值**。 这些是在源节点或类型节点中被明确定义为缺失的值（例如 unknown、99 或 –1）。 您还可以将空和空白视为空白值，这样将使得它们被标记为进行特殊处理并排除在大多数计算之外。 例如，您可以使用 @BLANK 函数将这些值与其他类型的缺失值一起视为空白值。

2. 在**方法**列中，指定要使用的方法。

   下列方法可用于输入缺失值：

   已修正。 替换为固定值（可以字段平均值、范围中间值，或者您指定的常数）。

   随机。 替换为基于正态分布或均匀分布产生的随机值。

   表达式。 用于指定定制表达式。 例如，您可以使用设置全局量节点创建的全局变量替换值。

算法。 基于 C&RT 算法替换为模型预测的值。 对于使用此方法输入的每个字段，都会有一个单独的 C&RT 模型，还有一个填充节点会使用该模型预测的值替换空白值和空值。 然后使用过滤节点删除该模型生成的预测字段。

3. 要生成缺失值超节点，请从菜单中选择：

   生成 > 缺失值超节点

   这将显示“缺失值超节点”对话框。

4. 选择**所有字段**或**仅选定字段**，并根据需要指定样本大小。 （指定的样本是百分比，缺省情况，将对所有记录取 10% 的样本。）
5. 单击**确定**将生成的超节点添加到流工作区中。
6. 将超节点附加到流中以应用变换。

在超节点中，将根据情况使用由模型块、填充和过滤节点形成的组合。 要了解超节点如何工作，可以编辑超节点并单击**放大**，并且可以在超节点中添加、编辑或删除特定节点以对行为进行微调。

处理离群值和极值

对于每个字段，将根据在“数据审核”节点中指定的检测选项显示列有离群值和极值个数的审核报告。 有关更多信息，请参阅主题 第 239 页的『数据审核的质量选项卡』。 您可以根据情况选择控制、放弃特定字段的这些值或使其无效，然后生成超节点以应用这些变换。

1. 在**操作**列中，根据需要指定对特定字段的离群值和极值的处理。

   下列操作可用于处理离群值和极值：

   - **强制**。 将离群值和极值替换为不会被视为极值的最接近值。 例如，如果将离群值定义为高于或低于三个标准差的任何值，则会将所有离群值替换为此范围中的最高值或最低值。
   - **废弃**。 废弃含指定字段的离群值或极值的记录。
   - **使无效**。 将离群值和极值替换为空值或系统缺失值。
   - **强制离群值/废弃极值**。 只废弃极值。
   - **强制离群值/使极值无效**。 仅使极值无效。

2. 要生成超节点，请从菜单中选择：

   生成 > 离群值和极值超节点

   这将显示“离群值超节点”对话框。

3. 选择**所有字段**或**仅选定字段**，然后单击**确定**以将生成的超节点添加到流工作区中。
4. 将超节点附加到流中以应用变换。

（可选）您可以编辑超节点并进行放大以进行浏览或进行更改。 在超节点内，根据情况使用一系列“选择”和/或“填充”节点废弃、强制或使值无效。

过滤含缺失数据的字段

从数据审核浏览器中，可以根据质量分析的结果通过使用“质量”对话框中的“生成过滤器”创建新的“过滤”节点。

方式。 为指定的字段选择所需的操作（**包括**或**排除**）。

- **所选字段**。 “过滤”节点将包括/排除在“质量”选项卡上选择的字段。 例如，您可以根据**完成百分比**列为表格排序，再通过按住 Shift 键并单击来选择完成率最低的字段，然后生成排除这些字段的过滤节点。
- **质量百分比高于以下值的字段：** “过滤”节点将包括/排除完整记录的百分比高于指定阈值的字段。 缺省阈值为 50%。

过滤空字段或无类型字段

请注意，在将数据值实例化之后，审核结果或 IBM SPSS Modeler 中的大多数其他输出会将无类型字段或空字段排除。 这些字段在建模时会被忽略，但它们可能会使数据过多或混乱。 如果这样，可以使用数据审核浏览器生成“过滤”节点，以通过该“过滤”节点从流中删除这些字段。

1. 要确保所有字段都包含在审核中（包括空字段或无类型字段），请单击上游源或“类型”节点中的**清除所有值**，或者将所有字段的值设置为 *<Pass>*。
2. 在数据审核浏览器中，请根据**完成百分比**列进行排序，选择含零个有效值（或某个其他阈值）的字段，并使用“生成”菜单生成可添加到流中的过滤节点。

选择含缺失数据的记录

从数据审核浏览器中，可以根据质量分析的结果创建新选择节点。

1. 在“数据审核”浏览器中，选择“质量”选项卡。
2. 从菜单中，选择：

   生成 > 缺失值选择节点

   这将显示“生成选择节点”对话框。

当记录处于以下状态时选择： 指定当记录**有效**或**无效**时是否应保存这些记录。

在以下位置查找无效值： 指定在何处检查无效值。

- **所有字段。** 选择字段将检查所有字段中是否有无效值。
- **在表中选择的字段。** 选择节点将只检查当前在“质量”输出表中选择的字段。
- **质量百分比高于以下值的字段：** 选择节点将检查完整记录的百分比大于指定阈值的字段。 缺省阈值为 50%。

如果在以下位置找到无效值，那么认为记录无效。 指定将记录确定为无效的条件。

- **以上字段中的任何一个。** 如果上述任何 指定字段包含某记录的无效值，则选择节点会将该记录视为无效。
- **以上的所有字段。** 如果上述所有 指定字段都包含某记录的无效值，则选择节点会将该记录视为无效。

生成其他用于数据准备的节点

在数据准备中使用的各种节点可直接从数据审核浏览器中生成，包括 Reclassify、Binning 和 Derive 节点。例如：

- 您可以根据 *claimvalue* 和 *farmincome* 的值派生新字段，方法为：在审核报告中选择这两者，并从“生成”菜单中选择**派生**。 这会将新节点添加到流画布中。
- 同样，您可以根据审核结果确定：将 *farmincome* 重新编码为基于百分位数的箱以提供更加集中的分析。要生成分级节点，请在显示中选择字段行，然后从“生成”菜单中选择**分级**。

一旦生成节点并将其添加到流画布中，必须将它附加到流中并打开该节点以指定所选字段的选项。

变换节点

将输入字段正态化是使用传统评分技术（如回归、logistic 回归和判别分析）之前的一个重要步骤。 这些技术采用的数据服从正态分布的假设，对于许多原始数据文件可能不适用。 处理现实世界数据的一种方法是：对原始数据元素作变换，使其更接近正态分布。 此外，可以轻松地在正态化字段之间进行比较例如，收入和年龄在原始数据文件中有着完全不同的尺度，但正态化后，可以轻松地解释每个尺度的相对影响。

“变换”节点提供输出查看器，使用该输出查看器，可以快速而直观地评估要使用的最佳变换。 您可以快速查看变量是否是正态分布，并在需要时选择所需的变换并进行应用。 可以选择多个字段并针对每个字段执行一次变换。

为字段选择首选变换后，可以生成执行这些变换的“派生”节点或“过滤”节点，并将这些节点附加到流中。“派生”节点创建新字段，而“过滤”节点变换现有字段。 有关更多信息，请参阅主题 第 245 页的『生成图形』。

“变换”节点的“字段”选项卡

在“字段”选项卡上，可指定要使用数据的哪些字段并查看可能的变换并应用它们。 仅能变换数字字段。 单击字段选择器按钮，并从显示的列表中选择一个或多个数字字段。

变换节点的选项选项卡

使用“选项”选项卡，可以指定要包括的变换类型。 您可以选择包括所有可用变换，或单独选择各个变换。

在后一种情况下，也可以输入一个数字以偏移逆变换和对数变换的数据。 如果数据中有很大比例的零，那么可能会导致平均值和标准差结果有偏差，此时作偏移处理将会非常有用。

例如，假设您有一个名为余额的字段，该字段中包含有一些零，并且您要针对该字段使用逆变换。 为避免不期望的偏差，请选择 **Inverse (1/x)** 并在**使用数据偏移**字段中输入 1。 （请注意，此偏移量与 IBM SPSS Modeler 中的 @OFFSET 序列函数所执行的偏移量无关。）

所有公式。 标识出所有应计算的可用变换并将其显示在输出中。

选择公式。 用于选择要计算并显示在输出中的不同变换。

- **反函数 (1/x)**。 指出逆变换应显示在输出中。
- **对数 (log n)**。 指出 $\log_n$ 变换应显示在输出中。
- **对数 (log 10)**。 指出 $\log_{10}$ 变换应显示在输出中。
- **指数**。 指出指数变换 (e^x) 应显示在输出中。
- **平方根**。 指出平方根变换应显示在输出中。

变换节点的输出选项卡

使用“输出”选项卡，可以指定输出格式和输出位置。 您还可以选择将结果显示在屏幕上，或将它们发送到其中一种标准文件中。 有关更多信息，请参阅主题 第 233 页的『输出节点的“输出”选项卡』。

变换节点的输出查看器

使用输出查看器，可以查看“变换”节点的执行结果。 该查看器是一种功能强大的工具，它在变换的缩略图视图中显示每个字段的多个变换，从而使您可以快速地比较字段。 您可以使用其“文件”菜单上的选项来保存、导出或打印输出。 有关更多信息，请参阅主题 第 230 页的『查看输出』。

对于所选变换以外的每个变换，会以下面的格式在其下显示图例：

```
Mean (Standard deviation)
```

为变换生成节点

输出查看器为数据准备提供了有用的起始点。 例如，您可能想要将字段年龄正态化，以便可以使用采用正态分布的评分技术（如 logistic 回归或判别分析）。 依据初始图形和汇总统计量，您可能会决定根据特定分布（例如，对数分布）变换年龄字段。 选择首选分布后，可以生成使用标准化变换的导出节点以用于评分。

可以从输出查看器中生成下列字段操作节点：

- 导出
- 填充

“派生”节点创建含所需变换的新字段，而“填充”节点变换现有字段。 节点以超节点的形式放置在画布上。

如果为不同的字段选择同一变换，那么“派生”节点或“填充”节点为应用该变换的所有字段包含该变换类型的公式。 例如，假设您选择了下表中显示的字段和变换来生成“派生”节点。

表 *43: “派生”*节点生成示例

字段	**转换**
AGE	当前分布
INCOME	日志
OPEN_BAL	逆模型
BALANCE	逆模型

超节点中包含下列节点：

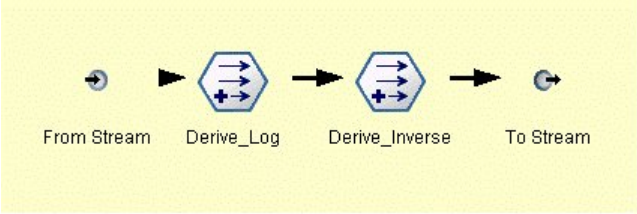


图 *74:* 工作区上的超节点

在此示例中，Derive_Log 节点具有收入字段的对数公式，而 Derive_Inverse 节点具有 *OPEN_BAL* 和余额字段的逆公式。

生成节点

1. 对于输出查看器中的每个字段，请选择所需的变换。
2. 从“生成”菜单中，根据需要选择**导出节点**或**填充节点**。

   如果这样做，会相应地显示“生成导出节点”或“生成填充节点”对话框。

   根据需要选择**非标准化变换**或**标准化变换（z 分）**。 第二个选项将 *z* 分应用到变换；*z* 分将值表示为与标准差中变量平均值的差值的函数。 例如，如果将对数变换应用到年龄字段并选择标准化变换，则生成的节点的最终公式为：

   ```
   (log(AGE)-Mean)/SD
   ```

一旦节点生成并显示在流工作区上：

1. 将它附加到流中。
2. 对于超节点，可以选择双击该节点以查看它的内容。
3. （可选）双击导出节点或填充节点以修改所选字段的选项。

生成图形

您可以在输出查看器中根据缩略直方图生成标准大小的直方图输出。

生成图形

1. 在输出查看器中双击缩略图。

   或者

   在输出查看器中选择缩略图。
2. 从“生成”菜单中，选择**图形输出**。

执行此操作将显示该直方图，并在其上叠放一条正态分布曲线。 这样，您便可以比较每个可用变换与正态分布的匹配程度。

注意：仅当创建输出的“变换”节点连接到流时，才能够生成图形。

其他操作

从输出查看器中，还可以：

- 按“字段”列为输出网格排序。
- 将输出导出到 HTML 文件中。 有关更多信息，请参阅主题 第 232 页的『导出输出』。

统计量节点

“统计量”节点提供与数字字段相关的基本汇总信息。 您可以获取各个字段的汇总统计量以及字段之间的相关。

统计量节点的设置选项卡

检查。 选择您需要其单独汇总统计量的字段。 您可以选择多个字段。

统计。 选择要报告的统计量。 可用选项包括**计数**、**平均值**、**和**、**最小值**、**最大值**、**极差**、**方差**、**标准差**、**平均值的标准误差**、**中位数**和**众数**。

相关。 选择您希望相关的字段。 您可以选择多个字段。 当选择相关字段时，将在输出中列出每个“检查”字段和相关字段之间的相关。

相关设置。 您可以指定用于在输出中显示相关强度的选项。

相关设置

IBM SPSS Modeler 可以使用描述性标签描述相关的特征以帮助突出显示重要关系。 **相关**度量两个连续（数值范围）字段之间的关系强度。 它的值介于 –1.0 和 1.0 之间。 值接近于 +1.0 表示强正相关，因此在两个字段之间，大值与大值相关，小值与小值相关。 值接近于 –1.0 表示强负相关，因此在两个字段之间，大值与小值相关，小值与大值相关。 值接近于 0.0 表示弱相关，因此，两个字段的值或多或少地相互独立。

通过使用“相关设置”对话框，您可以控制相关标签的显示，更改定义类别的阈值，以及更改用于每个范围的标签。 因为刻画相关的方式很大程度上依赖于问题域，所以您可能需要依据具体情况来自定义范围和标签。

在输出中显示相关强度标签。 缺省情况下，此选项处于选中状态。 取消选择此选项将在输出中省略描述标签。

相关强度。 有两个选项用于定义和标记相关强度：

- **按重要性 (1-p) 定义相关强度**。 根据重要性标记相关，重要性等于 1 减显著性（即，1 减去平均值的差值完全归结于机遇变异的概率）。 此值越接近于 1，两个字段不独立的机率越大，换句话说，它们之间存在某种关系。 一般情况下，建议根据重要性而不是绝对值标记相关，因为重要性考虑了数据的可变性，例如系数 0.6 可能在某个数据集中非常显著，而在另一个数据集中根本不显著。 缺省情况下，将介于 0.0 和 0.9 之间的重要性值标记为弱，将介于 0.9 和 0.95 之间的重要性值标记为中，将介于 0.95 和 1.0 之间的重要性值标记为强。
- **按绝对值定义相关强度**。 如上所述，根据 Pearson 相关系数（介于 –1 和 1 之间）的绝对值标记相关。 此度量的绝对值越接近于 1，相关就越强。 缺省情况下，将介于 0.0 和 0.3333 之间的相关（采用绝对值的形式）标记为弱，将介于 0.3333 和 0.6666 之间的相关标记为中，将介于 0.6666 和 1.0 之间的相关标记为强。 但是请注意，要将任何给定值的显著性从一个数据集扩展到另一个数据集都是非常困难的；因此，在大多数情况下，建议根据概率而不是绝对值定义相关。

统计量输出浏览器

统计量节点输出浏览器显示统计量分析的结果，并且可用于执行操作，包括选择字段、根据选择生成新节点，以及保存和打印结果。 “文件”菜单中提供了常用的保存、导出和打印选项，“编辑”菜单中提供了常用的编辑选项。 有关更多信息，请参阅主题 第 230 页的『查看输出』。

首次浏览统计量输出时，会展开结果。 要在查看结果后将其隐藏，请使用项目左侧的扩展器控件将要隐藏的特定结果折叠，或单击**全部折叠**按钮以折叠所有结果。 要在折叠结果后再次对其进行查看，请使用项目左侧的展开器控件显示结果，或单击**全部展开**按钮以显示所有结果。

输出包含每个检查字段的一部分，还包含所请求的统计量的表格。

- **计数**。 具有字段的有效值的记录数。
- **平均值**。 对所有记录的该字段值求平均。
- **总和** 对所有记录的该字段值求和。
- **最短** 字段的最小值。
- **最大值**。 字段的最大值。
- **范围**。 最大值与最小值之间的差。
- **方差**。 一种对字段值可变性的度量。 计算方法是：求出每个值和总均值之间的差值并平方之，然后对所有的平方值求和，再除以记录数。
- **标准差**。 字段值的可变性的另一个度量，其值为方差的平方根。

- **平均值的标准误差**。 一种对字段均值估计值的不确定性（如果该均值应用到新数据）的度量。
- **中位数**。 字段的“中间”值；即，根据字段值将数据的上半部分与下半部分拆分开的值。
- **方式**。 数据中最常见的单个值。

相关性。 如果指定了相关字段，那么输出还包含列出“检查”字段和每个相关字段之间的 Pearson 相关的部分，以及相关值的可选描述标签。 有关更多信息，请参阅主题 第 246 页的『相关设置』。

生成菜单。 “生成”菜单包含节点生成操作。

- **过滤**。 生成“过滤”节点以过滤掉与其他字段无关或相关弱的字段。

根据统计量生成过滤节点

从统计量输出浏览器生成的“过滤”节点将根据它们与其他字段的相关性过滤字段。 它的工作方式为：按绝对值的顺序为相关排序，获取一些最大的相关（根据“统计量”对话框中的条件集），并创建过滤器（该过滤器遍历在这些大相关中显示的所有字段）。

方式。 确定如何选择相关。 **包括**导致保留在指定的相关中显示的字段。 **排除**导致过滤掉字段。

包含/排除其中出现的字段。 定义用于选择相关的条件。

- **最大相关数**。 选择指定数量的相关并包括/排除在这些相关中出现的所有字段。
- **最大相关百分比 (%)**。 选择指定的相关百分比 (*n*%)，并包含/排除在其中任何相关中出现的字段。
- **相关大于**。 选择绝对值大于指定阈值的相关。

“平均值”节点

“平均值”节点在独立组之间或相关字段对之间进行平均值比较，以检验是否存在显著差别。 例如，您可以将开展促销前后的收入平均值进行比较，或将从参加促销的客户那获得的收入与从未参加促销的客户那获得的收入进行比较。

您可以根据您的数据以两种不同的方式比较平均值：

- **字段中的不同组之间**。 要比较独立组，请选择一个检验字段和一个分组字段。 例如，您可在开展促销时排除“拒不参加”客户样本，并将“拒不参加”组的收入平均值与所有其他组的收入平均值进行比较。 在这种情况下，您要指定一个检验字段（该字段指出每名客户的收入），以及一个标志或名义字段（该字段指出他们是否获得了优惠）。 这些样本是独立的，原因是：将每条记录分配给一个组或另一个组，并且无法将一个组的特定成员链接到另一个组的特定成员。 也可以指定含两个以上值的名义字段以比较多个组的平均值。 当执行该节点时，它会针对所选字段进行单因素 ANOVA 检验。 如果只有两个字段组，则单因素 ANOVA 结果与独立样本的 *t* 检验本质上一样。 有关更多信息，请参阅主题 第 247 页的『比较独立组的平均值』。
- **在字段对之间**。 比较两个相关字段的平均值时，组必须以某种方式配对，结果才有意义。 例如，可将同一组客户在开展促销前后的收入平均值进行比较，或在夫妻对之间比较某服务的使用率，以查看它们是否不同。 每条记录包含两个单独但相关的测量量，可以对它们进行有意义的比较。 当执行该节点时，它针对所选的每个字段对进行成对样本 *t* 检验。 有关更多信息，请参阅主题 第 247 页的『在成对字段之间比较平均值』。

比较独立组的平均值

在“平均值”节点中选择**在字段中的组之间**以比较两个或更多个独立组的平均值。

分组字段。 选择一个数字标志或名义字段，该字段含两个或两个以上的不同值，并且将记录分为要比较的组，如获得优惠的组和未获得优惠的组。 无论检验字段数是多少，都只能选择一个分组字段。

检验字段。 选择一个或多个包含要检验的测量量的数字字段。 对于您选择的每个字段，将进行单独检验。 例如，您可以检验给定促销对使用情况、收入和买卖的影响。

在成对字段之间比较平均值

在“平均值”节点中选择**在字段对之间**以在单独字段之间比较平均值。 这些字段必须以某种方式相关，结果才有意义，如促销前后的收入。 也可以选择多个字段对。

字段一。 选择包含要比较的第一个测量量的数字字段。 在前后研究中，该字段将为“之前”字段。

字段二。 选择要比较的第二个字段。

添加。 将所选对添加到“检验”字段对的列表中。

根据需要重复进行字段选择以将多个对添加到该列表中。

相关设置。 使您可以指定用于标注相关性强度的选项。 有关更多信息，请参阅主题 第 246 页的『相关设置』。

“平均值”节点选项

使用“选项”选项卡可以设置用于将结果标注为重要、边际或不重要的阈值 *p*。 您也可以编辑每个排秩的标签。 重要性可依据百分比尺度进行度量，并且可将重要性定义推广为 1 减去获取完全由机遇变异造成的相同或更为极端的结果（如两个字段的平均值差值）的概率。 例如，*p* 值大于 0.95 表示结果完全归结于机遇变异的机率小于 5%。

重要性标签。 您可以编辑用于标记输出中的每个字段对或组的标签。 缺省标签为重要、边际和不重要。

分界值。 指定每个等级的阈值。 通常情况下，大于 0.95 的 *p* 值将归为重要等级，而小于 0.9 则归为不重要等级，但可以根据需要调整这些阈值。

注意：许多节点中都提供了重要性度量。 具体计算依赖于节点以及所使用的目标和输入字段的类型，但仍旧可以对值进行比较（因为所有值都是根据百分比尺度测量的）。

“平均值”节点输出浏览器

平均值输出浏览器以交叉列表的形式显示数据，并且可用于执行标准操作，如一次一行地选择和复制表格，按任何列进行排序，以及保存和打印表格。 有关更多信息，请参阅主题 第 230 页的『查看输出』。

表格中的具体信息依赖于比较的类型（字段中的组或单独的字段）。

排序依据。 使您可以根据特定列对输出排序。 单击向上或向下箭头可更改排序的方向。 或者，可以单击任何列标题以便根据该列进行排序。 （要更改列中的排序方向，请再次单击。）

视图。 您可以选择**简单**或**高级**以控制显示中的详细信息级别。 高级视图包括简单视图中的所有信息，但是还额外提供了详细信息。

比较字段中的组的平均值输出

比较字段中的组时，分组字段的名称显示在输出表的上方，并且单独为每个组报告平均值和相关统计量。 该表格为每个检验字段包括一个单独的行。

以下各列将会显示：

- **字段**。 列出所选检验字段的名称。
- **平均值（按组）**。 显示分组字段的每个类别的平均值。 例如，可以将获得特殊优惠的客户（新促销）与未获得特殊优惠的客户（标准）进行比较。 在高级视图中，还会显示标准差、标准误差和计数。
- **重要性**。 显示重要性值和标签。 有关更多信息，请参阅主题 第 248 页的『“平均值”节点选项』。

高级输出

在高级视图中，还会显示以下各列。

- **F 检验**。 此检验基于组之间的方差与每个组内的方差的比率。 如果所有组的平均值相同，则您将预计 *F* 比率接近于 1（因为两者均是同一总体方差的估计值）。 此比率越大，组间的方差就越大，并且存在显著差别的机率越大。
- **自由度**。 显示自由度。

比较字段对的平均值输出

比较单独字段时，输出表为每个所选字段对包括一行。

- **字段 1/2**。 显示每个对中第一个字段和第二个字段的名称。 在高级视图中，还会显示标准差、标准误差和计数。
- **平均值 1/2**。 分别显示每个字段的平均值。
- **相关性**。 度量两个连续（数值范围）字段之间的关系强度。 值接近于 +1.0 表示强正相关，值接近于 –1.0 表示强负相关。 有关更多信息，请参阅主题 第 246 页的『相关设置』。
- **平均值差值**。 显示两个字段平均值之间的差值。
- **重要性**。 显示重要性值和标签。 有关更多信息，请参阅主题 第 248 页的『“平均值”节点选项』。

高级输出

高级输出增加了以下各列：

95% 置信区间。 范围的下限和上限，对总体中所有具有该样本量的区间而言，实际均值落入其中的可能性为 95%。

T 检验。 通过将平均值差值除以它的标准误差，可以获取 *t* 统计量。 此统计量的绝对值越大，平均值不相同的概率越大。

自由度。 显示统计量的自由度。

报告节点

使用“报告”节点，可以创建包含固定文本以及数据和从该数据派生的其他表达式的格式报告。 通过使用文本模板定义固定文本和数据输出构造，可以指定报告的格式。 通过使用模板中的 HTML 标记和在“输出”选项卡上设置选项，可以提供定制文本格式。 在使用模板中的 CLEM 表达式的报告中包含有数据值和其他条件输出。

报告节点的替代选项

报告节点最常用于列出流的记录或案例输出，如满足某个特定条件的所有记录。 就此而言，可将“报告”节点视为“表格”节点的结构性较差的替代选项。

- 如果您希望报告列出字段信息或在流而不是数据本身中定义的任何其他内容（如在“类型”节点中指定的字段定义），那么可以改用脚本。
- 要生成包括多个输出对象（如一个或多个流生成的模型、表格和图形的集合）并且可以成为采用多种格式（包括文本、HTML 和 Microsoft Word/Office）的输出的报告，可以使用 IBM SPSS Modeler 项目。
- 要在未使用脚本的情况下生成字段名称列表，可以使用前面带有“样本”节点（废弃所有记录）的“表格”节点。 这会生成一个不含行的表格，该表格可在导出时转置以在单一列中生成字段名称列表。 （要这样做，请在表节点中的“输出”选项卡上选择**转置数据**。）

报告节点的模板选项卡

创建模板。 要定义报告的内容，请在“报告”节点的“模板”选项卡上创建模板。 该模板包含数行文本，每一行都指定与报告内容相关的某些信息，并且用一些特殊标记行指出内容行的范围。 在每个内容行中，会在将该行发送到报告之前对括在方括号 ([]) 内的 CLEM 表达式求值。 模板中某个行的可能范围有三个：

已修正。 未标记的行被视为固定行。 在对固定行包含的所有表达式求值后，只将这些行向报告复制一次。 例如，行

```
This is my report, printed on [@TODAY]
```

将一个行复制到报告中，包含文本和当前日期。

全局 (iterate ALL)。 对于输入数据的每条记录，会将包含在特殊标记 #ALL 与 # 之间的行向报告复制一次。 CLEM 表达式（括在方括号中）根据每个输出行的当前记录进行求值。 例如，行

```
#ALL
For record [@INDEX], the value of AGE is [AGE]
#
```

将为每个记录包括一行，指出记录号和年龄。

生成所有记录的列表：

```
#ALL
[Age]    [Sex]    [Cholesterol]     [BP]
#
```

条件 (iterate WHERE)。 对于指定的条件为 true 的每条记录，会将特殊标记 #WHERE <*condition*> 和 # 之间包含的行复制到报告一次。 该条件是指 CLEM 表达式。 （在 WHERE 条件中，方括号是可选的。） 例如，行

```
#WHERE [SEX = 'M']
Male at record no. [@INDEX] has age [AGE].
#
```

会为每个性别值为 *M* 的记录向文件写入一行。 完整的报告将包含通过将模板应用到输入数据定义的固定行、全局行和条件行。

您可以使用各种类型的输出节点都具备的“输出”选项卡指定用于显示或保存结果的选项。 有关更多信息，请参阅主题 第 233 页的『输出节点的“输出”选项卡』。

以 HTML 或 XML 格式输出数据

您可以直接在模板中包括 HTML 或 XML 标记以使用这两种格式中的任意一种编写报告。 例如，以下模板生成 HTML 表。

```
This report is written in HTML.
Only records where Age is above 60 are included.

<HTML>
  <TABLE border="2">
      <TR>
        <TD>Age</TD>
        <TD>BP</TD>
        <TD>Cholesterol</TD>
        <TD>Drug</TD>
      </TR>

  #WHERE Age > 60
    <TR>
          <TD>[Age]</TD>
          <TD>[BP]</TD>
          <TD>[Cholesterol]</TD>
          <TD>[Drug]</TD>
        </TR>
  #
  </TABLE>
</HTML>
```

报告节点输出浏览器

报告浏览器向您显示所生成的报告的内容。 “文件”菜单中提供了常用的保存、导出和打印选项，“编辑”菜单中提供了常用的编辑选项。 有关更多信息，请参阅主题 第 230 页的『查看输出』。

设置全局量节点

设置全局量节点扫描数据并计算可在 CLEM 表达式中使用的汇总值。 例如，可以使用“设置全局量”节点为名为 *age* 的字段计算统计，然后通过插入 @GLOBAL_MEAN(age) 函数在 CLEM 表达式中使用 *age* 的总均值。

设置全局量节点的设置选项卡

要创建的全局量。 选择希望其全局量可用的字段。 您可以选择多个字段。 对于每个字段，请通过确保在字段名称旁边的列中选中所需的统计量来指定要计算的统计量。

- **MEAN**。 对所有记录的该字段值求平均。
- **SUM**。 对所有记录的该字段值求和。
- **最短** 字段的最小值。
- **MAX**。 字段的最大值。
- **SDEV**。 标准差，它是字段值的可变性的度量，其值为方差的平方根。

缺省操作。 在此处选择的选项将在向上面的全局列表添加新字段时使用。 要更改缺省统计量集，请根据情况选择或取消选择统计量。 也可以使用**应用**按钮将缺省运算应用到列表中的所有字段。

注: 某些操作不适用于非数字字段（例如，“合计”不适用于日期/时间字段）。 无法用于选定字段的操作将处于禁用状态。

在执行之前清除所有全局量。 选择此选项可在计算新值前删除所有全局量。 如果不选择此选项，则新计算出的值会替换较旧的值，但未重新计算的全局量仍保持可用。

显示执行后创建的全局量预览。 如果选择此选项，那么“流属性”对话框的“全局量”选项卡将在执行后显示，以显示计算出的全局量。

“模拟拟合”节点

“模拟拟合”节点将一组候选统计分布拟合到数据中的每个字段。 每个分布到字段的拟合将通过拟合度标准进行评估。 执行“模拟拟合”节点时，将构建一个“模拟生成”节点（或更新现有节点）。 将为每个字段分配其最佳拟合分布。 然后，可以使用“模拟生成”节点为每个字段生成模拟数据。

虽然“模拟拟合”节点是一个终端节点，但它不会向已生成的模型选用板添加模型，也不会向输出选项卡添加输出或图表或者导出数据。

注: 如果历史数据较为稀疏（即，缺失值非常多），那么拟合组件可能难以找到足够多的有效值将分布拟合到数据。 对于数据较为稀疏的情况，您应该先移除不需要的稀疏字段或插补缺失值，然后再进行拟合。 通过使用“数据审核”节点的**质量**选项卡上的选项，您可以查看完整记录的数目、标识稀疏字段并选择插补方法。 如果用于分布拟合的记录数不足，那么可以使用“平衡”节点来增加记录数。

使用“模拟拟合”节点可自动创建“模拟生成”节点

首次执行“模拟拟合”节点时，将使用指向“模拟拟合”节点的更新链接创建一个“模拟生成”节点。 再次执行“模拟拟合”节点时，只有在已移除更新链接的情况下才会创建新的“模拟生成”节点。另外，“模拟拟合”节点还可用于更新已连接的“模拟生成”节点。 结果取决于是否在这两个节点中存在相同的字段，以及是否在“模拟生成”节点中解锁了这些字段。 有关更多信息，请参阅主题 第 41 页的『“模拟生成”节点』。

“模拟拟合”节点只能具有一个指向“模拟生成”节点的更新链接。 要定义指向“模拟生成”节点的更新链接，请完成下列步骤：

1. 右键单击“模拟拟合”节点。
2. 从菜单中选择**定义更新链接**。
3. 单击要定义的更新链接所指向的“模拟生成”节点。

要移除“模拟拟合”节点与“模拟生成”节点之间的更新链接，请右键单击该更新链接，然后选择**移除链接**。

分布拟合

统计分布是某个变量可以使用的值的理论出现频率。 在“模拟拟合”节点中，会将一组理论统计分布与每个数据字段进行比较。 主题第 48 页的『分布』中描述了可用于拟合的分布。 根据拟合优度的度量（Anderson-Darling 标准或 Kolmogorov-Smirnov 标准）调整理论分布的参数，为数据提供最佳拟合。 通过“模拟拟合”节点实现的分布拟合的结果显示拟合了哪些分布、每个分布的最佳参数估计以及每个分布与数据的拟合度。 分布拟合期间，还可以计算具有数字存储类型的字段之间的相关性，以及具有分类分布的字段之间的偶然性。 分布拟合的结果将用于创建“模拟生成”节点。

将任何分布与数据进行拟合之前，会在前 1000 条记录中查找缺失值。 如果缺失值过多，那么无法进行分布拟合。 在这种情况下，您必须确定以下某个选项是否适用：

- 使用上游节点移除包含缺失值的记录。

- 使用上游节点针对缺失值对值进行插补。

分布拟合未排除用户缺失值。 如果您的数据包含用户缺失值，并且您希望从分布拟合中排除这些值，那么应该将这些值设置为系统缺失值。

拟合分布时，将不会考虑字段的角色。 例如，角色为**目标**的字段的处理方式与角色为**输入**、**无**、**两者**、**分区**、**分割**、**频率**和**标识**的字段相同。

分布拟合期间，将根据字段的存储类型和测量级别以不同方式对这些字段进行处理。 下表描述了分布拟合期间的字段处理。

表 *44:* 根据字段的存储类型和测量级别进行的分布拟合

存储器类型	**测量级别**					
	连续	分类	标志	名义	有序	无类型
String	不可能	对分类分布、散子分布和固定分布进行拟合				将忽略字段，并且不会将字段传递到“模拟生成”节点。
整数	对所有分布进行拟合。 将计算相关性和偶然性。	对分类分布进行拟合。 不计算相关性。			对二项式分布、负二项式分布和泊松分布进行拟合，并计算相关性。	
实数						
时间						
日期						
时间戳记						
未知	根据数据确定相应的存储类型。					

对于测量级别为有序的字段，其处理方式类似于连续字段，并且它们包含在“模拟生成”节点中的相关表内。 如果您要将二项式分布、负二项式分布或泊松分布以外的分布拟合到有序字段，那么必须将字段的测量级别更改为连续。 如果您先前为有序字段的每个值定义了标签，并且随后将测量级别更改为连续，那么这些标签将丢失。

在分布拟合到具有多个值的字段时，将以相同方式处理具有单个值的字段。 具有存储类型时间、日期或时间戳记的字段将作为数字进行处理。

将分布拟合到分割字段

如果您的数据包含分割字段，并且您希望对每个分割单独执行分布拟合，那么必须使用上游“重构”节点来变换数据。 使用“重构”节点可以为分割字段的每个值生成一个新字段。 随后，可以将此重构数据用于“模拟拟合”节点中的分布拟合。

“模拟拟合”节点的“设置”选项卡

源节点名称。 通过选择**自动**，您可以自动生成已生成（或已更新）的“模拟生成”节点的名称。 如果指定了定制名称，那么自动生成的名称将为“模拟拟合”节点中指定的名称（或者，如果“模拟拟合”节点中未指定任何定制名称，那么将为 Sim Gen）。 选择**定制**可在相邻的文本字段中指定定制名称。 除非对此文本字段进行编辑，否则缺省定制名称为 Sim Gen。

拟合选项 通过这些选项，您可以指定分布拟合到字段的方式以及评估分布拟合的方式。

- **要采样的个案数**。 此选项指定将分布拟合到数据集中的字段时要使用的观测值数。 选择**所有观测值**可将分布拟合到数据中的所有记录。 数据集非常大时，您可能需要考虑限制用于分布拟合的观测值数。 选择**限制**

为前 N 个观测值可仅使用前 N 个观测值。 单击箭头可指定要使用的观测值数。 或者，可以使用上游节点对用于分布拟合的记录进行随机抽样。

- **拟合优度条件（仅限连续字段）**。 对于连续字段，在将分布拟合到字段时，选择拟合优度的 Anderson-Darling 检验或 Kolmogorov-Smirnoff 检验对分布进行排序。 缺省情况下会选中 Anderson-Darling 检验，当您想要确保在尾部区域中尽可能实现最佳拟合时，尤其推荐使用此检验方式。 对于每个候选分布，计算了这两种统计信息，但仅使用所选统计信息对分布进行排序并确定最佳拟合分布。
- **分箱（仅限经验分布）**。 对于连续字段，经验分布是历史数据的累积分布函数。 它是每个值或值范围的概率，并且直接派生自数据。 通过单击箭头，可以指定用于计算连续字段的经验分布的分级数目。 缺省值为 100，最大值为 1000。
- **权重字段（可选）**。 如果数据集包含权重字段，请单击字段选择器图标并从列表中选择权重字段。 然后，将从分布拟合过程中排除权重字段。 列表显示了数据集中测量级别为连续的所有字段。 只能选择一个权重字段。

“模拟评估”节点

“模拟评估”节点是一个终端节点，用于评估指定字段、提供该字段的分布以及生成分布图和相关图。 此节点主要用于评估连续字段。 因此，它将对评估图进行补充，该图由“评估”节点生成并且可用于评估独立字段。另一差异在于，“模拟评估”节点跨多个迭代对单一预测进行评估，而“评估”节点对多个预测进行评估，其中每个预测包含单个迭代。 在为“模拟生成”节点中的分布参数指定了多个值的情况下，将生成迭代。 有关更多信息，请参阅主题 第 48 页的『迭代』。

“模拟评估”节点设计为与从“模拟拟合”和“模拟生成”节点中获取的数据配合使用。 但是，此节点可以与任何其他节点配合使用。 可以在“模拟生成”节点和“模拟评估”节点之间放置任意数目的处理步骤。

要点: “模拟评估”节点至少需要 1000 条具有目标字段的有效值的记录。

“模拟评估”节点的“设置”选项卡

在“模拟评估”节点的“设置”选项卡上，您可以指定数据集中每个字段的角色，并定制模拟所生成的输出。

选择一个项目。 使您能够在“模拟评估”节点的以下三个视图之间进行切换：字段、密度函数和输出。

“字段”视图

目标字段。 这是是必填字段。 单击箭头可从下拉列表中选择数据集的目标字段。 所选字段可以具有连续、有序或名义测量级别，但不能具有日期或未指定的测量级别。

迭代字段（可选）。 如果您的数据具有指示数据中每条记录所属迭代的迭代字段，那么必须在此处选择该字段。 这表示将分别对每次迭代进行评估。 只能选择具有连续、有序或名义测量级别的字段。

输入数据已按迭代排序。 只有在**迭代字段（可选）**字段中指定了迭代字段的情况下，才会启用此选项。 只有在您确定输入数据已按**迭代字段（可选）**中指定的迭代字段进行排序的情况下，才能选择此选项。

要绘制的最大迭代次数。 只有在**迭代字段（可选）**字段中指定了迭代字段的情况下，才会启用此选项。 单击箭头可指定要绘制的迭代的数目。 指定此数目可避免尝试在单个图表上绘制过多的迭代，迭代过多会导致难以解释绘制的图形。 最大迭代数可以设置的最低级别为 2；最高级别为 50。 要绘制的最大迭代数最初设置为 10。

相关性龙卷风的输入字段。 相关性龙卷风图是一个条形图，它显示了指定目标与每个指定输入之间的相关系数。 单击字段选择器图标可从可用模拟输入的列表中选择要包括在龙卷风图中的输入字段。 只能选择具有连续和有序测量级别的输入字段。 列表中未提供名义、无类型和日期输入字段，因此无法选择这些字段。

“密度函数”视图

通过此视图中的选项，您可以针对连续目标定制概率密度函数和累积分布函数的输出，以及针对分类目标定制预测值的条形图。

密度函数。 密度函数是根据模拟来探测结果集的主要方式。

- **概率密度函数 (PDF)**。 选择此选项可针对目标字段生成概率密度函数。 概率密度函数用于显示目标值的分布。 您可以使用概率密度函数来确定目标位于特定区域中的概率。 对于分类目标（具有名义或有序测量级别的目标），将生成一个条形图，用于显示落入每个目标类别内的观测值的百分比。

- **累积分布函数 (CDF)**。 选择此选项可针对目标字段生成累积分布函数。 累积分布函数用于显示目标值小于或等于指定值的概率。 此函数仅可用于连续目标。

参考线（连续）。 只有在选择了**概率密度函数 (PDF)** 和/或**累积分布函数 (CDF)** 的情况下，才会启用这些选项。 通过这些选项，您可以向概率密度函数和累积分布函数添加多条固定的垂直参考线。

- **平均值**。 选择此选项可在目标字段的均值处添加参考线。
- **中位数**。 选择此选项可在目标字段的中值处添加参考线。
- **标准差**。 选择此选项可在目标字段均数加上或减去指定标准差数目的位置添加参考线。 选择此选项将启用相邻的**数目**字段。 单击箭头可指定标准差的值。 标准差的最小值为 1，最大值为 10。 标准差的值最初设置为 3。
- **百分位数**。 选择此选项可在目标字段分布的两个百分位值处添加参考线。 选择此选项将启用相邻的**后几个值**和**前几个值**文本字段。 例如，在**前几个值**文本字段中输入值 90 将在目标的第 90 百分位数处添加参考线，90% 的观测值将落入该值以下。 同样，**后几个值**文本字段中的值 10 表示目标的第 10 百分位数，10% 的观测值将落入该值以下。
- **定制参考线**。 选择此选项可沿水平轴在指定值处添加参考线。 选择此选项将启用相邻的**值**表。 每次在**值**表中输入一个有效数字时，都会向表的底部追加一个新的空行。 有效数字是在目标字段的值范围内的数字

注: 如果在单个图表上显示了多个密度函数或分布函数（来自多个迭代），那么将分别对每个函数应用参考线（而不是定制线）。

分类目标（仅限于 PDF）。 只有在选择了**概率密度函数 (PDF)** 的情况下，才会启用这些选项。

- **要报告的类别值**。 对于具有分类目标字段的模型，此模型的结果是目标值落入每个类别的一组预测概率（每个类别对应一个概率）。 具有最高概率的类别将作为预测类别，并用于生成概率密度函数的条形图。 选择**预测类别**可生成条形图。 选择**预测概率**可针对目标字段的每个类别生成预测概率的分布直方图。 您也可以选择**两者**来生成这两种类型的图表。
- **敏感度分析的分组**。 包含敏感度分析迭代的模拟将针对分析所定义的每个迭代生成一个独立的目标字段（或模型中的预测目标字段）。 对于变化的分布参数的每个值，存在一个迭代。 如果存在迭代，那么分类目标字段的预测类别条形图将显示为一个包含所有迭代结果的复式条形图。 请选择**将类别分组到一起**或**将迭代分组到一起**。

“输出”视图

目标分布的百分位值。 通过这些选项，您可以选择创建目标分布的百分位值表，并指定要显示的百分位数。

创建一个百分位值表。 对于连续目标字段，选择此选项可获取目标分布的指定百分位数表。 请选择下列其中一个选项来指定百分位数:

- **四分位数** 四分位数是指目标字段分布的第 25 百分位数、第 50 百分位数和第 75 百分位数。 观测值被分为大小相等的四个组。
- **区间**。 如果需要数目相等的组（除 4 以外），请选择**区间**。 选择此选项将启用相邻的**数目**字段。 单击箭头可指定区间数。 区间的最小值为 2，最大值为 100。 区间的值最初设置为 10。
- **定制百分位数**。 选择**定制百分位数**可指定各个百分位数，例如第 99 百分位数。 选择此选项将启用相邻的**值**表。 每次在**值**表中输入一个有效数字（介于 1 到 100 之间）时，都会向表的底部追加一个新的空行。

“模拟评估”节点输出

执行“模拟评估”节点时，输出将添加到输出管理器中。 “模拟评估”输出浏览器显示了执行“模拟评估”节点的结果。 **文件**菜单中提供了常用的保存、导出和打印选项，而**编辑**菜单中提供了常用的编辑选项。 有关更多信息，请参阅主题 第 230 页的『查看输出』。 只有在选择了其中一个图表的情况下，才会启用**视图**菜单。 对于分布表或信息输出，未启用此菜单。 您可以从**视图**菜单中选择**编辑方式**来更改图表的布局和外观，或者选择**探索方式**来探索图表所表示的数据和值。 静态方式会将图表参考线固定在其当前位置，因此无法移动这些参考线。 只有在静态方式下才能复制、打印或导出带有参考线的图表。 要选择此方式，请在**视图**菜单中单击**静态方式**。

“模拟评估”输出浏览器窗口包含两个面板。 导航面板位于窗口左侧，用于显示执行“模拟评估”节点时生成的图表的缩略图表示。 选择缩略图后，图表输出将显示在窗口右侧的面板中。

导航面板

输出浏览器的导航面板包含根据模拟生成的图表的缩略图。 导航面板上显示的缩略图取决于目标字段的测量级别，以及在“模拟评估”节点对话框中选择的选项。 下表中提供了对这些缩略图描述。

表 *45:* 导航面板缩略图

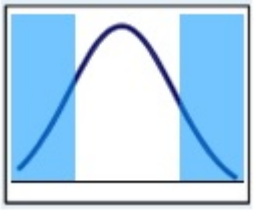
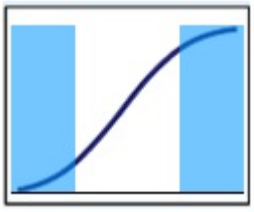
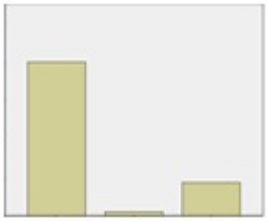
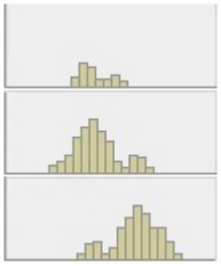
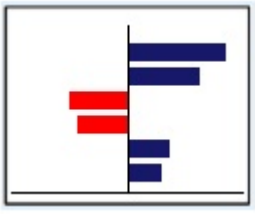
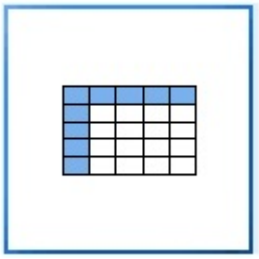
缩略图	描述	注释
	概率密度函数	只有在目标字段的测量级别为连续，并且在“模拟评估”节点对话框的“密度函数”视图中选择了**概率密度函数 (PDF)** 的情况下，才会显示此缩略图。 如果目标字段的测量级别为分类，那么将不会显示此缩略图。
	累积分布函数	只有在目标字段的测量级别为连续，并且在“模拟评估”节点对话框的“密度函数”视图中选择了**累积分布函数 (CDF)** 的情况下，才会显示此缩略图。 如果目标字段的测量级别为分类，那么将不会显示此缩略图。
	预测类别值	只有在目标字段的测量级别为分类，在“模拟评估”节点对话框的“密度函数”视图中选择了**概率密度函数 (PDF)**，以及在**要报告的类别值**区域中选择了**预测类别**或**两者**的情况下，才会显示此缩略图。 如果目标字段的测量级别为连续，那么将不会显示此缩略图。
	预测类别概率	只有在目标字段的测量级别为分类，在“模拟评估”节点对话框的“密度函数”视图中选择了**概率密度函数 (PDF)**，以及在**要报告的类别值**区域中选择了**预测概率**或**两者**的情况下，才会显示此缩略图。 如果目标字段的测量级别为连续，那么将不会显示此缩略图。
	龙卷风图	只有在“模拟评估”节点对话框的“字段”视图中的**相关性龙卷风的输入字段**字段中选择了一个或多个输入字段的情况下，才会显示此缩略图。

表 45: 导航面板缩略图 (继续)

缩略图	描述	注释
	分布表	只有在目标字段的测量级别为连续，并且在“模拟评估”节点对话框的“输出”视图中选择了**创建百分位值表**的情况下，才会显示此缩略图。 对于此图表，**视图**菜单处于禁用状态。 如果目标字段的测量级别为分类，那么将不会显示此缩略图。
	信息	始终显示此缩略图。 对于此输出，**视图**菜单处于禁用状态。

图表输出

可用的输出图表的类型取决于目标字段的测量级别、是否使用了迭代字段以及在“模拟评估”节点对话框中选择的选项。 根据模拟生成的很多图表都具有交互式功能，您可以使用此功能定制显示。 单击**图表选项**可以使用交互式功能。 所有模拟图表都采用图形板直观表示。

连续目标的概率密度函数图。 此图表同时显示概率和频率，概率刻度位于左侧的垂直轴上，而频率刻度位于右侧的垂直轴上。 此图表具有两条滑动垂直参考线，这些参考线将该图表划分为一些单独的区域。 此图表下方的表显示了每个区域中的分布所占的百分比。 如果同一图表上显示了多个密度函数（由于迭代），那么此表将为与每个密度函数关联的概率提供单独的一行，并提供额外的一列，用于显示迭代名称以及与每个密度函数关联的颜色。 迭代将根据迭代标签按字母顺序列示在表中。 如果迭代标签不可用，那么将改为使用迭代值。 此表无法进行编辑。

每条参考线都有一个滑块（倒三角形），您可以使用该滑块轻松移动此参考线。 每个滑块都有一个标签，用于指示其当前位置。 缺省情况下，滑块位于分布的第 5 百分位数和第 95 百分位数处。 如果存在多个迭代，那么滑块将位于表中列出的第一个迭代的第 5 百分位数和第 95 百分位数处。 您不能使这两条线穿过对方。

通过单击**图表选项**，可以使用一些附加功能。 特别是，您可以明确地设置滑块位置、添加固定的参考线，以及将图表视图由连续曲线更改为直方图。 有关更多信息，请参阅主题 第 257 页的『图表选项』。 右键单击图表可复制或导出该图表。

连续目标的累积分布函数图。 此图表具有两个可移动的垂直参考线和关联表，这些参考线和表与针对概率密度函数图描述的参考线和表相同。 存在多个迭代时，滑块控件和表的行为与概率密度函数相同。 用于标识哪个密度函数属于每个迭代的同一颜色也用于分布函数。

此图表还提供了对“**图表选项**”对话框的访问，通过该对话框您可以明确地设置滑块位置、添加固定的参考线以及指定累积分布函数是显示为递增函数（缺省值）还是递减函数。 有关更多信息，请参阅主题 第 257 页的『图表选项』。 右键单击图表可复制、导出或编辑该图表。 选择**编辑**将在浮动的图形板编辑器窗口中打开此图表。

分类目标的预测类别值图表。 对于分类目标字段，条形图显示了预测值。 预测值显示为预测落入每个类别的目标字段的百分比。 对于具有敏感度分析迭代的分类目标字段，预测目标类别的结果将显示为包含所有迭代的结果的复式条形图。 此图表按类别或按迭代进行聚类，具体取决于在“模拟评估”节点对话框的“密度函数”视图中的**敏感度分析分组**区域内选择了哪个选项。 右键单击图表可复制、导出或编辑该图表。 选择**编辑**将在浮动的图形板编辑器窗口中打开此图表。

分类目标的预测类别概率图表。 对于分类目标字段，直方图将显示目标的每个类别的预测概率分布。 对于具有敏感度分析迭代的分类目标字段，将按类别或按迭代显示直方图，具体取决于在“模拟评估”节点对话框

的“密度函数”视图中的**敏感度分析分组**区域内选择了哪个选项。 如果直方图按类别进行分组，那么您可以通过包含迭代标签的下拉列表来选择要显示的迭代。 另外，也可以通过右键单击图表并从**迭代**子菜单中选择迭代来选择要显示的迭代。 如果直方图按迭代进行分组，那么您可以通过包含类别名称的下拉列表来选择要显示的类别。 另外，也可以通过右键单击图表并从**类别**子菜单中选择类别来选择要显示的类别。

此图表仅可用于部分模型，并且必须在模型块上选中生成所有组概率的选项。 例如，在 Logistic 模型块上，您必须选中**追加所有概率**。 下列模型块支持此选项：

- Logistic、SVM、贝叶斯、神经网络和 KNN
- Logistic 回归的 Db2/ISW 数据库内挖掘模型、决策树和朴素贝叶斯

缺省情况下，未在这些模型块上选中生成所有组概率的选项。

龙卷风图。 龙卷风图是一个条形图，它显示了目标字段对每个指定输入的敏感度。 敏感度通过目标与每个输入之间的相关性进行度量。 图表标题包含目标字段的名称。 图表上的每个条形都表示目标字段与输入字段之间的相关性。 该图表中包含的模拟输入是在“模拟评估”节点对话框的“字段”视图中的**相关性龙卷风的输入字段**字段中选择的输入。 每个条形都通过相关性值进行标注。 这些条形按相关性绝对值从大到小的顺序排序。 如果存在迭代，那么将针对每个迭代生成一个单独的图表。 每个图表都具有一个子标题，其中包含迭代的名称。

分布表。 此表包含目标字段的值，指定百分比的观测值将落入该值以下。 此表为“模拟评估”节点对话框的“输出”视图中指定的每个百分位值提供了一行。 百分位值可以是四分位数、空间相等的其他百分位数或分别指定的百分位数。 分布表为每个迭代提供了一列。

信息。 此部分提供在评估中使用的字段和记录的整体摘要。 它还显示了输入字段和记录计数，可以针对每个迭代细分这些字段和计数。

图表选项

您可以在“图表选项”对话框中定制根据模拟生成的概率密度函数和累积分布函数的活动图表显示。

视图。 **视图**下拉列表仅适用于概率密度函数图。 您可以使用此列表将图表视图由连续曲线切换为直方图。 在同一图表上显示了多个密度函数（来自多个迭代）的情况下，此功能将处于禁用状态。 如果存在多个密度函数，那么只能以连续曲线形式查看这些密度函数。

顺序。 **顺序**下拉列表仅适用于累积分布函数图。 它指定累积分布函数显示为递增函数（缺省值）还是递减函数。 如果显示为递减函数，那么水平轴上指定点处的函数值是目标字段位于该点右侧的概率。

滑块位置。 **上限**文本字段包含右滑动参考线的当前位置。 **下限**文本字段包含左滑动参考线的当前位置。 通过在**上限**和**下限**文本字段中输入值，您可以明确地设置滑块的位置。 **下限**文本字段中的值必须严格小于**上限**文本字段中的值。 可以选择**负无穷大**来移除左参考线，从而将位置有效地设置为负无穷大。 此操作将禁用**下限**文本字段。 可以选择**无穷大**来移除右参考线，从而将其位置有效地设置为无穷大。 此操作将禁用**上限**文本字段。 您不能同时移除这两条参考线；选择**负无穷大**将禁用**无穷大**复选框，反之亦然。

参考线。 您可以向概率密度函数和累积分布函数添加多条固定的垂直参考线。

- **平均值**。 可以在目标字段的均数处添加参考线。
- **中位数**。 可以在目标字段的中位数处添加参考线。
- **标准差**。 可以在目标字段均数加上或减去指定标准差数目的位置添加参考线。 您可以在相邻的文本字段中输入要使用的标准差值。 标准差的最小值为 1，最大值为 10。 标准差的值最初设置为 3。
- **百分位数**。 通过在**后几个值**和**前几个值**文本字段中输入值，您可以在目标字段分布的一个或两个百分位值处添加参考线。 例如，**前几个值**文本字段中的值 95 表示第 95 百分位数，95% 的观测值将落入该值以下。 同样，**后几个值**文本字段中的值 5 表示第 5 百分位数，5% 的观测值将落入该值以下。 对于**后几个值**文本字段，最小百分位值为 0，最大百分位值为 49。 对于**前几个值**文本字段，最小百分位值为 50，最大百分位值为 100。
- **定制位置**。 您可以沿水平轴在指定值处添加参考线。 通过从网格中删除该条目，您可以移除定制参考线。

在您单击**确定**后，滑块、滑块上方的标签、参考线以及图表下方的表将进行更新，以反映“图表选项”对话框中选择的选项。 单击**取消**可以在不进行任何更改的情况下关闭此对话框。 可以通过取消选择“图表选项”对话框中的关联选项并单击**确定**来移除参考线。

注: 如果在单个图表上显示了多个密度函数或分布函数（由于结果来自敏感度分析迭代），那么将分别对每个函数应用参考线（而不是定制线）。 将只显示第一个迭代的参考线。 参考线标签包含迭代标签。 迭代标签派生自上游，通常派生自“模拟生成”节点。 如果迭代标签不可用，那么将改为使用迭代值。 对于具有多个迭代的累积分布函数，**均数**、**中位数**、**标准差**和**百分位数**选项处于禁用状态。

“扩展输出”节点

如果在“扩展输出”节点对话框的**输出**选项卡上选择了**输出到屏幕**，那么将在输出浏览器窗口中显示屏幕上的输出。 这还会将输出添加到输出管理器。 输出浏览器窗口具有它自己的菜单集，使用这些菜单，可以打印或保存输出，也可以将输出导出为其他格式。 **编辑**菜单仅包含**复制**选项。 “扩展输出”节点的输出浏览器有两个选项卡：用于显示文本输出的**文本输出**选项卡，以及用于显示图形和图表的**图形输出**选项卡。

如果在“扩展输出”节点对话框的**输出**选项卡上选择了**输出到文件**，那么在成功执行“扩展输出”节点后，不会显示输出浏览器窗口。

“扩展输出”节点 -“语法”选项卡

选择语法类型 - **R** 或 **Python for Spark**。 请参阅以下部分以获取更多信息。 语法就绪时，您可以单击**运行**来执行“扩展输出”节点。 输出对象将添加到输出管理器，或者选择性地添加到**输出**选项卡上的**文件名**字段中指定的文件中。

R 语法

R 语法。 您可以在此字段中输入或粘贴用于数据分析的定制 R 脚本语法。

转换标志字段。 指定标志字段的处理方式。 共有两个选项：**将字符串转换为因子，将整数和实数转换为双精度数**和**逻辑值（True 和 False）**。 如果选择**逻辑值（True 和 False）**，那么标志字段的原始值将丢失。 例如，如果某个字段的值为 Male 和 Female，那么这些值将更改为 True 和 False。

将缺失值转换为 R "不可用" 值 (NA)。 选中时，任何缺失值都将转换为 R NA 值。 R 使用值 NA 来标识缺失值。 您使用的某些 R 函数可能有一个参数，可用于控制当数据包含 NA 时函数的行为方式。 例如，该函数可能会允许您选择自动排除包含 NA 的记录。 如果未选择此选项，那么所有缺失值都将按原样传递到 R，并可能导致执行 R 脚本时发生错误。

将日期/时间字段转换为特殊时区控制的 R 类。 如果选择此选项，那么会将带有日期或日期时间格式的变量转换为 R 日期/时间对象。 必须选择下列选项之一：

- 将具有日期或日期时间格式的 **RPOSIXct.** 变量将转换为 R POSIXct 对象。
- **R POSIXlt （列表）**。 将具有日期或日期时间格式的变量转换为 R POSIXlt 对象。

注: POSIX 格式是高级选项。 仅当您的 R 脚本指定以需要这些格式的方式处理日期时间字段时才使用这些选项。 POSIX 格式不适用于具有时间格式的变量。

Python 语法

Python 语法。 您可以针对数据分析向此字段中输入或粘贴定制的 Python 脚本语法。 有关 Python for Spark 的更多信息，请参阅 Python for Spark 和 Python for Spark 的脚本编制。

“扩展输出”节点 -“控制台输出”选项卡

控制台输出选项卡包含当“语法”选项卡上的 R 脚本或 Python for Spark 脚本运行时接收到的任何输出（例如，如果使用 R 脚本，当执行**语法**选项卡上的 **R 语法**字段中的 R 脚本时，它显示从 R 控制台接收到的输出）。 此输出可能包括执行 R 或 Python 脚本时生成的 R 或 Python 错误消息或警告。 输出可主要用于调试脚本。 **控制台输出**选项卡还包含 **R 语法**或 **Python 语法**字段中的脚本。

每次执行“扩展输出”脚本时，都会使用从 R 控制台或 Python for Spark 接收到的输出来覆盖**控制台输出**选项卡的内容。 输出不能编辑。

“扩展输出”节点 -“输出”选项卡

输出名称。 指定执行此节点时生成的输出的名称。 选择**自动**时，输出名称将自动设置为“R 输出”或“Python 输出”，具体取决于脚本类型。 （可选）可以选择**定制**以指定其他名称。

输出到屏幕。 选择此选项以在新窗口中生成并显示输出。 这还会将输出添加到输出管理器。

输出到文件。 选择此选项可将输出保存到文件。 执行此操作将启用**输出图形**和**输出文件**单选按钮。

输出图形。 只有在选择了**输出到文件**的情况下才会启用此选项。 选择此选项可以将执行“扩展输出”节点所产生的任何图形保存到文件中。 请在**文件名**字段中指定要用于生成的输出的文件名。 单击省略符按钮 (**...**) 以选择特定文件和位置。 请在**文件类型**下拉列表中指定文件类型。 可用的文件类型如下所示：

- 输出对象 (`.cou`)
- HTML (`.html`)

输出文本。 只有在选择了**输出到文件**的情况下才会启用此选项。 选择此选项可以将执行“扩展输出”节点所产生的任何文本输出保存到文件中。 请在**文件名**字段中指定要用于生成的输出的文件名。 单击省略符按钮 (**...**) 以指定特定文件和位置。 请在**文件类型**下拉列表中指定文件类型。 可用的文件类型如下所示：

- HTML (`.html`)
- 输出对象 (`.cou`)
- 文本文档 (`.txt`)

扩展输出浏览器

如果在“扩展输出”节点对话框的**输出**选项卡上选择了**输出到屏幕**，那么将在输出浏览器窗口中显示屏幕上的输出。 这还会将输出添加到输出管理器。 输出浏览器窗口具有它自己的菜单集，使用这些菜单，可以打印或保存输出，也可以将输出导出为其他格式。 **编辑**菜单仅包含**复制**选项。 “扩展输出”节点的输出浏览器有两个选项卡：

- **文本输出**选项卡显示文本输出
- **图形输出**选项卡显示图形和图表

如果在“扩展输出”节点对话框的**输出**选项卡上选择了**输出到文件**，而不是**输出到屏幕**，那么在成功执行“扩展输出”节点后，不会显示输出浏览器窗口。

扩展输出浏览器 -“文本输出”选项卡

文本输出选项卡显示执行“扩展输出”节点的**语法**选项卡上的 R 脚本或 Python for Spark 脚本时生成的任何文本输出。

注： 由于执行扩展输出脚本而产生的 R 或 Python for Spark 错误消息或警告始终显示在“扩展输出”节点的**控制台输出**选项卡上。

扩展输出浏览器 -“图形输出”选项卡

Python Spark 的 "扩展输出" 节点现在包含一个类似于 R 的 **图形输出** 选项卡。 此选项卡显示执行 "扩展输出" 节点的 **语法** 选项卡上的 R 脚本或 Python for Spark 脚本时生成的任何图形或图表。 例如，如果您的 R 脚本包含对 R `plot` 函数的调用，那么产生的图形会显示在此选项卡上。 现在，您可以使用如下脚本来生成图形：

```
import spss.pyspark.runtime

import numpy
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

ascontext = spss.pyspark.runtime.getContext()
indf = ascontext.getSparkInputData()
sns.pairplot(indf.toPandas(), hue='K')
sns.pairplot(indf.toPandas(), hue='Age')
plt.show()
```

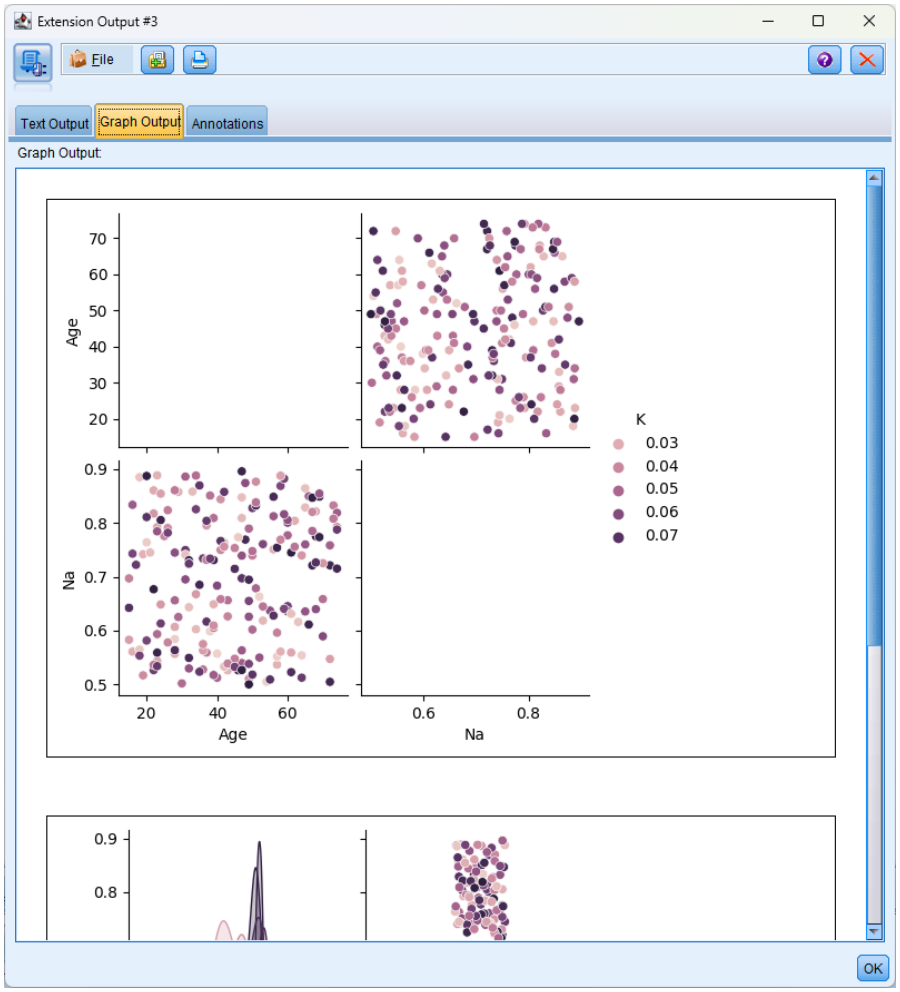


图 *75:* “图形输出”选项卡

KDE 节点

Kernel Density Estimation (KDE)© 使用 Ball Tree 或 KD Tree 算法以进行效率查询，并且游走于无监督学习、特征工程和数据建模。 基于相邻元素的方法（例如，KDE）是最流行且最有用的一些密度估算方法。 可在任意数量的维度执行 KDE，但是在实践当中，高维数可能导致性能下降。 SPSS Modeler 中的 KDE 建模和 KDE 模拟节点公开 KDE 库的核心特征和常用参数。 节点使用 Python 进行实现。 [1]

要使用 KDE 节点，必须设置上游“类型”节点。 KDE 节点将从“类型”节点（或者上游源节点的“类型”选项卡）读取输入值。

KDE 建模节点位于 SPSS Modeler 的“建模”选项卡和 Python 选项卡上。 “KDE 建模”节点生成一个模型块，并且块的评分值是来自输入数据的核心密度值。

KDE 模拟节点位于“输出”选项卡和 Python 选项卡上。 “KDE 模拟”节点生成 KDE Gen 源节点，后者可创建一些使用相同分布作为输入数据的记录。 KDE Gen 节点包含“设置”选项卡，可在其中指定节点将创建的记录数（缺省值为 1）并生成随机种子。

有关 KDE 的更多信息，包括示例，请参阅 KDE 文档 (http://scikit-learn.org/stable/modules/density.html#kernel-density-estimation)。 [1]

[1] “User Guide”。 *Kernel Density Estimation*. Web. © 2007-2018, scikit-learn developers.

KDE 建模节点和 KDE 模拟节点字段

“字段”选项卡指定要在分析中使用的字段。

使用预定义角色。 此选项使用上游“类型”节点（或上游源节点的“类型”选项卡）中的输入设置。

使用定制字段分配。 要手动分配输入，请选择此选项。

字段。 使用方向按钮可以将此列表中的项手动分配到屏幕右侧的“输入”列表。 图标指示每个字段的有效测量级别。 要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

输入。 选择一个或多个字段作为聚类输入。 KDE 只能处理连续字段。

KDE 节点构建选项

使用“构建选项”选项卡以指定 KDE 节点的构建选项，包括用于内核密度参数和集群标签的**基本选项**，以及**高级选项**，例如，容差、叶大小以及是否使用广度优先方法。 有关这些选项的更多信息，请参阅以下在线资源：

- 内核密度估算 Python API 参考 [1]
- 内核密度估算用户指南 [2]

基本

带宽。 指定内核的带宽。

内核。 选择要使用的内核。 KDE 建模节点的可用内核为 **Gaussian**、**Tophat**、**Epanechnikov**、**Eponential**、**Linear** 或 **Cosine**。 KDE 模拟节点的可用内核为 **Gaussian** 或 **Tophat**。 有关这些可用内核的详细信息，请参阅内核密度估算用户指南。[2]

算法。 对于要使用的树算法，选择 **Auto**、**Ball Tree** 或 **KD Tree**。 有关更多信息，请参阅 Ball Tree[3] 和 KD Tree。[4]

度量。 选择距离度量。 可用度量为 **Euclidean**、**Braycurtis**、**Chebyshev**、**Canberra**、**Cityblock**、**Dice**、**Hamming**、**Infinity**、**Jaccard**、**L1**、**L2**、**Matching**、**Manhattan**、**P**、**Rogerstanimoto**、**Russellrao**、**Sokalmichener**、**Sokalsneath**、**Kulsinski** 或 **Minkowski**。 如果选择 **Minkowski**，那么根据需要设置 **P 值**。

此下拉列表中可用的度量将根据选择的算法的不同而不同。 另外，请注意，密度输出的标准化仅针对 Euclidean 距离度量正确。

高级

绝对容差。 指定期望的结果的绝对容差。 较大的容差通常将导致更快的运行时间。 缺省值为 **0.0**。

相对容差。 指定期望的结果的相对容差。 较大的容差通常将导致更快的运行时间。 缺省值为 **1E-8**。

叶大小。 指定底层树的叶大小。 缺省值为 **40**。 更改叶大小可能会显著影响性能和所需的内存。 有关 Ball Tree 和 KD Tree 算法的更多信息，请参阅 Ball Tree[3] 和 KD Tree。[4]

广度优先。 如果想要使用广度优先方法，那么选择 **True**，或者选择 **False** 以使用深度优先方法。

下表显示 SPSS Modeler KDE 节点对话框中的设置与 Python KDE 库参数之间的关系。

表 *46:* 映射到 *Python* 库参数的节点属性

SPSS Modeler 设置	**脚本名称（属性名称）**	**KDE 参数**
输入	`inputs`	
带宽(B)	`bandwidth`	`bandwidth`
内核(K)	`kernel`	`kernel`
算法	`algorithm`	`algorithm`
指标	`metric`	`metric`
P 值	`pValue`	`pValue`
绝对容差	`atol`	`atol`

表 46: 映射到 *Python* 库参数的节点属性 *(继续)*		
SPSS Modeler 设置	**脚本名称（属性名称）**	**KDE 参数**
相对容差	`rtol`	`Rtol`
叶大小	`leafSize`	`leafSize`
广度优先	`breadthFirst`	`breadthFirst`

[1] "API Reference." *sklearn.neighbors.KernelDensity*。 Web. © 2007-2018, scikit-learn developers.

[2] “User Guide”。 *Kernel Density Estimation*. Web. © 2007-2018, scikit-learn developers.

[3] “Ball Tree”。 *Five balltree construction algorithms*。 © 1989, Omohundro, S.M., 国际计算机科学研究所技术报告。

[4] “K-D Tree”。 *Multidimensional binary search trees used for associative searching*。 © 1975, Bentley, J.L., ACM 的通信。

KDE 建模节点和 KDE 模拟节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

IBM SPSS Statistics 帮助应用程序

如果在您的计算机上安装并许可了 IBM SPSS Statistics 的兼容版本，则可以使用 统计信息“变换”、统计信息“模型”、统计信息“输出”或 统计信息“导出”节点将 IBM SPSS Modeler 配置为使用 IBM SPSS Statistics 功能来处理数据。

有关 IBM SPSS Modeler 当前版本的产品兼容性的信息，请访问公司支持站点 (*http://www.ibm.com/support*)。

要配置 IBM SPSS Modeler 以便与 IBM SPSS Statistics 和其他应用程序一起使用，选择：

工具 > 选项 > 帮助应用程序

IBM SPSS Statistics Interactive。 输入要直接在“统计信息导出”节点所生成的数据文件上启动 IBM SPSS Statistics 时使用的命令的完整路径和名称（例如，`C:\Program Files\IBM\SPSS\Statistics\<nn>\stats.exe`）。 有关更多信息，请参阅第 277 页的『Statistics 导出节点』。

连接。 如果 IBM SPSS Statistics Server 位于 IBM SPSS Modeler Server 所在的主机，那么可以在两个应用程序之间启用一个连接，以便在分析过程中将数据保留在服务器上以提高效率。 选择 **服务器** 以启用下列 **端口** 选项。 缺省设置为 **本地** 。

端口。 为 IBM SPSS Statistics Server 指定服务器端口。

IBM SPSS Statistics Location Utility。 要启用 IBM SPSS Modeler 以使用 Statistics 变换、Statistics 模型和 Statistics 输出节点，您必须在运行流的计算机上拥有 IBM SPSS Statistics 安装和许可的一个副本。

- 如果以本地（独立）方式运行 IBM SPSS Modeler，那么 IBM SPSS Statistics 的许可副本必须位于本地计算机上。 单击该按钮以指定想要用于许可的本地 IBM SPSS Statistics 安装的位置。
- 此外，如果是针对远程 IBM SPSS Modeler Server 以分布式方式运行，那么还需要在 IBM SPSS Modeler Server 主机上运行一个实用程序来创建 *statistics.ini* 文件，此文件向 IBM SPSS Statistics 指出 IBM SPSS Modeler Server 的安装路径。 为此，请在命令提示符下切换至 IBM SPSS Modeler Server *bin* 目录，然后运行以下命令（针对 Windows 系统）：

  ```
  statisticsutility -location=<IBM SPSS Statistics_installation_path>/
  ```

  或者对于 UNIX 系统，运行以下命令：

  ```
  ./statisticsutility -location=<IBM SPSS Statistics_installation_path>/bin
  ```

本地计算机上没有 IBM SPSS Statistics 的许可副本时，您仍然可以对 IBM SPSS Statistics 服务器运行“Statistics 文件”节点，但尝试运行其他 IBM SPSS Statistics 节点将显示一条错误消息。

备注

如果在运行 IBM SPSS Statistics 过程节点时遇到困难，可考虑以下提示：

- 如果 IBM SPSS Modeler 中使用的字段名超过八个字符（对于 IBM SPSS Statistics 12.0 之前的版本），或 64 个字符（对于 IBM SPSS Statistics 12.0 及之后的版本），或者字段名中包含无效字符，则在将这些字段名读入到 IBM SPSS Statistics 之前有必要将其重命名或截断。 有关更多信息，请参阅第 278 页的『重命名或过滤 IBM SPSS Statistics 的字段』。
- 如果 IBM SPSS Statistics 是在 IBM SPSS Modeler 之后安装，那么可能需要指定 IBM SPSS Statistics 位置（如上所述）。

第 7 章 导出节点

导出节点概述

“导出”节点提供一种导出各种格式的数据以与其他软件工具相互作用的机制。

可用的导出节点有：

“数据库导出”节点将数据写入符合 ODBC 标准的关系数据源。 为了写入 ODBC 数据源，该数据源必须存在，并且您必须对其具有写入许可权。

“平面文件导出”节点将数据输出到定界文本文件中。 它对于导出可由其他分析或电子表格软件读取的数据非常有用。

Statistics 导出节点以 IBM SPSS Statistics *.sav* 或 *.zsav* 格式输出数据。 `.sav` 或 `.zsav` 文件可以由 IBM SPSS Statistics Base 和其他产品读取。 这也是用于 IBM SPSS Modeler 中的高速缓存文件的格式。

数据收集 导出节点以 数据收集 市场调查软件使用的格式输出数据。 要使用此节点，必须安装 数据收集 Data Library。

IBM Cognos 导出节点以 Cognos 数据库可以读取的格式导出数据。

IBM Cognos TM1 导出节点以 Cognos TM1 数据库可以读取的格式导出数据。

“SAS 导出”节点以 SAS 格式输出数据，以便将该数据读入 SAS 或者与 SAS 兼容的软件包。 提供三种 SAS 文件格式：SAS for Windows/OS2、SAS for UNIX 或 SAS V7/8。

Excel 导出节点在 Microsoft Excel 中输出数据。`xlsx` 文件格式。 您还可选择在执行完此节点后自动启动 Excel 并打开导出的文件。

“XML 导出”节点将数据以 XML 格式输出到文件。 还可选择创建 XML 源节点，以将导出的数据读取回到流中。

JSON 导出节点以 JSON 格式输出数据。

数据库导出节点

可使用“数据库”节点将数据写入与 ODBC 兼容的关系数据源，请参阅“数据库”源节点相关说明。 有关更多信息，请参阅主题 第 13 页的『“数据库源”节点』。

请使用以下常用步骤将数据写入数据库：

1. 为要使用的数据库安装 ODBC 驱动程序并配置数据源。
2. 在“数据库”节点的“导出”选项卡中，指定要写入的数据源和表。 可创建新表或将数据插入现有表。
3. 根据需要指定其他选项。

在下面的几个主题中将对这些步骤进行更详细地说明。

数据库节点的“导出”选项卡

注： 您可以导出到的某些数据库可能不支持表中长度超过 30 个字符的列名。 如果显示提醒您的表有不正确的列名的错误消息，请将名称的大小减小至少于 30 个字符。

数据源。 显示所选数据源。 输入名称或者从下拉列表中选择名称。 如果列表中未显示所需的数据库，则选择**添加新的数据库连接**并从“数据库连接”对话框选定数据库。 有关更多信息，请参阅 第 14 页的『添加数据库连接』。

表名称。 输入接收数据的表名称。 如果选择 **插入到表中** 选项，

- 您可以通过单击 **选择** 按钮来选择数据库中的现有表。
- 如果提供当前不存在的表名，那么将创建具有指定名称的新表，并将数据插入其中。

创建表。 选中此项可创建一个新的数据库表或覆盖现有的数据库表。

插入表中。 选择此选项以便：

- 将数据插入现有数据库表中的新行，或者
- 将数据插入到不存在的表中。 在这种情况下，将创建新表，并将数据作为新创建的表中的新行插入。

合并表。 （如可用）选择此选项以使用相应源数据字段中的值更新所选数据库列。 选择此选项会启用**合并**按钮，在其显示的对话框中您可以将源数据字段映射到数据库列。

删除现有表。 选中此项可在创建新表时删除所有名称相同的现有表。

删除现有行。 选择此选项可在插入表时先将现有行从表中删除然后导出。

注： 如果选择上述任意两个选项，那么在执行节点时将收到**覆盖警告**消息。 要想不显示此警告，请取消选择“用户选项”对话框的“通知”选项卡上的**当节点覆盖数据表时发出警告**选项。

缺省字符串大小。 上游“类型”节点中标记为无类型的字段将作为字符串字段写入数据库。 请指定无类型字段要使用的字符串大小。

单击**模式**可打开一个对话框，您可在其中设置各种导出选项（对于支持此功能的数据库）、设置所需字段的 SQL 数据类型，并指定创建数据库索引的主要关键字。 有关更多信息，请参阅 第 267 页的『数据库导出模式选项』。

单击 **索引** 可指定创建导出表索引的选项，以提高数据库的运行性能。 有关更多信息，请参阅 第 269 页的『数据库导出索引选项』。

单击 **高级** 可指定批量加载和数据库提交选项。 有关更多信息，请参阅 第 270 页的『数据库导出高级选项』。

将表和列名加上引号。 选择将 CREATE TABLE 语句发送到数据库时使用的选项。 必须为包含空格和非标准字符的表和列添加引号。

- **根据需要**。 选择此选项，以允许 IBM SPSS Modeler 自动根据各个情况确定是否需要添加引号。
- **始终**。 选中此项将始终为表名和列名称添加引号。
- **永不**。 选中此项将禁用引号。

为此数据生成导入节点。 选中此项可在将数据导出到指定数据源和表时生成此数据的数据源节点。 执行此操作后，此节点即被添加到流工作区。

数据库导出合并选项

此对话框使您能够将字段从源数据映射到目标数据库表中的列。 其中源数据字段被映射到数据库列，在运行流时列值被替换为源数据值。 未映射的源字段在数据库中保持不变。

映射字段。 您可在这里指定源数据字段与数据库列之间的映射。 如果源数据字段与数据库中的列具有相同的名称，那么自动将其映射。

- **映射**。 将按钮左侧字段列表中选中的源数据字段映射到右侧列表中选中的数据库列。 您可一次映射多个字段，但两个列表中选中的条目数必须相同。
- **取消映射**。 删除一个或多个选中的数据库列的映射。 在对话框右侧的表格中选择字段或数据库列后，将激活此按钮。
- **添加**。 将按钮左侧字段列表中选中的一个或多个源数据字段添加到右侧的列表中以待映射。 在左侧列表中选择字段，并且右侧列表中不存在具有该名称的字段时，将激活此按钮。 单击此按钮可以将选定字段映射到具有同一名称的新数据库列。 在数据库列名后会显示“**<NEW>**”一词，指示这是一个新字段。

合并行。 使用键字段（如交易标识）合并在键字段中具有相同值的记录。 此选项等同于数据库的“相等连接”。 关键字值必须为这些主要键字段的值；也就是说，它们必须是唯一的，且不得包含空值。

- **可能的关键字**。 列出所有输入数据源中的所有字段。 从此列表选择一个或多个字段，并使用箭头按钮将其添加为用于合并记录的键字段。 具有相应映射数据库列的任何映射字段都可用作键，但作为新数据库列添加的字段（在名称后显示 **<NEW>**）不可用。
- **用于合并的键**。 基于关键字段值，列出所有用于从所有输入数据源中合并记录的字段。 要从列表中移除关键字段，请选择一个关键字段，然后使用箭头按钮将其返回到“可能的关键字”列表中。 如果选择了多个关键字段，那么下面的选项将启用。
- **仅包括数据库中存在的记录**。 执行部分连接；如果记录同时在数据库和流中，那么将更新映射字段。
- **将记录添加到数据库中**。 执行外部连接；流中的所有记录将被合并（如果数据库中存在相同记录）或添加（如果数据库中尚不存在该记录）。

要映射源数据字段到新数据库列

1. 单击左侧列表中**映射字段**下的源字段名称。
2. 单击**添加**按钮完成映射。

要映射源数据字段到现有数据库列

1. 单击左侧列表中**映射字段**下的源字段名称。
2. 单击右侧**数据库列**下的列名称。
3. 单击**映射**按钮完成映射。

要删除映射

1. 在右侧列表中的“字段”下单击您要移除映射的字段名称。
2. 单击**解除映射**按钮。

在任何列表中取消选择字段

按下 CTRL 键并单击字段名。

数据库导出模式选项

在数据库导出模式对话框中，可以设置数据库导出选项（适用于支持这些选项的数据库），设置字段的 SQL 数据类型，指定主要键字段，并定制导出后生成的 CREATE TABLE 语句。

该对话框由几个部分组成：

- 上面的部分（如显示）包含导出到数据库的选项（适用于支持这些选项的数据库）。 如果没有连接到此类数据库，则不会显示此部分。
- 中间的文本字段显示用于生成 CREATE TABLE 命令的模板，该字段的缺省格式为：

  ```
  CREATE TABLE <table-name> <(table columns)>
  ```

- 下方的表格用于指定每个字段的 SQL 数据类型，并指明如下所述的哪些字段为主要关键字。 对话框将根据表中的规范自动生成 <table-name> 和 <(table columns)> 参数的值。

设置数据库导出选项

如果显示此部分，那么可以指定多种导出到数据库的设置。 支持此功能的数据库类型如下。

- SQL Server Enterprise 和 Developer 版本。 有关更多信息，请参阅主题 第 268 页的『用于 SQL Server 的选项』。
- Oracle Enterprise 或 Personal 版本。 有关更多信息，请参阅主题 第 269 页的『用于 Oracle 的选项』。

自定义 CREATE TABLE 语句

使用此对话框的文本字段部分，可将其他特定于数据库的选项添加到 CREATE TABLE 语句。

1. 请选中**自定义 CREATE TABLE 命令**复选框，以激活文本窗口。
2. 将任意特定于数据库的选项添加到语句中。 请确保保留文本 <table-name> 和 (<table-columns>) 参数，因为 IBM SPSS Modeler 会将这些替换为真实表名和列定义。

设置 SQL 数据类型

缺省情况下，IBM SPSS Modeler 允许数据库服务器自动指定 SQL 数据类型。 要覆盖字段的自动类型，请查找与该字段相对应的行并从模式表的 类型 列的下拉列表中选择所需的类型。 可使用 Shift-单击选择多个行。

对于采用长度、精度或小数位自变量（BINARY、VARBINARY、CHAR、VARCHAR、NUMERIC 和 NUMBER）的类型，您应该指定长度，而不是让数据库服务器分配自动长度。 例如，指定合理的长度值（例如，VARCHAR(25)）可确保将覆盖 IBM SPSS Modeler 中的存储类型（如果这是您的意图）。 要覆盖自动指定的值，请从“类型”下拉列表中选择**指定**，并将类型定义替换为所需的 SQL 类型定义语句。

最简单的方法是先选择最接近所需类型定义的类型，然后选择 **指定** 编辑该定义。 例如，如果要将 SQL 数据类型设置为 VARCHAR(25)，则可以先从“类型”下拉列表中将类型设置为 **VARCHAR(length)**，然后选择**指定**并将文本长度替换为值 25。

主要关键字

如果导出表中的每一行必须对应一列唯一值或多列值组合，则可以通过为每个字段选中适用的 **主要关键字** 复选框来指定。 大多数数据库都不允许对表进行会导致某个主要关键字的约束条件无效的修改，并将自动创建有助于加强此约束的主要关键字索引。 （您可以在“索引”对话框中创建其他字段的索引（可选操作）。 有关更多信息，请参阅主题 第 269 页的『数据库导出索引选项』。）

用于 SQL Server 的选项

使用压缩。 如选中，使用压缩为导出创建表格。

压缩对象。 选择压缩层级。

- **行**。 启用行级别压缩（例如，SQL 中 CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = ROW); 的等效项）。
- **页面**。 启用页面级压缩（例如，SQL 中的 CREATE TABLE MYTABLE(…) WITH (DATA_COMPRESSION = PAGE);）。

用于 Oracle 的选项

Oracle 设置 -“基本”选项

使用压缩。 如选中，使用压缩为导出创建表格。

压缩对象。 选择压缩层级。

- **缺省值**。 启用缺省压缩（例如，SQL 中的 CREATE TABLE MYTABLE(…) COMPRESS;）。 在此情况下，它与**基本**选项的效果相同。
- **基本**。 启用基本压缩（例如，SQL 中的 CREATE TABLE MYTABLE(…) COMPRESS BASIC;）。

Oracle 设置 -“高级”选项

使用压缩。 如选中，使用压缩为导出创建表格。

压缩对象。 选择压缩层级。

- **缺省值**。 启用缺省压缩（例如，SQL 中的 CREATE TABLE MYTABLE(…) COMPRESS;）。 在此情况下，它与**基本**选项的效果相同。
- **基本**。 启用基本压缩（例如，SQL 中的 CREATE TABLE MYTABLE(…) COMPRESS BASIC;）。
- **OLTP**。 启用 OLTP 压缩（例如，SQL 中的 CREATE TABLE MYTABLE(…)COMPRESS FOR OLTP;）。
- **查询低/高**。 （仅限 Exadata 服务器）对查询启用混合列式压缩（例如，SQL 中的 CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY LOW; 或 CREATE TABLE MYTABLE(…)COMPRESS FOR QUERY HIGH;）。 查询压缩非常适合用在数据仓储环境中；HIGH 提供比 LOW 更高的压缩比。
- **归档低/高**。 （仅限 Exadata 服务器）对归档启用混合列式压缩（例如，SQL 中的 CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE LOW; 或 CREATE TABLE MYTABLE(…)COMPRESS FOR ARCHIVE HIGH;）。 归档压缩非常适合用于压缩那些需要长时间存储的数据；HIGH 提供比 LOW 更高的压缩比。

数据库导出索引选项

利用“索引”对话框，可创建从 IBM SPSS Modeler 导出的数据库表的索引。 您可以根据需要指定要包含的字段集，并定制 CREATE INDEX 命令。

该对话框由两部分组成：

- 顶部的文本字段会显示一个模板，可用于生成一个或多个 CREATE INDEX 命令，缺省情况下，该命令遵循以下格式：

  ```
  CREATE INDEX <index-name> ON <table-name>
  ```

- 对话框下方的表用于指定要创建的各个索引。 可指定每个索引的名称以及要包含的字段或列。 该对话框会相应地自动生成 <index-name> 和 <table-name> 参数的值。

  例如，可以为字段 *empid* 和 *deptid* 生成如下所示的单索引 SQL 语句：

  ```
  CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)
  ```

  可以添加多行创建多索引。 为每一行生成单独的 CREATE INDEX 命令。

自定义 CREATE INDEX 命令

（可选）您可以为所有索引或仅特定索引定制 CREATE INDEX 命令。 通过此选项可以灵活地调整设置，以兼容特定数据库要求或选项，还可以根据需要对所有索引或仅对特定的单个索引进行定制。

- 选择对话框上方的**自定义 CREATE INDEX 命令**选项，可修改所有后续添加的索引使用的模板。 请注意，系统不会自动对已添加到表中的索引应用更改。
- 选择表中的一行或多行，然后单击对话框上方的 **更新选定的索引** ，即可将当前自定义应用到所有选定的行。
- 选择每一行的**自定义**复选框可仅修改相应索引的命令模板。

请注意，<index-name> 和 <table-name> 参数的值是由对话框根据表规范自动生成的，无法直接编辑。

BITMAP 关键字。 如果使用的是 Oracle 数据库，那么可以定制模板来创建位图索引，而不是标准索引，方法如下：

```
CREATE BITMAP INDEX <index-name> ON <table-name>
```

位图索引对于创建包含少量不同值的列的索引很有用。 所需 SQL 语句如下所示：

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

UNIQUE 关键字。 大多数数据库都支持在 CREATE INDEX 命令中使用 UNIQUE 关键字。 此关键字将强制执行一个唯一性约束，类似于对基表执行的主要关键字约束。

```
CREATE UNIQUE INDEX <index-name> ON <table-name>
```

请注意，不必对实际指定为主要关键字的字段执行此约束。 大多数的数据库都会自动为 CREATE TABLE 命令中指定为主键字段的任何字段创建索引，因此无需为这些字段显式创建索引。 有关更多信息，请参阅主题第 267 页的『数据库导出模式选项』。

FILLFACTOR 关键字。 可以对某些索引的物理参数进行优调。 例如，利用 SQL Server，用户可以根据日后更改表所产生的维护成本来选用索引大小（首次创建后）。

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

备注

- 如果已存在指定名称的索引，那么索引创建将失败。 任何故障最初都将被当作警告处理，以便创建后续索引，然后在所有索引都已尝试后在消息日志中重新报告为错误。
- 为获得最佳性能，应在将数据加载到表中后创建索引。 索引必须至少包含一列。
- 在执行节点前，可以在消息日志中预览生成的 SQL。
- 对于写入数据库的临时表（即启用节点高速缓存时），指定主要关键字和索引的选项不可用。 但是，系统可以根据数据在下游节点中的使用方式在临时表中创建相应的索引。 例如，如果逐个将高速缓存数据添加到 *DEPT* 列，则有必要为此列高速缓存表创建索引。

索引和查询优化

在某些数据库管理系统中，对数据库表进行创建、加载和建立索引操作之后，还需要执行进一步的操作，优化程序才能利用索引来提高对新表的查询速度。 例如，在 Oracle 中，基于成本的查询优化程序要求先对表进行分析，然后才能在查询优化系统中使用表的索引。 Oracle 的内部 ODBC 属性文件（用户不可见）包含导致发生此情形的选项，显示如下：

```
# 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'
```

在 Oracle 中创建表（无论是否定义主要关键字或索引）时，都会执行此步骤。 如有必要，可以类似方法自定义其他数据库的 ODBC 属性文件 - 请联系支持获得帮助。

数据库导出高级选项

单击数据库导出节点对话框的“高级”按钮时，将显示一个新的对话框，可在其中指定将结果导出到数据库的具体方法。

使用批量提交。 选中此选项可关闭到数据库的逐行提交。

批量大小。 指定在提交到内存前发送到数据库的记录数。 此数值越低时，数据完整性越高，但会降低数据传输速度。 可优化调整该数值，以实现最佳的数据库性能。

使用批量加载。 指定一种将数据从 IBM SPSS Modeler 直接批量加载到数据库的方法。 可能需要进行多次尝试以为特定场景选择合适的批量加载选项。

- **通过 ODBC**。 选择此项可使用 ODBC API 执行多行插入，这比正常导出到数据库的速度更快。 请从以下选项选择逐行绑定或逐列绑定。
- **通过外部加载程序**。 选中此项可使用特定于数据库的定制批量加载程序。 选中此项将激活下面各个选项。

高级 ODBC 选项。 这些选项只有在选中**通过 ODBC** 时才可用。 注意：只有部分 ODBC 驱动程序支持此功能。

- **逐行**。 选择逐行绑定可使用 SQLBulkOperations 调用将数据加载到数据库。 通常，与逐条插入数据的参数化插入相比，逐行绑定的速度更快。
- **逐列**。 选中此项可使用逐列绑定将数据加载到数据库。 逐列绑定通过将每个数据库列（在参数化 INSERT 语句中）绑定到 *N* 个值的数组，提高性能。 执行 INSERT 语句一次可将 *N* 行插入数据库中。 此方法可以大大提高处理速度。

外部加载程序选项。 指定**通过外部加载程序**时，将显示各种选项，这些选项可用于将数据集导出到文件，然后指定并执行定制加载程序以将数据从该文件加载到数据库。 IBM SPSS Modeler 可连接到许多常用数据库系统的外部加载程序接口。 该软件随附多个脚本，这些脚本与技术文档一起位于 scripts 子目录下。 请注意，要使用此功能，必须将 Python 2.7 安装在 IBM SPSS Modeler 或 IBM SPSS Modeler Server 所在的机器上，并且必须在 options.cfg 文件中设置 python_exe_path 参数。 有关更多信息，请参阅主题 第 271 页的『批量加载程序设计』。

- **使用定界符**。 指定已导出文件中应使用的定界符。 选择 **制表符** 将以制表符定界，选择 **空格** 则以空格定界。 选择 **其他** 可指定其他符号，比如半角逗号 (,)。
- **指定数据文件**。 选中此项可输入批量加载时写入数据文件所用的路径。 缺省情况下，将在服务器的临时目录中创建一个临时文件。
- **指定加载程序**。 选中此项可指定一个批量加载程序。 缺省情况下，该软件会搜索 IBM SPSS Modeler 安装程序的 scripts 子目录，查找给定数据库要执行的 Python 脚本。 该软件随附多个脚本，这些脚本与技术文档一起位于 scripts 子目录下。
- **生成日志**。 选中此项可在指定目录下生成日志文件。 日志文件包含的错误信息在批量加载操作失败时可派上用场。
- **检查表大小**。 选中此项可执行表检查，以确保表大小会随从 IBM SPSS Modeler 导出行数的增多而相应地增大。
- **额外的加载程序选项**。 指定加载程序的其他参数。 对于含有空格的参数，请使用双引号。

双引号包含在可选参数中，并使用反斜杠进行转义。 例如，指定为 -comment “This is a \” comment\"" 的选项包含 -comment 标志和注释自身（呈现为 This is a “comment”）。

还可包含单个反斜杠，并通过使用另一个反斜杠进行转义。 例如，指定为 -specialdir “C:\\Test Scripts\\” 的选项包含标志 -specialdir 和目录（呈现为 C:\Test Scripts\）。

批量加载程序设计

数据库导出节点在“高级选项”对话框中提供了用于批量加载的选项。 批量加载程序可以将数据从文本文件加载到数据库。

选项**使用批量加载 - 通过外部加载程序**可通过配置 IBM SPSS Modeler 来完成以下三项操作：

- 创建任何需要的数据库表。
- 将数据导出到文本文件。
- 激活批量加载程序，以将数据从此文件加载到数据库表。

一般而言，批量加载程序自身不是数据库加载实用程序（例如，Oracle 的 sqlldr 实用程序），而是一个小型脚本或程序，该脚本或程序可以构造正确的参数，创建任何特定于数据库的辅助文件（比如控制文件），然后激活数据库加载实用程序。 以下各节中的信息将帮助您编辑现有批量加载程序。

另外，您也可以编写自己的程序以用于批量加载。 有关更多信息，请参阅主题 第 274 页的『开发批量加载程序』。 请注意，标准的技术支持协议未涵盖此内容，您应该与 IBM 服务代表联系以获取帮助。

批量加载脚本

IBM SPSS Modeler 附带了大量的批量加载程序，它们适合不同的数据库，并使用 Python 脚本实现。 当运行包含选择了**通过外部加载程序**选项的数据库导出节点的流时，IBM SPSS Modeler 将通过 ODBC 来创建数据库表（如需要），并将数据导出至运行 IBM SPSS Modeler Server 的主机上的临时文件，然后调用批量加载脚本。 此脚本进而执行 DBMS 供应商提供的实用程序，以便将数据从临时文件导入至数据库中。

注释:

- IBM SPSS Modeler 安装不包含 Python 运行时解释器，因此需要单独安装 Python 。 有关更多信息，请参阅主题 第 270 页的『数据库导出高级选项』。
- 针对下表中列出的数据库提供了脚本 (在 IBM SPSS Modeler 安装目录的 `\scripts` 文件夹中)。
- 目前， IBM SPSS Modeler 提供的批量加载程序脚本不支持 LDAP。

表 *47:* 提供的批量加载程序脚本

数据库	脚本名称	进一步信息
IBM Db2	`db2_loader.py`	有关更多信息，请参阅主题 第 272 页的『向 IBM Db2 数据库批量加载数据』。
IBM Netezza	`netezza_loader.py`	有关更多信息，请参阅主题 第 273 页的『向 IBM Netezza 数据库批量加载数据』。
Oracle	`oracle_loader.py`	有关更多信息，请参阅主题 第 273 页的『向 Oracle 数据库批量加载数据』。
SQL Server	`mssql_loader.py`	有关更多信息，请参阅主题 第 274 页的『向 SQL Server 数据库批量加载数据』。
Teradata	`teradata_loader.py`	有关更多信息，请参阅主题 第 274 页的『向 Teradata 数据库批量加载数据』。

向 IBM Db2 数据库批量加载数据

以下几点说明如何通过使用“数据库导出高级选项”对话框中的“外部加载程序”选项来配置从 IBM SPSS Modeler 到 IBM Db2 数据库的批量加载。

确保安装了 Db2 命令行处理器 (CLP) 实用程序

脚本 `db2_loader.py` 调用 Db2 LOAD 命令。 请确保命令行处理器（在 UNIX 上为 db2，在 Windows 上为 `db2cmd`）安装在要执行 `db2_loader.py` 的服务器上（通常是运行 IBM SPSS Modeler Server 的主机）。

检查本地数据库别名是否与实际数据库名称相同

Db2 本地数据库别名是 Db2 客户端软件所使用的名称，在本地或远程 Db2 实例中，此名称指代数据库。 如果本地数据库别名与远程数据库的名称不同，请提供其他加载程序选项:

```
-alias <local_database_alias>
```

例如，在主机 GALAXY 上的远程数据库名为 STARS，但在运行 IBM SPSS Modeler Server 的主机上 Db2 本地数据库别名为 STARS_GALAXY。 使用其他加载程序选项

```
-alias STARS_GALAXY
```

非 ASCII 字符数据编码

如果要批量加载非 ASCII 格式的数据，应确保在 `db2_loader.py` 配置部分中的代码页变量在系统中正确设置。

空字符串

空字符串将作为 NULL 值导出到数据库。

向 IBM Netezza 数据库批量加载数据

以下几点说明如何通过使用“数据库导出高级选项”对话框中的“外部加载程序”选项来配置从 IBM SPSS Modeler 到 IBM Netezza 数据库的批量加载。

确保安装了 Netezza nzload 实用程序

脚本 *netezza_loader.py* 调用 Netezza 实用程序 *nzload*。 确保在要执行 *netezza_loader.py* 的服务器上安装并正确配置 *nzload*。

导出非 ASCII 数据

如果导出中包含非 ASCII 格式的数据，那么可能需要向“数据库导出高级选项”对话框的**其他加载程序选项**字段添加 `-encoding UTF8`。 这将确保正确上载非 ASCII 数据。

日期、时间和时间戳记格式数据

在流属性中，将日期格式设为 **DD-MM-YYYY**，并将时间格式设为 **HH:MM:SS**。

空字符串

空字符串将作为 NULL 值导出到数据库。

在现有表中插入数据时，流与目标表中的列顺序不同

如果流中的列顺序与目标表不同，那么数据值将被插入错误的列中。 使用 Field Reorder 节点可确保流中的列顺序与目标表中的列顺序匹配。 有关更多信息，请参阅主题 第 136 页的『字段重排节点』。

跟踪 nzload 进度

在本地方式下运行 IBM SPSS Modeler 时，向“数据库导出高级选项”对话框的**其他加载程序选项**字段添加 `-sts`，即可在 *nzload* 实用程序打开的命令窗口中查看每 1000 行的状态信息。

向 Oracle 数据库批量加载数据

以下几点说明如何通过使用“数据库导出高级选项”对话框中的“外部加载程序”选项来配置从 IBM SPSS Modeler 到 Oracle 数据库的批量加载。

确保安装了 Oracle 实用程序

脚本 *oracle_loader.py* 调用 Oracle 实用程序 *sqlldr*。 请注意，*sqlldr* 未自动包含在 Oracle 客户端中。 确保在要执行 *oracle_loader.py* 的服务器上安装 *sqlldr*。

指定数据库 SID 或服务名称

如果要将数据导出至非本地 Oracle 服务器，或者本地 Oracle 服务器有多个数据库，则需要在“数据库导出高级选项”对话框的**其他加载程序选项**字段中指定以下选项，以传入 SID 或服务名称：

```
-database <SID>
```

编辑 oracle_loader.py 的配置部分

在 UNIX（或 Windows）系统上，编辑 *oracle_loader.py* 脚本开始处的配置部分。 在这里，可根据情况指定 ORACLE_SID、NLS_LANG、TNS_ADMIN 和 ORACLE_HOME 环境变量值，以及 *sqlldr* 实用程序的完整路径。

日期、时间和时间戳记格式数据

在流属性中，通常应将日期格式设为 **YYYY-MM-DD**，并将时间格式设为 **HH:MM:SS**。

如果要使用其他日期与时间格式，请参阅 Oracle 文档并编辑 *oracle_loader.py* 脚本文件。

非 ASCII 字符数据编码

如果要批量加载非 ASCII 格式的数据，那么应该确保在系统中正确地设置了环境变量 NLS_LANG。 这将由 Oracle 加载实用程序 *sqlldr* 来读取。 例如，对于 Windows，Shift-JIS 的 NLS_LANG 的正确值是 Japanese_Japan.JA16SJIS。 有关 NLS_LANG 的更多详细信息，请查阅 Oracle 文档。

空字符串

空字符串将作为 NULL 值导出到数据库。

向 SQL Server 数据库批量加载数据

以下几点说明如何通过使用“数据库导出高级选项”对话框中的“外部加载程序”选项来配置从 IBM SPSS Modeler 到 SQL Server 数据库的批量加载。

确保安装了 SQL Server bcp.exe 实用程序
脚本 *mssql_loader.py* 调用 SQL Server 实用程序 *bcp.exe*。 确保在要执行 *mssql_loader.py* 的服务器上安装 *bcp.exe*。

不得使用使用空格作为分隔符
请避免在“数据库导出高级选项”对话框中选择空格作为分隔符。

建议的“检查表大小”选项
建议在“数据库导出高级选项”对话框中启用**检查表大小**选项。 系统并不是总能检测到批量加载过程中的故障，启用此选项将执行额外的检查，以确保加载正确数目的行。

空字符串
空字符串将作为 NULL 值导出到数据库。

指定标准 SQL Server 命名实例
在某些情况下，SPSS Modeler 可能会由于主机名未限定而无法访问 SQL Server 并显示以下错误：

执行外部批量加载程序时发生错误。 日志文件可能提供了更多详细信息。

要更正此错误，请向**其他加载程序选项**字段添加以下字符串（包括双引号）：

```
"-S mhreboot.spss.com\SQLEXPRESS"
```

向 Teradata 数据库批量加载数据

以下几点说明如何通过使用“数据库导出高级选项”对话框中的“外部加载程序”选项来配置从 IBM SPSS Modeler 到 Teradata 数据库的批量加载。

确保安装了 Teradata fastload 实用程序

脚本 *teradata_loader.py* 调用 Teradata 实用程序 *fastload*。 确保在要执行 *teradata_loader.py* 的服务器上安装并正确配置 *fastload*。

数据只能批量加载到空表中

对于批量加载，只能使用空表作为目标。 如果目标表包含批量加载之前的任何数据，那么操作将失败。

日期、时间和时间戳记格式数据

在流属性中，将日期格式设为 **YYYY-MM-DD**，并将时间格式设为 **HH:MM:SS**。

空字符串

空字符串将作为 NULL 值导出到数据库。

Teradata 进程标识 (tdpid)

缺省情况下，*fastload* 会将数据导出至 `tdpid=dbc` 的 Teradata 系统中。 通常，在 HOSTS 文件中会有一个条目，用于将 `dbccop1` 与 Teradata 服务器的 IP 地址进行关联。 要使用其他服务器，请在“数据库导出高级选项”对话框的**其他加载程序选项**字段中指定以下选项以遍历此服务器的 `tdpid`：

```
-tdpid <id>
```

表名和列名中的空格

如果表名或列名包含空格。那么批量加载操作将失败。 如有可能，请重新命名表或列以移除空格。

开发批量加载程序

该主题说明如何开发可在 IBM SPSS Modeler 中运行的批量加载程序，以便将文本文件数据加载到数据库中。 请注意，标准的技术支持协议未涵盖此内容，您应该与 IBM 服务代表联系以获取帮助。

使用 Python 构建批量加载程序

缺省情况下，IBM SPSS Modeler 会根据数据库类型搜索缺省的批量加载程序。 请参阅 第 272 页的表 47。

可使用脚本 *test_loader.py* 来协助开发批量加载程序。 有关更多信息，请参阅主题 第 276 页的『测试批量加载程序』。

传递给批量加载程序的对象

IBM SPSS Modeler 写入两个将传递给批量加载程序的文件。

- **数据文件**。 该文件为文本格式，包含要加载的数据。
- **模式文件**。 该文件为 XML 文件，描述列名与类型，并提供数据文件的格式信息（例如，用作字段间分隔符的字符）。

此外，IBM SPSS Modeler 还将传递其他信息（例如，表名、用户名和密码等）作为调用批量加载程序时的参数。

注意：为了向 IBM SPSS Modeler 表明加载操作完成，批量加载程序将删除模式文件。

传递给批量加载程序的参数

下表中列出了遍历给程序的自变量。

表 *48:* 传递给批量加载程序的参数

自变量	**描述**
schemafile	模式文件的路径。
data file	数据文件的路径。
servername	DBMS 服务器的名称；可以为空。
databasename	DBMS 服务器内的数据库名称；可以为空。
username	用于登录数据库的用户名。
password	用于登录数据库的密码。
tablename	要加载的表的名称。
ownername	表所有者的名称（也称为模式名称）。
logfilename	日志文件的名称（如果为空，那么不会创建日志文件）。
rowcount	数据集中的行数。

遍历这些标准参数之后，将向批量加载程序遍历在“数据库导出高级选项”的**其他加载程序选项**字段中指定的任何选项。

数据文件格式

数据将以文本格式写入数据文件，每个字段之间以在“数据库导出高级选项”上指定的分隔符进行分隔。 下面是制表符分隔的数据文件的可能显示方式的示例。

```
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
```

采用 IBM SPSS Modeler Server（或 IBM SPSS Modeler，如未连接到 IBM SPSS Modeler Server）所用的本地编码来写入该文件。 某些格式通过 IBM SPSS Modeler 流设置来控制。

模式文件格式

模式文件是用于描述数据文件的 XML 文件。 下面是前述数据文件附带的一个示例。

```
<?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>
```

下面的两个表列出了模式文件的 <table> 和 <column> 元素的属性。

表 *49: <table>* 元素的属性

属性	**描述**
delimiter	字段分隔符（TAB 将表示为 \t）。
commit_every	批量大小间隔（与在“数据库导出高级选项”对话框上相同）。
date_format	用于表示日期的格式。
time_format	用于表示时间的格式。
append_existing	true（如果要加载的表已包含数据）；否则，值为 false。
delete_datafile	true（如果批量加载程序在完成加载时应删除数据文件）。

表 *50: <column>* 元素的属性

属性	**描述**
name	列名称。
encoded_name	列名称已转换为与数据文件相同的编码，并输出为一系列两位十六进制数字。
type	列的数据类型，值为下列其中一个：integer、real、char、time、date 和 datetime。
size	对于 char 数据类型，这是列的最大宽度（字符数）。

测试批量加载程序

可使用位于 IBM SPSS Modeler 安装目录的 *\scripts* 文件夹下的测试脚本 *test_loader.py* 对批量加载进行测试。 在尝试开发、调试用于 IBM SPSS Modeler 的批量加载程度或脚本以及对其进行故障诊断时，此操作非常有用。

要使用测试脚本，请继续以下操作。

1. 运行 *test_loader.py* 脚本，将模式与数据文件复制到文件 *schema.xml* 与 *data.txt*，并创建 Windows 批处理文件 (*test.bat*)。
2. 编辑 *test.bat* 文件以选择要测试的批量加载程序或脚本。
3. 从命令行运行 *test.bat* 以测试选定的批量加载脚本或脚本。

注意：实际上，运行 *test.bat* 不会将数据加载数据库。

平面文件导出节点

平面文件导出节点可用于将数据写入定界文本文件。 特别适用于导出其他分析或电子表格软件可以读取的数据。

如果数据包含地理空间信息，那么您可以将其导出为平面文件；如果生成了“变量文件”源节点以便在同一个流中使用，那么所有的存储、测量和地理空间元数据都将保留在新的源节点中。 但是，如果您导出该数据，然后在另一个流中将其导入，那么您必须执行一些额外的步骤以便在新的源节点中设置地理空间元数据。 有关更多信息，请参阅主题 第 22 页的『“变量文件”节点』。

注: 无法以过往的缓存格式写文件，这是因为 IBM SPSS Modeler 已不再使用该缓存文件格式。 IBM SPSS Modeler 缓存文件现以 IBM SPSS Statistics *.sav* 格式保存，您可以通过 统计信息 导出节点写入文件。 有关更多信息，请参阅主题 第 277 页的『Statistics 导出节点』。

“平面文件导出”选项卡

导出文件。 指定文件名。 输入文件名，或单击“文件选择器”按钮浏览文件位置。

写方式。 如果选择了**覆盖**，那么将覆盖指定文件中的任何现有数据。 如果选择**追加**，则输出将被添加到现有文件的末尾，同时保留该文件包含的所有数据。

- **包含字段名**。 如果选中此选项，那么字段名称将写入到输出文件的第一行。 此选项只适用于 **覆盖** 写入模式。

每条记录后新建一行。 如果选中此选项，那么每条记录都将对应写入到输出文件的新的一行中。

字段分隔符。 指定要在生成的文本文件中的字段值之间插入的字符。 选项有 **逗号** 、 **制表符** 、 **空格** 和 **其他** 。 如果选择 **其他** ，则在文本框中输入所需的定界符。

符号引号。 指定用于为符号字段值添加引号的方式。 选项有 **无** （不为值添加引号）、 **单引号 (')** 、 **双引号 (")** 和 **其他** 。 如果选择 **其他** ，则在文本框中输入所需的引号。

编码。 指定使用的文本编码方法。 您可以选择系统缺省值、流缺省值或 UTF-8。

- 系统缺省值在 Windows 控制面板中指定，如果以分布式模式运行，则在服务器计算机上指定。
- 流缺省值在“流属性”对话框中指定。

十进制符号。 指定小数在数据中的表示方法。

- **流缺省值**。 将使用当前流缺省设置定义的小数分隔符。 通常为计算机语言环境中定义的小数分隔符。
- **句点 (.)**。 句点字符将用作十进制分隔符。
- **逗号 (,)**。 逗号字符将用作十进制分隔符。

为此数据生成导入节点。 选中此选项可自动生成将读取已导出数据文件的“变量文件”源节点。 有关更多信息，请参阅主题 第 22 页的『“变量文件”节点』。

Statistics 导出节点

“统计信息导出”节点使您能够以 IBM SPSS Statistics *.sav* 格式导出数据。IBM SPSS Statistics *.sav* 文件可由 IBM SPSS Statistics Base 和其他模块读取。 这也是 IBM SPSS Modeler 缓存文件所使用的格式。

将 IBM SPSS Modeler 字段名称映射到 IBM SPSS Statistics 变量名称某些时候可能导致错误，因为 IBM SPSS Statistics 变量名称限制为 64 个字符，而且不能包含特定字符，例如，空格、美元符号 ($) 和连字符 (–)。可通过两种方式针对以下限制进行调整:

- 可通过单击“过滤”选项卡，根据 IBM SPSS Statistics 变量名要求重命名字段。 有关更多信息，请参阅主题 第 278 页的『重命名或过滤 IBM SPSS Statistics 的字段』。
- 选择同时从 IBM SPSS Modeler 导出字段名和标签。

注意：IBM SPSS Modeler 以 Unicode UTF-8 格式写入 *.sav* 文件。 IBM SPSS Statistics 16.0 版以后的版本只支持 Unicode UTF-8 格式的文件。 为了防止数据可能损坏，使用 Unicode 编码保存的 *.sav* 文件不得用于 16.0 之前的 IBM SPSS Statistics 版本。 有关详细信息，请参阅 IBM SPSS Statistics 帮助。

多重响应集。 导出文件后，将自动保留流中定义的任何多重响应集。 借助“过滤器”选项卡，您可以查看和编辑任意节点的多重响应集。 有关更多信息，请参阅主题 第 114 页的『编辑多重响应集』。

Statistics 导出节点 - “导出”选项卡

导出文件 指定文件的名称。 输入文件名，或单击“文件选择器”按钮浏览文件位置。

文件类型 请选择是以正常的 *.sav* 格式还是压缩的 *.zsav* 格式保存文件。

使用密码对文件加密 要使用密码保护文件，请选中此框；系统将提示您在另一对话框中输入并确认**密码**。

注: 受密码保护的文件只能由 SPSS Modeler V16 或更高版本，或者由 SPSS 统计信息 V21 或更高版本打开。

导出字段名称 指定将变量名称和标签从 SPSS Modeler 导出至 SPSS 统计信息 *.sav* 或 *.zsav* 文件时的处理方法。

- **名称和变量标签** 选择此选项可同时导出 SPSS Modeler 字段名称和字段标签。 名称将以 SPSS 统计信息 变量名方式导出，而标签则以 SPSS 统计信息 变量标签方式导出。
- **作为变量标签的名称** 选中此选项可将 SPSS Modeler 字段名称用作 SPSS 统计信息 中的变量标签。 SPSS Modeler 允许在字段名称中包含不适用于 SPSS 统计信息 变量名称的字符。 为了防止创建无效的 SPSS 统计信息 名称，请选择**将名称用作变量标签**或使用“过滤”选项卡调整字段名。

启动应用程序 如果计算机上安装了 SPSS 统计信息，则可以选择此选项对已保存数据文件直接激活应用程序。 “帮助应用程序”对话框中必须指定用于启动应用程序的选项。 有关更多信息，请参阅主题 第 262 页的『IBM SPSS Statistics 帮助应用程序』。 要在不打开外部程序的情况下直接创建 SPSS 统计信息 *.sav* 或 *.zsav* 文件，请取消选择此选项。

注: 当以服务器（分布式）方式同时运行 SPSS Modeler 和 SPSS 统计信息 时，写出数据并启动 SPSS 统计信息 会话不会自动打开 SPSS 统计信息 客户端以显示读入活动数据集的数据集。 变通方法是 SPSS 统计信息 客户端启动时立刻在其中打开数据文件。

针对此数据生成一个导入节点 选中此选项可自动生成将读取已导出数据文件的“Statistics 文件”源节点。 有关更多信息，请参阅主题 第 26 页的『Statistics 文件节点』。

重命名或过滤 IBM SPSS Statistics 的字段

将数据从 IBM SPSS Modeler 导出或部署到外部应用程序（比如 IBM SPSS Statistics）之前，可能需要重命名或调整字段名。 “统计信息变换”、“统计信息输出”和“统计信息导出”对话框提供了一个“过滤”选项卡，用于帮助执行此过程。

“过滤”选项卡基本功能将在后文介绍。 有关更多信息，请参阅主题 第 113 页的『设置过滤选项』。 本主题提供将数据读入 IBM SPSS Statistics 的提示。

要调整字段名称以符合 IBM SPSS Statistics 命名规则：

1. 在“过滤”选项卡上，单击“过滤选项菜单”工具栏按钮（工具栏上的第一个按钮）。
2. 选择“为 IBM SPSS Statistics 重命名”。
3. 在“为 IBM SPSS Statistics 重命名”对话框中，您可以选择使用**井号 (#)** 字符或**下划线 (_)** 替换文件名中的无效字符。

重命名多重响应集。 如果您要调整多重响应集（可通过 Statistics 文件源节点导入 IBM SPSS Modeler）的名称，请选择此项。 多重响应集可用于记录对每个问题都具有多个值的数据，例如，在调查响应时。

数据收集 导出节点

数据收集 导出节点基于 数据收集 数据模型，以 数据收集 市场调查软件中使用的格式保存数据。 此格式可以将观测值数据（对调查期间所收集的问题的实际响应）与元数据（描述如何收集和组织观测值数据）进行区分。 元数据包括以下信息，例如问题文本、变量名称和说明、多重响应集、不同文本的转换以及观测值数据结构的定义。 有关更多信息，请参阅主题 第 26 页的『数据收集 节点』。

元数据文件。 指定调查表定义文件的名称 (*.mdd*)，将在该文件中保存已导出的元数据。 缺省的调查表将基于字段类型信息创建。 例如，名义（集合）字段可以表示为单个问题，其中，将字段说明用作问题文本，并为每个已定义的值采用单独复选框。

合并元数据。 指定元数据是将覆盖现有版本还是与现有元数据合并。 如果选择了合并选项，那么将在每次运行流时创建新版本。 由此可以在调查表发生更改时跟踪它的各种版本。 每个版本都可看作是用于收集观测值数据特定集合的元数据的一个快照。

启用系统变量。 指定系统变量是否包含在导出的 .mdd 文件中。 这些变量中包括 *Respondent.Serial* 、*Respondent.Origin* 和 *DataCollection.StartTime* 等。

观测值数据设置。 指定从其中导出观测值数据的 IBM SPSS Statistics 数据 (*.sav*) 文件。 请注意，所有对变量和值名称的限制均适用于此处，例如您可能需要切换至“过滤”选项卡，并使用“过滤选项”菜单上的“为 IBM SPSS Statistics 重命名”选项，以更正字段名称中的无效字符。

为此数据生成导入节点。 选择此选项可自动生成将读取已导出数据文件的数据收集源节点。

多重响应集。 导出文件后，将自动保留流中定义的任何多重响应集。 借助“过滤器”选项卡，您可以查看和编辑任意节点的多重响应集。 有关更多信息，请参阅主题 第 114 页的『编辑多重响应集』。

Analytic Server 导出节点

Analytic Server 的“导出”节点使您能够将来自分析的数据写入现有 Analytic Server 数据源。 例如，此数据源可以是 Hadoop 分布式文件系统 (HDFS) 或数据库中的文本文件。

通常，具有 Analytic Server 导出节点的流还会以 Analytic Server 源节点开头，并被提交到 Analytic Server，然后在 HDFS 上执行。 另外，具有“本地”数据源的流可以 Analytic Server 导出节点结束，以便上载相对较小的数据集（不超过 100,000 条记录）与 Analytic Server 配合使用。

如果要使用您自己的 Analytic Server 连接而不是管理员定义的缺省连接，请取消选择**使用缺省 Analytic Server**，并选择您的连接。

数据源。 选择包含要使用的数据的数据源。 数据源包含与该源关联的文件和元数据。 单击**选择**以显示可用数据源的列表。 有关更多信息，请参阅主题 第 9 页的『选择数据源』。

如果您需要创建新数据源或编辑现有数据源，请单击**启动数据源编辑器 ...**。

方式。 选择**追加**表示添加到现有数据源，选择**覆盖**表示替换数据源的内容。

为此数据生成“导入”节点。 选择此选项可在数据导出至指定数据源时生成此数据的源节点。 这会将此节点添加到流画布中。

请注意，使用多个 Analytic Server 连接在控制数据流方面十分有用。 例如，使用 Analytic Server 源节点和导出节点时，您可能希望在流的不同分支中使用不同的 Analytic Server 连接，以便在每个分支运行时，使用自己的 Analytic Server，并且不会将任何数据拉取到 IBM SPSS Modeler Server。 请注意，如果某个分支包含多个 Analytic Server 连接，那么会将数据从 Analytic Server 拉取到 IBM SPSS Modeler Server。

IBM Cognos 导出节点

IBM Cognos 导出节点允许您以 UTF-8 格式，将数据从 IBM SPSS Modeler 流导出到 Cognos Analytics。 如此一来，Cognos 就可以利用来自 IBM SPSS Modeler 的已进行转换或评分的数据。 例如，您可使用 Cognos Report Studio 创建基于所导出数据（包括预测和置信度值）的报告。 然后，可将该报告保存在 Cognos 服务器上，并分发给 Cognos 用户。

注: 只能导出关系数据，无法导出 OLAP 数据。

要将数据导出到 Cognos，您需指定以下信息：

- Cognos 连接 - 到 Cognos Analytics 服务器的连接
- ODBC 连接 - 与 Cognos 服务器所使用的 Cognos 数据服务器的连接

在 Cognos 连接中指定要使用的 Cognos 数据源。 该数据源必须使用与 ODBC 数据源相同的登录信息。

将实际流数据导出到数据服务器，同时将数据包元数据导出到 Cognos 服务器。

就像任何其他导出节点一样，您还可使用节点对话框的“发布”选项卡，发布流以便使用 IBM SPSS Modeler Solution Publisher 进行部署。

注: Cognos 源节点仅支持 Cognos CQM 数据包。 不支持 DQM 数据包。

Cognos 连接

您可以在此处指定与要用于导出的 Cognos Analytics 服务器的连接。 此程序涉及将元数据导出到 Cognos 服务器上的新数据包，同时将流数据导出到 Cognos 数据服务器。

连接。 单击**编辑**按钮会显示一个对话框，您可在该对话框中定义数据所要导出到的 Cognos 服务器的 URL 及其他详细信息。 如果您已通过 IBM SPSS Modeler 登录 Cognos 服务器，那么还可以编辑当前连接的详细信息。 有关更多信息，请参阅第 32 页的『Cognos 连接』。

数据源。 这是将数据导出到的 Cognos 数据源（通常为数据库）的名称。 下拉列表显示您可从当前连接访问的所有 Cognos 数据源。 单击**刷新**按钮以更新此列表。

文件夹。 Cognos 服务器上的文件夹路径及名称，导出数据包将在该文件夹中创建。

数据包名称。 要包含导出元数据的指定文件夹中的数据包名称。 这必须是一个带有单独查询主体的新数据包；不能导出到现有数据包。

方式。 指定您希望如何执行导出：

- **立即发布数据包**。 （缺省值）单击**运行**时立即执行导出操作。
- **导出操作脚本**。 创建一个可稍后运行的 XML 脚本（例如使用 Framework Manager）以执行导出。 在**文件**字段中的脚本键入路径和文件名，或使用**编辑**按钮指定脚本文件的名称和位置。

为此数据生成导入节点。 选中此项可在将数据导出到指定数据源和表时生成此数据的源节点。 单击**运行**时，此节点即被添加到流工作区。

ODBC 连接

可在此处指定将要导出流数据至的 Cognos 数据服务器（即数据库）的连接。

注: 您必须确保此处指定的数据源与 **Cognos 连接**面板上指定的数据源相同。 您还必须确保 Cognos 连接数据源使用与 ODBC 数据源相同的登录信息。

数据源。 显示所选数据源。 输入名称或者从下拉列表中选择名称。 如果列表中未显示所需的数据库，则选择**添加新的数据库连接**并从“数据库连接”对话框选定数据库。 有关更多信息，请参阅 第 14 页的『添加数据库连接』。

表名称。 输入接收数据的表名称。 如果选择 **插入到表中** 选项，

- 您可以通过单击 **选择** 按钮来选择数据库中的现有表。
- 如果提供当前不存在的表名，那么将创建具有指定名称的新表，并将数据插入其中。

创建表。 选中此项可创建一个新的数据库表或覆盖现有的数据库表。

插入表中。 选择此选项以便：

- 将数据插入现有数据库表中的新行，或者
- 将数据插入到不存在的表中。 在这种情况下，将创建新表，并将数据作为新创建的表中的新行插入。

合并表。 （如可用）选择此选项以使用相应源数据字段中的值更新所选数据库列。 选择此选项会启用**合并**按钮，在其显示的对话框中您可以将源数据字段映射到数据库列。

删除现有表。 选中此项可在创建新表时删除所有名称相同的现有表。

删除现有行。 选择此选项可在插入表时先将现有行从表中删除然后导出。

注: 如果选择上述任意两个选项，那么在执行节点时将收到**覆盖警告**消息。 要想不显示此警告，请取消选择“用户选项”对话框的“通知”选项卡上的**当节点覆盖数据表时发出警告**选项。

缺省字符串大小。 上游“类型”节点中标记为无类型的字段将作为字符串字段写入数据库。 请指定无类型字段要使用的字符串大小。

单击**模式**可打开一个对话框，您可在其中设置各种导出选项（对于支持此功能的数据库）、设置所需字段的 SQL 数据类型，并指定创建数据库索引的主要关键字。 有关更多信息，请参阅 第 267 页的『数据库导出模式选项』。

单击 **索引** 可指定创建导出表索引的选项，以提高数据库的运行性能。 有关更多信息，请参阅 第 269 页的『数据库导出索引选项』。

单击 **高级** 可指定批量加载和数据库提交选项。 有关更多信息，请参阅 第 270 页的『数据库导出高级选项』。

将表和列名加上引号。 选择将 CREATE TABLE 语句发送到数据库时使用的选项。 必须为包含空格和非标准字符的表和列添加引号。

- **根据需要**。 选择此选项，以允许 IBM SPSS Modeler 自动根据各个情况确定是否需要添加引号。
- **始终**。 选中此项将始终为表名和列名称添加引号。
- **永不**。 选中此项将禁用引号。

为此数据生成导入节点。 选中此项可在将数据导出到指定数据源和表时生成此数据的源节点。 单击**运行**时，此节点即被添加到流工作区。

IBM Cognos TM1 导出节点

通过 IBM Cognos 导出节点，您可以将数据从 SPSS Modeler 流导出至 Cognos TM1。 这样，Cognos Analytics 可利用来自 SPSS Modeler 的转换或评分数据。

注: 您只能导出测量值而不能导出上下文维度数据；或者，您可以向多维数据集添加新元素。

要将数据导出到 Cognos Analytics，需要指定以下各项：

- 与 Cognos TM1 服务器的连接。
- 数据所要导出到的多维数据集。
- 从 SPSS 数据名到对等 TM1 维度和测量值的映射。

注: TM1 用户需要下列许可权：对多维数据集的写特权、对维度的读特权以及对维度元素的写特权。 此外，需要 IBM Cognos TM1 10.2 FP3 或更高版本，SPSS Modeler 才能导入和导出 Cognos TM1 数据。 基于先前版本的现有流仍能够正常运行。

该节点不需要管理员凭证。 但如果仍要使用 17.1 之前的旧的遗存 TM1 节点，仍需要管理员凭证。

SPSS Modeler 仅支持通过 IntegratedSecurityMode 1、4 和 5 使用 Cognos TM1。

就像任何其他导出节点一样，您还可使用节点对话框的“发布”选项卡，发布流以便使用 IBM SPSS Modeler Solution Publisher 进行部署。

注: 必须先验证 `tm1s.cfg` 文件中的某些设置，然后才能在 SPSS Modeler 中使用 TM1“源”或“导出”节点；这是 TM1 服务器根目录中的 TM1 服务器配置文件。

- HTTPPortNumber - 设置有效的端口号；通常介于 1 到 65535 之间。 请注意，这不是后续在节点中的连接内指定的端口号；这是 TM1 使用的内部端口，缺省情况下处于禁用状态。 如果需要，请联系 TM1 管理员以确认该端口的有效设置。
- UseSSL - 如果将此项设置为 *True*，那么将使用 HTTPS 作为传输协议。 在这种情况下，您必须将 TM1 证书导入 SPSS Modeler Server JRE。

连接到 IBM Cognos TM1 多维数据集以导出数据

要将数据导出到 IBM Cognos TM1 数据库，请在 IBM Cognos TM1 对话框的**连接**选项卡上指定服务器连接详细信息，并选择关联的多维数据集和数据详细信息。

注: 将数据导出到 TM1 时，将废弃实际的“null”值。 零 (0) 值将作为有效值导出。 另请注意，在“映射”选项卡上，只有存储类型为字符串的字段才能映射到维度。 在导出到 TM1 之前，必须使用 IBM SPSS Modeler 客户端将非字符串数据类型转换为字符串。

连接类型。 选择**管理服务器**或 **TM1 服务器**。 请注意，已从 Planning Analytics on Cloud 中移除了管理服务器，因此如果您有连接到旧管理服务器的旧流，那么可以转而将其修改为指向 Planning Analytics on

Cloud。 如果在此处选择**管理服务器**，那么必须输入服务器的 URL（REST API 的主机名）和服务器的名称。 如果选择 **TM1 服务器**，请继续到以下部分。

TM1 服务器 URL。 输入要连接的 TM1 服务器将要安装到的管理主机 URL。 管理主机定义为所有 TM1 服务器的单个 URL。 通过此 URL 可以发现并访问环境中安装并运行的所有 IBM Cognos TM1 服务器。 单击**登录**。 如果先前未连接值此服务器，那么系统将提示您输入**用户名**和**密码**；另外，您可以搜索先前输入的已保存为**已存储的凭证**的登录详细信息。

选择要导出的 TM1 多维数据集显示可以向其导出数据的 TM1 服务器内多维数据集的名称。

为了选择要导出的数据，请选中多维数据集，然后单击右箭头，从而将该多维数据集移至**导出到多维数据集**字段。 选择多维数据集后，请使用**映射**选项卡将 TM1 维度和测量映射到相关 SPSS 字段或固定值（选择操作）。

映射 IBM Cognos TM1 数据以进行导出

选择 TM1 管理主机以及关联的 TM1 服务器和多维数据集之后，请使用“IBM Cognos TM1 导出”对话框的“映射”选项卡将 TM1 维度和度量映射到 SPSS 字段或者将 TM1 维度设置为固定值。

注: 只有存储类型为字符串的字段才能映射到维度。 在导出到 TM1 之前，必须使用 IBM SPSS Modeler 客户端将非字符串数据类型转换为字符串。

字段 列出 SPSS 数据文件中可供导出的数据字段名称。

TM1 维度 显示“连接”选项卡中选择的 TM1 多维数据集及其常规维度、测量维度及所选测量维度的元素。 选择要映射到 SPSS 数据字段的 TM1 维度（即，测量）的名称。

在“映射”选项卡上，提供了下列选项。

选择测量维度 从所选多维数据集的维度列表中，选择某个维度作为测量维度。

选择测量维度之外的某个维度，并单击**选择**时，将显示一个对话框，其中包含所选维度的叶元素。 您只能选择叶元素。 所选元素带有标签 **S**。

映射 将所选 SPSS 数据字段映射到所选 TM1 维度或度量（常规维度或者度量维度中的特定度量或元素）。 映射的字段标注了 **M**。

取消映射 取消所选 SPSS 数据字段与所选 TM1 维度或度量之间的映射。 请注意，一次只能将单个映射取消映射。 取消映射的 SPSS 数据字段将移回到左侧列中。

新建 在 TM1 测量维度中新建测量。 这将显示一个对话框，您可以在其中输入新的 **TM1 测量名称**。 此选项仅适用于度量维度，而不适用于常规维度。

有关 TM1 的更多信息，请参阅 IBM Cognos TM1 文档 (http://www-01.ibm.com/support/knowledgecenter/SS9RXT_10.2.2/com.ibm.swg.ba.cognos.ctm1.doc/welcome.html)。

SAS 导出节点

SPSS Modeler Professional 和 SPSS Modeler Premium 中提供了此功能。

使用 SAS 导出节点可以 SAS 格式写入数据，以将数据读入 SAS 软件包或 SAS 兼容软件包。 可以使用三种 SAS 文件格式进行导出：SAS for Windows/OS2、SAS for UNIX 或 SAS。

SAS 导出节点“导出”选项卡

导出文件。 指定文件名。 输入文件名，或单击“文件选择器”按钮浏览文件位置。

EXPORT。 指定导出文件格式。 选项有 **SAS for Windows/OS2**、**SAS for UNIX** 或 **SAS Version 7/8/9**。

导出字段名称。 选择从 IBM SPSS Modeler 导出字段名称和标签以供 SAS 使用的选项。

- **名称和变量标签**。 选择此选项可同时导出 IBM SPSS Modeler 字段名称和字段标签。 名称将以 SAS 变量名方式导出，而标签则以 SAS 变量标签方式导出。
- **作为变量标签的名称**。 选中此选项可将 IBM SPSS Modeler 字段名称用作 SAS 中的变量标签。 IBM SPSS Modeler 允许在字段名称中包含不适用于 SAS 变量名的字符。 为了防止创建无效的 SAS 名称，请选择 **名称和变量标签** 。

为此数据生成导入节点。 选中此选项可自动生成将读取已导出数据文件的 SAS 源节点。 有关更多信息，请参阅主题 第 34 页的『SAS 源节点』。

注: 字符串的最大允许长度为 255 字节。 如果字符串超过 255 字节，那么在导出时将被截断。

Excel 导出节点

Excel 导出节点以 Microsoft Excel `.xlsx` 格式输出数据。 也可以选择在执行节点时自动启动 Excel 并打开导出的文件。

Excel 节点“导出”选项卡

文件名。 输入文件名，或单击“文件选择器”按钮浏览文件位置。 缺省文件名为 *excelexp.xlsx*。

文件类型。 支持 Excel .xlsx 文件格式。

创建新文件。 新建 Excel 文件。

插入到现有文件中。 内容从**从单元格开始**字段指定的单元格开始替换。 电子表格中的其他单元格保留其原始内容。

包含字段名。 指定字段名是否可以包含在工作表的第一行中。

起始单元格。 用于首个导出记录（若选中了**包括字段名**，则为首个字段名）的单元格位置。 数据填入右侧并从此初始单元格向下填充。

选择工作表。 指定您要导出数据的目标工作表。 您可以按索引或按名称标识工作表:

- **按索引**。 如果您要创建新文件，指定从 0 到 9 的数字，以标识您要导出的目标工作表，开头的 0 表示第一个工作表，1 表示第二个工作表，依此类推。 如果在此位置已存在工作表，您可以使用 10 或更大的值。
- **按名称**。 如果您要创建新文件，指定用于工作表的名称。 如果您正在插入到现有文件，若工作表存在则数据插入此工作表，否则将创建具有此名称的新工作表。

启动 Excel。 指定执行节点时是否在导出文件中自动启动 Excel。 请注意，以分布式模式运行 IBM SPSS Modeler Server 时，输出结果将保存至服务器文件系统中，并在客户端上启动 Excel，其中显示已导出文件的副本。

为此数据生成导入节点。 选择此选项可自动生成将读取已导出数据文件的 Excel 源节点。 有关更多信息，请参阅主题 第 35 页的『Excel 源节点』。

“扩展导出”节点

通过“扩展导出”节点，您可以运行 R 或 Python for Spark 脚本来导出数据。

“扩展导出”节点 -“语法”选项卡

选择语法类型 - **R** 或 **Python for Spark**。 请参阅以下部分以获取更多信息。 语法就绪时，您可以单击**运行**来执行“扩展导出”节点。

R 语法

R 语法。 您可以在此字段中输入或粘贴用于数据分析的定制 R 脚本语法。

转换标志字段。 指定标志字段的处理方式。 共有两个选项：**将字符串转换为因子，将整数和实数转换为双精度数**和**逻辑值（True 和 False）**。 如果选择**逻辑值（True 和 False）**，那么标志字段的原始值将丢失。 例如，如果某个字段的值为 Male 和 Female，那么这些值将更改为 True 和 False。

将缺失值转换为 R "不可用" 值 (NA)。 选中时，任何缺失值都将转换为 R NA 值。 R 使用值 NA 来标识缺失值。 您使用的某些 R 函数可能有一个参数，可用于控制当数据包含 NA 时函数的行为方式。 例如，该函数可能会允许您选择自动排除包含 NA 的记录。 如果未选择此选项，那么所有缺失值都将按原样传递到 R，并可能导致执行 R 脚本时发生错误。

将日期/时间字段转换为特殊时区控制的 R 类。 如果选择此选项，那么会将带有日期或日期时间格式的变量转换为 R 日期/时间对象。 必须选择下列选项之一:

- 将具有日期或日期时间格式的 **RPOSIXct.** 变量将转换为 R POSIXct 对象。
- **R POSIXlt （列表）。** 将具有日期或日期时间格式的变量转换为 R POSIXlt 对象。

注: POSIX 格式是高级选项。 仅当您的 R 脚本指定以需要这些格式的方式处理日期时间字段时才使用这些选项。 POSIX 格式不适用于具有时间格式的变量。

Python 语法

Python 语法。 您可以针对数据分析向此字段中输入或粘贴定制的 Python 脚本语法。 有关 Python for Spark 的更多信息，请参阅 Python for Spark 和 Python for Spark 的脚本编制。

“扩展导出”节点 -“控制台输出”选项卡

控制台输出选项卡包含当“语法”选项卡上的 R 脚本或 Python for Spark 脚本运行时接收到的任何输出（例如，如果使用 R 脚本，当执行**语法**选项卡上的 **R 语法**字段中的 R 脚本时，它显示从 R 控制台接收到的输出）。 此输出可能包括执行 R 或 Python 脚本时生成的 R 或 Python 错误消息或警告。 输出可主要用于调试脚本。 **控制台输出**选项卡还包含 **R 语法**或 **Python 语法**字段中的脚本。

每次执行“扩展导出”脚本时，都会使用从 R 控制台或 Python for Spark 接收到的输出来覆盖**控制台输出**选项卡的内容。 输出不能编辑。

XML 导出节点

XML 导出节点允许您以使用 UTF-8 编码的 XML 格式输出数据。 还可选择创建 XML 源节点，以将导出的数据读取回到流中。

XML 导出条件。 您要导出数据的目标 XML 文件的完整路径和文件名。

使用 XML 模式。 如果您要使用模式或 DTD 来控制导出数据的结构，请选择此复选框。 这将激活下面所描述的**映射**按钮。

如果您不使用模式或 DTD，则对导出数据使用以下缺省结构：

```
<records>
     <record>
          <fieldname1>value</fieldname1>
          <fieldname2>value</fieldname2>
      :
          <fieldnameN>value</fieldnameN>
     </record>
     <record>
      :
      :
     </record>
      :
      :
</records>
```

字段名中的空格用下划线替换；例如，“My Field”将成为 <My_Field>。

映射。 如果您选择使用 XML 模式，该按钮会打开一个对话框，从中可以指定使用 XML 结构的哪个部分开始每个新记录。 有关更多信息，请参阅主题 第 285 页的『XML 映射记录选项』。

映射的字段。 表示已映射的字段数。

为此数据生成导入节点。 选中此选项可自动生成一个 XML 源节点，该节点会将已导出数据文件读取回到流中。 有关更多信息，请参阅主题 第 36 页的『XML 源节点』。

写入 XML 数据

当指定 XML 元素时，字段值会放入元素标记内：

```
<element>value</element>
```

当映射属性时，字段值会作为属性值放置：

```
<element attribute="value">
```

如果字段映射到 <records> 元素上面的元素，则字段仅写入一次，并作为所有记录的常量。 该元素的值将来自第一个记录。

如果要写入空值，那么可通过指定空内容来完成。 对于元素，此为：

```
<element></element>
```

对于属性，则为：

```
<element attribute="">
```

XML 映射记录选项

“记录”选项卡允许您指定使用 XML 结构的哪个部分来开始每个新记录。 要正确映射到模式，您需要指定记录定界符。

XML 结构。 显示前面屏幕中指定的 XML 模式的结构的层级树。

记录（XPath 表达式）。 要设置记录定界符，请选择 XML 结构中的元素，然后单击右箭头按钮。 每次在源数据中遇到此元素时，都将在输出文件中创建新的记录。

注意：如果您选择了 XML 结构中的根元素，那么只能写入单个记录，所有其他记录将被跳过。

XML 映射字段选项

当使用模式文件时，“字段”选项卡可用于将数据集中的字段映射到 XML 结构中的元素或属性。

只要元素或属性名称是唯一的，就会自动映射与元素或属性名匹配的字段名称。 因此，如果同时存在名为 field1 的元素和属性，则不会自动映射。 如果在结构中只有一个名为 field1 的项目，则自动映射在流中具有此名称的字段。

字段。 模型中的字段列表。 选择一个或多个字段作为映射的源部分。 您可以使用列表底部的按钮选择所有字段，或具有特定测量级别的所有字段。

XML 结构。 选择 XML 结构中的元素作为映射目标。 要创建映射，请单击“映射”。 然后将显示映射。 已通过此方式映射的字段数显示在列表下方。

要删除映射，选择 XML 结构列表中的项目，然后单击**解除映射**。

显示属性。 显示或隐藏 XML 结构中的 XML 元素的属性（如果有）。

XML 映射预览

在“预览”选项卡上，单击**更新**以查看将写入的 XML 的预览。

如果映射不正确，返回到“记录”或“字段”选项卡以纠正错误，然后再次单击**更新**以查看结果。

JSON 导出节点

使用 JSON 导出节点以使用 UTF-8 编码输出 JSON 格式的数据。 还可以选择创建 JSON 源节点以将导出的数据读取回流中。

在 SPSS Modeler 将数据写入 JSON 导出文件时，其执行以下转换。

表 *51: JSON* 数据导出转换

SPSS Modeler 数据存储	**JSON 值**
String	字符串
整数	number(int)
实数	number(real)
日期	字符串
时间	字符串
时间戳记	字符串
列表	不支持。 将排除列表字段。
缺失值	null

JSON 导出文件。 将导出数据的 JSON 文件的完整路径和文件名。

JSON 字符串格式。 指定 JSON 字符串格式。 如果想要 JSON 导出节点输出名称和值对集合，请选择**记录**。 或者，如果仅想导出值（无名称），那么选择**值**。

JSON 字符串格式。 指定 JSON 字符串格式。 选择“记录”，JSON 导出节点将输出名称和值对集合。 或者，如果仅想导出值（无名称），那么选择“值”。

为此数据生成导入节点。 选中此选项可自动生成一个 JSON 源节点，该节点会将已导出数据文件读取回到流中。 有关更多信息，请参阅 第 52 页的『JSON 源节点』。

“公共导出”节点选项卡

通过单击相应的选项卡可以为所有导出节点指定下列选项：

- **“发布”选项卡**。 用于发布流的结果。
- **“注解”选项卡**。 用于所有节点，此选项卡提供的选项可用于重命名节点、提供定制的工具提示及存储长的注解。

发布流

系统使用以下任何一个标准导出节点直接从 IBM SPSS Modeler 发布流：数据库、平面文件、统计信息 导出、扩展导出、数据收集导出、SAS 导出、Excel 和 XML 导出节点。 导出节点的类型决定每次使用 IBM SPSS Modeler Solution Publisher Runtime 或外部应用程序执行发布的流时要写入的结果格式。 例如，如果您想每次运行发布的流时将结果写入数据库，则请使用数据库导出节点。

发布流

1. 以普通方式打开或构建一个流，并在最后附加一个导出节点。
2. 在导出节点中的“发布”选项卡上，指定已发布文件的根名（即，`.pim`、`.par` 和 `.xml` 扩展将附加到的文件名）。
3. 单击**发布**以发布该流，或选择**发布流**以便每次执行该节点时自动发布流。

已发布的名称。 指定已发布图像和参数文件的根名。

- **图像文件** (`*.pim`) 提供了 Runtime 执行发布的流时所需的所有信息，这些信息与导出时完全相同。 如果您确信不需要更改流的任何设置（如输入数据源或输出数据文件），则可以只部署该图像文件。
- **参数文件** (`*.par`) 包含有关数据源、输出文件和执行选项的可配置信息。 如果您希望能够在不重新发布流的情况下控制流的输入或输出，则同时需要参数文件和图像文件。
- **元数据文件** (`*.xml`) 用于描述图像及其数据模型的输入和输出。 它旨在供内嵌 runtime 库并需要了解输入数据和输出数据结构的应用程序使用。

  注: 仅当您选择了**发布元数据**选项时，才会生成此文件。

发布参数。 如果需要，可以在 *.par 文件中包含流参数。 在通过编辑 *.par 文件或通过运行时 API 执行图像时，可以更改这些流参数值。

此选项用于启用**参数**按钮。 单击该按钮时，将显示“发布参数”对话框。

通过在**发布**列中选择相关选项来选择您要包含在已发布图像中的参数。

在流执行时。 指定执行节点时是否自动发布流。

- **导出数据**。 以标准方式执行导出节点，但不发布流。 （基本上，节点在 IBM SPSS Modeler 中的执行方式与 IBM SPSS Modeler Solution Publisher 不可用时的执行方式相同。） 如果您选择此选项，那么只有通过单击导出节点对话框上的**发布**进行明确发布时，才会发布该。 另外，您还可以通过使用工具栏上的发布工具，或通过使用脚本来发布当前流。
- **发布流**。 使用 IBM SPSS Modeler Solution Publisher 发布流以用于部署。 如果您希望每次执行节点时都自动发布流，则请选择此选项。

注:

- 如果您计划使用新数据或更新后的数据运行发布的流，则要注意的重要一点是，输入文件中字段的顺序必须与发布的流中指定的源节点输入文件中的字段顺序相同。
- 发布到外部应用程序时，请考虑对无关的字段进行过滤，或者对字段进行重命名以符合输入要求。 通过在导出节点之前使用过滤节点，可以完成上述两个操作。

第 8 章 IBM SPSS StatisticsNodes

IBM SPSS Statistics 节点 - 概述

作为 IBM SPSS Modeler 及其数据挖掘功能的补充，IBM SPSS Statistics 允许您进一步执行统计分析和数据管理。

安装 IBM SPSS Statistics 的兼容、受许可副本后，您可以从 IBM SPSS Modeler 与它连接，并执行 IBM SPSS Modeler 不支持的复杂、多步数据操作与分析。 对于高级用户，还提供了一个使用命令语法对分析进行进一步修改的选项。 请参阅“发行说明”以了解关于版本兼容性的信息。

如果可用，IBM SPSS Statistics 节点显示在节点选用板的专门部分中。

注: 建议您在使用 IBM SPSS Statistics 的“转换”、“模型”和“输出”节点之前，在“类型”节点中实例化数据。 使用 AUTORECODE 语法命令时，这还是一项要求。

IBM SPSS Statistics 选用板包含下列节点：

Statistics 文件节点从 IBM SPSS Statistics 使用的 *.sav* 或 *.zsav* 文件格式以及保存在 IBM SPSS Modeler 中的缓存文件（也使用同一格式）读取数据。

Statistics 转换节点针对 IBM SPSS Modeler 中的数据源运行所选的 IBM SPSS Statistics 语法命令。 此节点需要 IBM SPSS Statistics 的许可副本。

Statistics 模型节点使您能够通过运行生成 PMML 的 IBM SPSS Statistics 过程分析和处理数据。 此节点需要 IBM SPSS Statistics 的许可副本。

Statistics 输出节点可调用 IBM SPSS Statistics 过程以分析您的 IBM SPSS Modeler 数据。 可以访问许多不同的 IBM SPSS Statistics 分析过程。 此节点需要 IBM SPSS Statistics 的许可副本。

Statistics 导出节点以 IBM SPSS Statistics *.sav* 或 *.zsav* 格式输出数据。 `.sav` 或 `.zsav` 文件可以由 IBM SPSS Statistics Base 和其他产品读取。 这也是用于 IBM SPSS Modeler 中的高速缓存文件的格式。

注: 如果您的 SPSS 统计信息 副本仅授权给单一用户使用，而您运行的流带有两个或两个以上分支，并且每个分支均包含 SPSS 统计信息 节点，那么您可能会得到许可授权错误。 当某个分支的 SPSS 统计信息 会话尚未结束，而另一个分支试图起动时，就会出现此错误。 如可行，应重新设计流，以确保带有 SPSS 统计信息 节点的多个分支不会同时执行。

Statistics 文件节点

可以使用 Statistics 文件节点从已保存的 IBM SPSS Statistics 文件（*.sav* 或 *.zsav*）中直接读取数据。 现在可使用该格式替换 IBM SPSS Modeler 早期版本中的缓存文件。 如果想要导入已保存的缓存文件，则应使用 IBM SPSS Statistics 文件节点。

导入文件。 指定文件名。 可以输入文件名或单击省略按钮 (**...**) 来选择文件。 一旦选定了一个文件，即可显示此文件的路径。

文件受密码加密。 如果您知道该文件受密码保护，请选中此框；系统将提示您输入**密码**。 如果该文件受密码保护，但您未输入密码，那么在尝试切换至另一选项卡、刷新数据、预览节点内容或尝试执行包含节点的流时，将显示一条警告消息。

注: 受密码保护的文件只能由 IBM SPSS Modeler V16 或更高版本打开。

变量名称。 选择从 IBM SPSS Statistics *.sav* 或 *.zsav* 文件导入变量名称和标签时的处理方法。 在您使用 IBM SPSS Modeler 的整个过程中，所选的包含在此处的元数据会保留，并且可以再次导出以在 IBM SPSS Statistics 中使用。

- **读取名称和标签**。 选中此选项可将变量名称和标签同时读入 IBM SPSS Modeler。 缺省情况下将选中此选项，并且变量名称将显示在“类型”节点中。 根据流属性对话框中指定的选项，标签将显示在图表、模型浏览器和其他类型的输出中。 缺省情况下，将禁止在输出中显示标签。
- **读取标签作为名称**。 选择从 IBM SPSS Statistics *.sav* 或 *.zsav* 文件中读取描述性变量标签（而不是短字段名称），并在 IBM SPSS Modeler 中将这些标签用作变量名称。

值。 选择从 IBM SPSS Statistics *.sav* 或 *.zsav* 文件导入值和标签时的处理方法。 在您使用 IBM SPSS Modeler 的整个过程中，所选的包含在此处的元数据会保留，并且可以再次导出以在 IBM SPSS Statistics 中使用。

- **读取数据和标签**。 选中此选项可将实际值和值标签同时读入 IBM SPSS Modeler。 缺省情况下将选中此选项，并且这些值本身将显示在“类型”节点中。 根据流属性对话框中指定的选项，值标签将显示在表达式构建器、图表、模型浏览器和其他类型的输出中。
- **读取标签作为数据**。 如果要使用 *.sav* 或 *.zsav* 文件中的值标签而不是用于表示值的数字或符号代码，请选中此选项。 例如，对于含性别字段（其值 1 和 2 实际上分别代表男性和女性）的数据，选中此选项可将该字段转换为字符串，并将男性和女性作为实际值导入。

  选中此选项前考虑 IBM SPSS Statistics 数据中的缺失值非常重要。 例如，如果数字字段仅对缺失值使用标签 (0 = *No Answer* , -99 = *Unknown*)，则选中上述选项将仅导入值标签 *No Answer* 和 *Unknown* ，并会将字段转换为字符串。 在这种情况下，应在类型节点中导入值本身并设置缺失值。

使用字段格式信息来确定存储。 如果取消选中此复选框，那么将使用整数存储导入 *.sav* 文件中格式化为整数的字段值（即，在 IBM SPSS Statistics 的“变量视图”中指定为 F*n*.0 的字段）。 将除字符串外的所有其他字段值作为实数导入。

如果选中了此框（缺省情况），那么无论是否已在 *.sav* 文件中格式化为整数，除字符串以外的所有字段值都将作为实数导入。

多重响应集。 导入文件后，IBM SPSS Statistics 文件中定义的任何多重响应集都将自动被保留。 借助“过滤器”选项卡，您可以查看和编辑任意节点的多重响应集。 有关更多信息，请参阅主题 第 114 页的『编辑多重响应集』。

Statistics 转换节点

使用 Statistics 转换节点，可以使用 IBM SPSS Statistics 命令语法完成数据转换。 这样便有可能完成 IBM SPSS Modeler 不支持的若干变换，并实现复杂多步变换的自动化，包括通过单一节点创建多个字段。 这种节点与 Statistics 输出节点类似，只是数据将返回 IBM SPSS Modeler 进行进一步的分析，而输出节点中的数据将作为请求的输出对象（如图形或表格）返回。

要使用此节点，必须在计算机上安装 IBM SPSS Statistics 的兼容版本并许可使用。 有关更多信息，请参阅 第 262 页的『IBM SPSS Statistics 帮助应用程序』。 请参阅在线发行说明以获得有关兼容性的信息。

如有必要，可以使用“过滤”选项卡过滤或重命名字段，以便它们符合 IBM SPSS Statistics 命名标准。 有关更多信息，请参阅第 278 页的『重命名或过滤 IBM SPSS Statistics 的字段』。

语法参考。 有关特定 IBM SPSS Statistics 过程的详细信息，请参阅 IBM SPSS Statistics 软件拷贝随附的 *IBM SPSS Statistics* 命令语法参考指南。 要从“语法”选项卡查看该指南，请选择**语法编辑器**选项，然后单击**启动 Statistics 语法帮助**按钮。

注意：此节点并不支持所有 IBM SPSS Statistics 语法。 有关更多信息，请参阅主题 第 291 页的『允许的语法』。

Statistics 转换节点 - “语法”选项卡

IBM SPSS Statistics 对话框选项

如果不熟悉某个过程的 IBM SPSS Statistics 语法，那么在 IBM SPSS Modeler 中创建语法的最简单方式是选择 **IBM SPSS Statistics 对话**选项，选择该过程的对话框，完成对话框并单击“确定”。 这样可以将语法放入当前在 IBM SPSS Modeler 中使用的 IBM SPSS Statistics 节点的“语法”选项卡中。 然后，可以运行流以获得过程输出。

IBM SPSS Statistics 语法编辑器选项

选中这一项。 在对话框上面输入语法命令后，使用此按钮可验证您的输入。 所有不正确的语法将在对话框下面标示。

为确保检查过程不会过长，当您验证语法时，会对数据的代表性样本进行检查（而不是检查整个数据集），以确保输入有效。

允许的语法

如果有大量继承自 IBM SPSS Statistics 的语法或熟悉 IBM SPSS Statistics 的数据准备功能，您可以使用 Statistics 转换节点运行很多现有转换。 提示：使用该节点可按可预测的方式变换数据 - 例如，通过运行循环命令或通过对数据进行更改、添加、排序、过滤或选择。

可使用的命令示例如下：

- 根据二项式分布计算随机数：

```
COMPUTE newvar = RV.BINOM(10000,0.1)
```

- 将某个变量重新编码为新变量：

```
RECODE Age (Lowest thru 30=1) (30 thru 50=2) (50 thru Highest=3) INTO
AgeRecoded
```

- 替换缺失值：

```
RMV Age_1=SMEAN(Age)
```

下表列出了 Statistics 转换节点所支持的 IBM SPSS Statistics 语法。

命令名称

ADD VALUE LABELS

APPLY DICTIONARY

AUTORECODE

BREAK

CD

CLEAR MODEL PROGRAMS

CLEAR TIME PROGRAM

CLEAR TRANSFORMATIONS

COMPUTE

COUNT

CREATE

DATE

DEFINE-!ENDDEFINE

DELETE VARIABLES

命令名称

DO IF

DO REPEAT

ELSE

ELSE IF

END CASE

END FILE

END IF

END INPUT PROGRAM

END LOOP

END REPEAT

EXECUTE

FILE HANDLE

FILE LABEL

FILE TYPE-END FILE TYPE

FILTER

FORMATS

IF

INCLUDE

INPUT PROGRAM-END INPUT PROGRAM

INSERT

LEAVE

LOOP-END LOOP

MATRIX-END MATRIX

MISSING VALUES

N OF CASES

NUMERIC

PERMISSIONS

PRESERVE

RANK

RECODE

RENAME VARIABLES

RESTORE

RMV

SAMPLE

SELECT IF

SET

SORT CASES

命令名称

STRING

SUBTITLE

TEMPORARY

TITLE

UPDATE

V2C

VALIDATEDATA

VALUE LABELS

VARIABLE ATTRIBUTE

VARSTOCASES

VECTOR

Statistics 模型节点

Statistics 模型节点使您能够通过运行生成 PMML 的 IBM SPSS Statistics 过程分析和处理数据。 然后，创建的模型块可按常规方式在 IBM SPSS Modeler 流中进行评分等操作。

要使用此节点，必须在计算机上安装 IBM SPSS Statistics 的兼容版本并许可使用。 有关更多信息，请参阅第 262 页的『IBM SPSS Statistics 帮助应用程序』。 请参阅在线发行说明以获得有关兼容性的信息。

可用的 IBM SPSS Statistics 分析程序取决于您有的许可证类型。

Statistics 模型节点 - “模型”选项卡

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

选择对话框。 单击以显示您可以选择并运行的可用 IBM SPSS Statistics 过程列表。 此列表仅显示那些生成 PMML 且您已获得许可的过程，并且不包括用户编写的过程。

1. 单击所需过程；显示相关 IBM SPSS Statistics 对话框。
2. 在 IBM SPSS Statistics 对话框中输入过程详细信息。
3. 单击**确定**返回到 Statistics 模型节点；在“模型”选项卡中显示 IBM SPSS Statistics 语法。
4. 在任何时候，要返回到 IBM SPSS Statistics 对话框，例如，要修改您的查询，请单击过程选择按钮右侧的 IBM SPSS Statistics 对话框显示按钮。

Statistics 模型节点 - 模型块汇总

在运行 Statistics 模型节点时，它执行相关的 IBM SPSS Statistics 过程并创建您可在 IBM SPSS Modeler 流中进行评分的模型块。

模型块的“摘要”选项卡显示有关字段、构建设置和模型估计过程的信息。 结果以树形视图显示，通过单击指定项可以扩展或合并树形视图。

查看模型按钮以 IBM SPSS Statistics 输出查看器的修改形式显示结果。 有关该查看器的更多信息，请参阅 IBM SPSS Statistics 文档。

“文件”菜单中提供了常用的导出和打印选项。 有关更多信息，请参阅主题第 230 页的『查看输出』。

Statistics 输出节点

Statistics 输出节点可调用 IBM SPSS Statistics 过程以分析您的 IBM SPSS Modeler 数据。 您可以在浏览器窗口中查看结果，或以 IBM SPSS Statistics 输出文件格式保存结果。 从 IBM SPSS Modeler 中，可以访问许多不同的 IBM SPSS Statistics 分析过程。

要使用此节点，必须在计算机上安装 IBM SPSS Statistics 的兼容版本并许可使用。 有关更多信息，请参阅第 262 页的『IBM SPSS Statistics 帮助应用程序』。 请参阅在线发行说明以获得有关兼容性的信息。

如有必要，可以使用“过滤”选项卡过滤或重命名字段，以便它们符合 IBM SPSS Statistics 命名标准。 有关更多信息，请参阅第 278 页的『重命名或过滤 IBM SPSS Statistics 的字段』。

语法参考。 有关特定 IBM SPSS Statistics 过程的详细信息，请参阅 IBM SPSS Statistics 软件拷贝随附的 *IBM SPSS Statistics* 命令语法参考指南。 要从“语法”选项卡查看该指南，请选择**语法编辑器**选项，然后单击**启动 Statistics 语法帮助**按钮。

Statistics 输出节点 - “语法”选项卡

使用此选项卡为要用于分析您的数据的 SPSS 统计信息 过程创建语法。 语法由两个部分组成：**语句**和关联的**选项**。 语句指定要执行的分析或操作和要使用的字段。 选项指定所有其他内容，包括要显示的统计量、要保存的导出字段，等等。

SPSS 统计信息 对话框选项

如果不熟悉某个过程的 IBM SPSS Statistics 语法，那么在 IBM SPSS Modeler 中创建语法的最简单方式是选择 **IBM SPSS Statistics 对话**选项，选择该过程的对话框，完成对话框并单击“确定”。 这样可以将语法放入当前在 IBM SPSS Modeler 中使用的 IBM SPSS Statistics 节点的“语法”选项卡中。 然后，可以运行流以获得过程输出。

可以选择性地生成一个“Statistics 文件”源节点，用于导入生成的数据。 此节点非常有用，例如，除显示输出之外，某个过程将字段（例如评分）写入活动数据集。

注:

- 当以非英语语言生成输出时，建议在语法中指定该语言。
- “统计量输出”节点中不支持“输出样式”选项。

创建语法

1. 单击**选择对话框**按钮。
2. 选择其中一个选项:
   - **分析** 列出 SPSS 统计信息 分析菜单的内容；选择您要使用的过程。
   - **其他** 如果显示，则列出在 SPSS 统计信息 的自定义对话框构建器中创建的对话框，以及未出现在“分析”菜单上且您具有许可的任何其他 SPSS 统计信息 对话框。 如果没有适用的对话框，那么将不显示此选项。

注: 不会显示“自动数据准备”对话框。

如果某个 SPSS 统计信息 自定义对话框创建有新字段，则这些字段不能在 SPSS Modeler 中使用，因为 Statistics 输出节点为终端节点。

也可以选中**为结果数据生成导入节点**复选框来创建 Statistics 文件源节点，以便用于将结果数据导入其他流。 该节点放置在屏幕工作区中，数据包含在由**文件**字段指定的 `.sav` 文件中（缺省位置是 SPSS Modeler 安装目录）。

语法编辑器选项

要保存针对频繁使用的过程创建的语法，请执行下列操作:

1. 单击**文件选项**按钮（工具栏上的第一个按钮）。
2. 从菜单中选择**保存**或**另存为**。

3. 将文件保存为 .sps 文件。

要使用先前创建的语法文件，并同时替换语法编辑器的当前内容（如果有），请执行下列操作：

1. 单击“文件选项”按钮（工具栏上的第一个按钮）。
2. 从菜单中选择**打开**。
3. 选择 .sps 文件以将其内容粘贴到“输出”节点“语法”选项卡中。

要插入先前保存的语法而不替换当前内容，请执行下列操作：

1. 单击“文件选项”按钮（工具栏上的第一个按钮）。
2. 从菜单中选择**插入**
3. 选择 .sps 文件以将其内容粘贴到“输出”节点的光标指定位置。

也可以选中**为结果数据生成导入节点**复选框来创建 Statistics 文件源节点，以便用于将结果数据导入其他流。 该节点放置在屏幕工作区中，数据包含在由**文件**字段指定的 .sav 文件中（缺省位置是 SPSS Modeler 安装目录）。

单击**运行**时，结果会显示在 SPSS 统计信息 输出查看器中。 有关查看器的更多信息，请参阅 SPSS 统计信息 文档。

注： 不支持以下项（及其对应 SPSS 统计信息 对话框选项）的语法。 它们不会影响输出。

- OUTPUT ACTIVATE
- OUTPUT CLOSE
- OUTPUT DISPLAY
- OUTPUT EXPORT
- OUTPUT MODIFY
- OUTPUT NAME
- OUTPUT NEW
- OUTPUT OPEN
- OUTPUT SAVE

Statistics 输出节点 - “输出”选项卡

使用“输出”选项卡，可以指定输出的格式和位置。 您可以选择将结果显示在屏幕上，也可以将其发送给其中一种可用字段类型。

输出名称。 指定执行节点时使用的输出名称。 **自动** 根据生成输出的节点选择名称。 （可选）可以选择**定制**以指定其他名称。

输出到屏幕（缺省选项）。 创建要在线查看的输出对象。 当执行输出节点时，该输出对象将显示在管理器窗口的“输出”选项卡上。

输出到文件。 运行节点时将输出保存到文件。 如果选择此选项，请在**文件名**字段中输入文件名（或导航到某目录，并使用文件选择器按钮指定文件名）并选择文件类型。

文件类型。 选择您要发送输出的目标文件类型。

- **HTML 文档 (*.html)**。 以 HTML 格式写入输出。
- **IBM SPSS Statistics 查看器文件 (*.spv)**。 以可由 IBM SPSS Statistics 输出查看器读取的格式写入输出。
- **IBM SPSS Statistics Web 报告文件 (*.spw)**。 以 IBM SPSS Statistics Web 报告格式写入输出，此类文件可以发布到 IBM SPSS 协作和部署服务 存储库，随后在 Web 浏览器中查看。 有关更多信息，请参阅主题 第 231 页的『发布到 Web』。

注： 如果选择**输出到屏幕**，那么 IBM SPSS Statistics OMS 指令 VIEWER=NO 将不起作用；另外，脚本编写 API（基本和 *Python SpssClient* 模块）将在 IBM SPSS Modeler 中不可用。

Statistics 导出节点

“统计信息导出”节点使您能够以 IBM SPSS Statistics *.sav* 格式导出数据。IBM SPSS Statistics *.sav* 文件可由 IBM SPSS Statistics Base 和其他模块读取。 这也是 IBM SPSS Modeler 缓存文件所使用的格式。

将 IBM SPSS Modeler 字段名称映射到 IBM SPSS Statistics 变量名称某些时候可能导致错误，因为 IBM SPSS Statistics 变量名称限制为 64 个字符，而且不能包含特定字符，例如，空格、美元符号 ($) 和连字符 (–)。可通过两种方式针对以下限制进行调整：

- 可通过单击“过滤”选项卡，根据 IBM SPSS Statistics 变量名要求重命名字段。 有关更多信息，请参阅主题 第 278 页的『重命名或过滤 IBM SPSS Statistics 的字段』。
- 选择同时从 IBM SPSS Modeler 导出字段名和标签。

注意：IBM SPSS Modeler 以 Unicode UTF-8 格式写入 *.sav* 文件。 IBM SPSS Statistics 16.0 版以后的版本只支持 Unicode UTF-8 格式的文件。 为了防止数据可能损坏，使用 Unicode 编码保存的 *.sav* 文件不得用于 16.0 之前的 IBM SPSS Statistics 版本。 有关详细信息，请参阅 IBM SPSS Statistics 帮助。

多重响应集。 导出文件后，将自动保留流中定义的任何多重响应集。 借助“过滤器”选项卡，您可以查看和编辑任意节点的多重响应集。 有关更多信息，请参阅主题 第 114 页的『编辑多重响应集』。

Statistics 导出节点 - “导出”选项卡

导出文件 指定文件的名称。 输入文件名，或单击“文件选择器”按钮浏览文件位置。

文件类型 请选择是以正常的 *.sav* 格式还是压缩的 *.zsav* 格式保存文件。

使用密码对文件加密 要使用密码保护文件，请选中此框；系统将提示您在另一对话框中输入并确认**密码**。

注: 受密码保护的文件只能由 SPSS Modeler V16 或更高版本，或者由 SPSS 统计信息 V21 或更高版本打开。

导出字段名称 指定将变量名称和标签从 SPSS Modeler 导出至 SPSS 统计信息 *.sav* 或 *.zsav* 文件时的处理方法。

- **名称和变量标签** 选择此选项可同时导出 SPSS Modeler 字段名称和字段标签。 名称将以 SPSS 统计信息 变量名方式导出，而标签则以 SPSS 统计信息 变量标签方式导出。
- **作为变量标签的名称** 选中此选项可将 SPSS Modeler 字段名称用作 SPSS 统计信息 中的变量标签。 SPSS Modeler 允许在字段名称中包含不适用于 SPSS 统计信息 变量名称的字符。 为了防止创建无效的 SPSS 统计信息 名称，请选择**将名称用作变量标签**或使用“过滤”选项卡调整字段名。

启动应用程序 如果计算机上安装了 SPSS 统计信息，则可以选择此选项对已保存数据文件直接激活应用程序。 “帮助应用程序”对话框中必须指定用于启动应用程序的选项。 有关更多信息，请参阅主题 第 262 页的『IBM SPSS Statistics 帮助应用程序』。 要在不打开外部程序的情况下直接创建 SPSS 统计信息 *.sav* 或 *.zsav* 文件，请取消选择此选项。

注: 当以服务器（分布式）方式同时运行 SPSS Modeler 和 SPSS 统计信息 时，写出数据并启动 SPSS 统计信息 会话不会自动打开 SPSS 统计信息 客户端以显示读入活动数据集的数据集。 变通方法是 SPSS 统计信息 客户端启动时立刻在其中打开数据文件。

针对此数据生成一个导入节点 选中此选项可自动生成将读取已导出数据文件的“Statistics 文件”源节点。 有关更多信息，请参阅主题 第 26 页的『Statistics 文件节点』。

重命名或过滤 IBM SPSS Statistics 的字段

将数据从 IBM SPSS Modeler 导出或部署到外部应用程序（比如 IBM SPSS Statistics）之前，可能需要重命名或调整字段名。 “统计信息变换”、“统计信息输出”和“统计信息导出”对话框提供了一个“过滤”选项卡，用于帮助执行此过程。

“过滤”选项卡基本功能将在后文介绍。 有关更多信息，请参阅主题 第 113 页的『设置过滤选项』。 本主题提供将数据读入 IBM SPSS Statistics 的提示。

要调整字段名称以符合 IBM SPSS Statistics 命名规则：

1. 在“过滤”选项卡上，单击“过滤选项菜单”工具栏按钮（工具栏上的第一个按钮）。

2. 选择“为 IBM SPSS Statistics 重命名”。
3. 在“为 IBM SPSS Statistics 重命名”对话框中，您可以选择使用**井号 (#)** 字符或**下划线 (_)** 替换文件名中的无效字符。

重命名多重响应集。 如果您要调整多重响应集（可通过 Statistics 文件源节点导入 IBM SPSS Modeler）的名称，请选择此项。 多重响应集可用于记录对每个问题都具有多个值的数据，例如，在调查响应时。

第 9 章 超节点

超节点概述

IBM SPSS Modeler 可视化编程界面容易学习的其中一个原因是，每个节点都有明确定义的功能。 但是，要进行复杂处理，可能需要一个较长的节点序列。 最后，这可能会使流画布混乱不堪，并导致难以遵循流图。 要避免长而复杂的流混乱不堪，有两种方法：

- 您可以将处理序列拆分为几个相互融汇的流。 例如，第一个流创建第二个流作为输入使用的数据文件。 第二个流创建第三个流作为输入使用的文件，依此类推。 您可以通过将多个流保存在 **工程** 中来对它们进行管理。 工程提供多个流及其输出的组织。 但是，工程文件只包含对它所包含的对象的引用，并且仍旧有多个流文件需要您进行管理。
- 处理复杂的流过程时，可以创建有着更为清晰的流程的 **超节点** 作为备选方案。

超节点通过封装数据流的组成部分将多个节点组成一个节点。 这会为数据挖掘程序提供许多便利：

- 流更加整洁并且更易管理。
- 节点可以合并为业务特定的超节点。
- 可将超节点导出到库中以在多个数据挖掘项目中重复使用。

超节点的类型

超节点在数据流中由星形图标表示。 该图标中具有阴影，用于表示超节点的类型以及流必须流入或流出的方向。

存在三种类型的超节点：

- 源超节点
- 过程超节点
- 终端超节点

源超节点

源超节点包含一个类似于普通源节点的数据源，并且可在可使用普通源节点的任意位置使用。 源超节点的左侧具有阴影，表示它在左侧“关闭”，并且数据从 超节点流动到下游。

源超节点只在右侧具有一个连接点，表示数据离开超节点并流向流中。

过程超节点

过程超节点只包含过程节点，并且没有阴影，这表示数据可 流入 这种类型的超节点和从这种类型的超节点流出 。

过程超节点在左侧和右侧都有连接点，表示数据进入超节点并离开以流回到流中。 虽然超节点可以包含附加流段，甚至额外的流，但这两个连接点都必须通过一条连接 开始流 点和 结束流 点的路径流动。

注意：过程超节点有时也称为操纵超节点。

终端超节点

终端超节点包含一个或多个终端节点（图、表等）并且使用方式可与终端节点相同。 终端超节点的右侧具有阴影，表示它在右侧“关闭”并且数据只能流入 终端超节点中。

终端超节点只在左侧具有一个连接点，表示数据从流进入超节点，并在超节点内终止。

终端超节点也可以包含脚本，脚本用于指定超节点内所有终端节点的执行顺序。 有关更多信息，请参阅主题 第 304 页的『超节点和脚本编写』。

创建超节点

创建超节点时，通过将多个节点封装为一个节点可“缩短”数据流。 一旦在工作区上创建或加载了流，便有多种方法来创建超节点。

多项选择

创建超节点的最简单方法是选择要封装的所有节点：

1. 使用鼠标选择流画布上的多个节点。 也可以通过在按住 Shift 键的同时单击来选择流或流的一部分。

   注: 所选择的节点必须来自连续流或分支流。 不能选择不相邻或未以某种方式连接的节点。
2. 然后，使用下面三种方法之一来封装所选节点：
   - 单击工具栏上的“超节点”图标（形状像星形）。
   - 右键单击超节点，然后从上下文菜单中选择：

     创建超节点 > 从选择
   - 从“超节点”菜单中，选择：

     创建超节点 > 从选择

所有这三个选项均将节点封装到一个超节点中，该超节点具有阴影，用于根据它的内容反映它的类型：源、过程或终端。

单选

也可以通过以下方式创建超节点：选择一个节点并使用菜单选项确定超节点的开始和结束，或封装所选节点的全部下游内容。

1. 单击确定封装开始的节点。
2. 从“超节点”菜单中，选择：

   创建超节点 > 从此处

也可以通过选择流部分的开始和结束来封装节点，以更加交互的方式创建超节点：

1. 单击要包括在超节点中的第一个或最后一个节点。
2. 从“超节点”菜单中，选择：

   创建超节点 > 选择...
3. 或者，可以通过右键单击所需的节点来使用上下文菜单选项。
4. 光标会变为“超节点”图标，表示必须选择流中的其他点。 逆流或顺流移动到“超节点”段的“另一端”并单击某节点。 此操作将使用“超节点”星形图标替换所有节点。

注: 所选择的节点必须来自连续流或分支流。 不能选择不相邻或未以某种方式连接的节点。

嵌套超节点

超节点可以嵌套在其他超节点中。 用于每类超节点（源、过程和终端）的规则也应用于嵌套超节点。 例如，具有嵌套的过程超节点必须具有通过所有嵌套超节点的连续数据流，才能保持为过程超节点。 如果其中一个嵌套超节点是终端节点，则数据将不再通过层次结构流动。

终端超节点和源超节点可以包含其他类型的嵌套超节点，但用于创建超节点的基本规则同样适用。

锁定超节点

一旦您已创建超节点，您可以使用密码将其锁定，以防修改。 例如，如果您创建流或部分流，您可以进行此操作，因为在您组织中其他人使用的固定值模板，这些人设置 IBM SPSS Modeler 查询的经验较少。

当锁定超节点时，用户仍然可以在“参数”选项卡上为任何已经定义的参数输入值，且不输入密码也可执行锁定的超节点。

注: 不能使用脚本执行锁定和解锁。

锁定和解锁超节点

警告: 丢失的密码不能恢复。

您可以从三个选项卡中的任何选项卡锁定或解锁超节点。

1. 单击**锁定节点**。
2. 输入并确认密码。
3. 单击**确定**。

受密码保护的超节点在流画布上通过超节点图标左上角的小挂锁符号标识。

解锁超节点

1. 要永久删除密码保护，请单击**解锁节点**。 将提示您输入密码。
2. 输入密码并单击**确定**。 超节点不再受密码保护，挂锁符号不再显示在流中图标的旁边。

对于在包含锁定超节点的 SPSS Modeler V16 和 V17.0 中保存的流，在其他环境中打开流（例如，在 IBM SPSS 协作和部署服务 中），或者在 Mac 上，在 SPSS Modeler 安装的 JRE 不同时，首先必须使用上次保存所在的旧环境上的 V17.1 或更高版本进行打开、解锁和重新保存。

某些时候，在解锁 V18 之前的流中的超节点时，将显示不正确的密码错误。 要解决此问题，请在具有相同系统本地设置的相同平台上使用与上次打开相同的 IBM SPSS Modeler 版本（或更新版本）重新打开并解锁节点。 然后，在 V18 或更高版本中进行打开。 解锁节点，然后重新保存流。

编辑锁定的超节点

如果您尝试定义参数或放大以显示锁定的超节点，将提示您输入密码。

输入密码并单击**确定**。

您现在能够根据需要随时编辑参数定义并放大和缩小，直到您关闭超节点所在的流。

注意这并不会删除密码保护，只允许您访问以处理超节点。 有关更多信息，请参阅主题 第 301 页的『锁定和解锁超节点』。

编辑超节点

一旦您已创建了超节点，您可以通过放大更加仔细地检查；如果超节点已锁定，那么系统将提示您输入密码。 有关更多信息，请参阅主题 第 301 页的『编辑锁定的超节点』。

要查看超节点的内容，可以使用 IBM SPSS Modeler 工具栏中的放大图标或以下方法：

1. 右键单击超节点。
2. 从上下文菜单中，选择 **放大** 。

所选超节点的内容将在略有不同的 IBM SPSS Modeler 环境中显示，其中的连接器显示通过流或流段进行的数据流动。 在流画布的此级别上，可以执行多项任务：

- 修改超节点类型：源、过程或终端。
- 创建参数或编辑参数的值。 参数在脚本和 CLEM 表达式中使用。
- 指定超节点及其子节点的缓存选项。
- 创建或修改超节点脚本（仅限终端超节点）。

修改超节点类型

在一些情况下，改变超节点的类型会很有用。 只有在放大超节点时，此选项才可用，并且它仅在该级别适用于超节点。 下表对三种类型的超节点进行了说明。

表 *52:* 超节点的类型	
超节点的类型	**描述**
源超节点	一个传出的连接
过程超节点	两个连接：一个传入，一个传出
终端超节点	一个传入的连接

要更改超节点的类型

1. 确保放大超节点。
2. 从“超节点”菜单中，选择**超节点类型**，然后选择类型。

添加注解和重命名超节点

您可以在超节点显示在流中时重命名超节点，以及编写在项目或报告中使用的注解。 要访问这些属性，请执行以下操作：

- 右键单击超节点（缩小），然后选择 **重命名并注解** 。
- 或者，从“超节点”菜单中，选择**重命名并注解**。 此选项在缩小模式和放大模式下均可用。

在这两种情况下，会打开一个对话框，其中“注解”选项卡处于选定状态。 使用此处的选项可定制显示在流工作区上的名称，并提供与超节点操作相关的文档。

使用带有超节点的注释

从带注释的节点或块创建超节点时，如果您想在超节点中出现注释，那么必须在创建超节点的选择中包括注释。 如果您在选择中省略了注释，那么注释将在创建超节点时在流上保留。

当您展开包括注释的超节点时，注释恢复到创建超节点之前的位置。

当您展开包括注释对象的超节点，但注释不包含在超节点中时，对象恢复到之前的位置，但不会再次添加注释。

超节点参数

在 IBM SPSS Modeler 中，您可以设置用户定义的变量（例如 `Minvalue`），在脚本或 CLEM 表达式中使用时可以指定这些变量的值。 这些变量被称为 **参数** 。 您可以为流、会话和超节点设置参数。 当在该超节点或任何嵌套节点中构建 CLEM 表达式时，超节点的所有参数集均可用。 嵌套超节点的参数集不适用于它们的父级超节点。

为超节点创建和设置参数分为两个步骤：

1. 为超节点定义参数。
2. 然后，为超节点的每个参数指定值。

然后，可在任何封装节点的 CLEM 表达式中使用这些参数。

定义超节点参数

在缩小模式和放大模式下均可定义超节点的参数。 所定义的参数适用于所有封装节点。 要定义超节点的参数，首先需要访问“超节点”对话框的“参数”选项卡。 使用下列方法之一可打开该对话框：

- 双击流中的超节点。
- 从“超节点”菜单中，选择**设置参数**。
- 或者，当进行放大以向超节点推进时，请从上下文菜单中选择 **设置参数** 。

一旦打开了对话框，“参数”选项卡便会显示，其中包含以前定义的所有参数。

定义新的参数

单击 **定义参数** 按钮以打开对话框。

名称。 参数名在这里列出。 可以通过在本字段中输入名称来创建新的参数。 例如，要为最小温度创建参数，可以键入 `minvalue` 。 请勿包含表示 CLEM 表达式中的参数的 `$P-` 前缀。 该名称也用于在 CLEM 表达式构建器中显示。

长名称。 列出每个所创建参数的描述性名称。

存储器。 在列表中选择存储类型。 存储类型表明数据值在参数中如何存储。 例如，当所使用的值包含希望保留的先导 0 时（例如 008），应选择 **字符串**作为存储类型。 否则，先导 0 将从值中剥离。 有效的存储类型为字符串、整数、实数、时间、日期及时间戳 。 对于日期参数，注意其值必须用下一段落所示的 ISO 标准符号指定。

值。 列出每个参数的当前值。 根据需要调整参数。 请注意，对于日期参数，必须以 ISO 标准表示法（即，`YYYY-MM-DD`）来指定值。不接受以其他格式指定的日期。

类型（可选）。 如果计划将该流部署到外部应用程序，则从列表中选择测量级别。 否则，建议保留 类型列的值。 如果您想为参数指定值约束（如数值范围的上限或下限），请从列表中选择**指定**。

注意只能通过用户界面为参数设置长名称、存储类型和类型选项。 不能用脚本设置这些选项。

单击位于右侧的箭头可使选中的参数在可用参数列表中上下移动。 使用删除按钮（标记为 *X*）可删除选中的参数。

设置超节点参数的值

一旦为超节点定义了参数，便可以使用 CLEM 表达式或脚本中的参数指定值。

指定超节点的参数

1. 双击“超节点”图标以打开“超节点”对话框。
2. 或者，从“超节点”菜单中，选择**设置参数**。
3. 单击 **参数** 选项卡。 注意：此对话框中的字段是通过单击此选项卡上的**定义参数**按钮定义的字段。
4. 在文本框中为您创建的每个参数输入值。 例如，可以将 *minvalue* 值设置为相关的特定阈值。 然后，可以在许多操作中使用此参数，如果选择高于或低于此阈值的记录以进行进一步的研究。

使用超节点参数访问节点属性

超节点参数也可用于定义封装节点的节点属性（也称为 **通道参数** ）。 例如，假设您要指定某超节点使用可用的随机抽样数据对封装在其中的“神经网络”节点进行适当时间长度的训练。 使用参数，可指定时间长度和抽样百分比的值。

假设此示例超节点包含名为抽样的“样本”节点和名为训练的“神经网络”节点。 您可以使用节点对话框将样本节点的 **抽样** 设置指定为 **随机 %** ，将神经网络节点的 **停止条件** 设置指定为 **时间** 。 一旦指定了这些选项，便可使用参数访问节点属性并为超节点指定特定值。 在“超节点”对话框中，单击**定义参数**并创建如下表中所示的参数。

表 *53:* 要创建的参数

参数	值	长名称
Train.time	5	要训练的时间（分钟）
Sample.random	10	随机抽样百分比

注意：参数名称（如 *Sample.random*）使用正确的语法引用节点属性，其中 *Sample* 表示节点的名称，*random* 是节点属性。

定义了这些参数后，可以轻松地修改“样本”节点和“神经网络”节点属性的值，而不必重新打开每个对话框。相反，只需从“超节点”菜单中选择**设置参数**便可以访问“超节点”对话框的“参数”选项卡，在该选项卡上，可为**随机 %**和**时间**”指定新值。 这有助于在模型构建的多次迭代期间研究数据。

超节点和缓存

从超节点中，可以缓存除终端节点以外的所有节点。 通过右键单击节点并从“缓存”上下文菜单中选择其中一个选项，以控制缓存。 此菜单选项对超节点外部以及封装在超节点内的节点均适用。

以下是一些用于超节点缓存的指导准则：

- 如果任何封装在超节点中的节点启用了缓存，则超节点也会启用缓存。
- 禁用超节点缓存将禁用 所有 被封装节点的缓存。
- 启用超节点缓存会实际启用最后一个可缓存子节点的缓存。 换而言之，如果最后一个子节点是“选择”节点，那么将对该“选择”节点启用缓存。 如果最后一个子节点是终端节点（不允许缓存），则将启用比邻的支持缓存的上游节点。
- 一旦为超节点的子节点设置了缓存，则该缓存节点中的所有上游活动（如添加或编辑节点）都将刷新缓存。

超节点和脚本编写

您可以使用 SPSS Modeler 脚本编写语言来编写操纵并执行终端超节点内容的简单程序。 例如，您可能想要指定复杂流的执行顺序。 例如，如果超节点包含需要在“散点图”节点之前执行的“设置全局量”节点，那么可以创建首先执行“设置全局量”节点的脚本。 接着，可在执行“散点图”节点时使用此节点所计算的值（例如，平均值或标准偏差）。

“超节点”对话框的“脚本”选项卡只适用于终端超节点。

要为终端超节点打开“脚本编写”对话框，请执行以下操作：

- 右键单击超节点工作区，然后选择 **超节点脚本** 。
- 或者，在放大模式和缩小模式下，均可以从“超节点”菜单中选择**超节点脚本**。

注: 仅当在对话框中选定**运行此脚本**时，才会对流和超节点执行超节点脚本。

《脚本编写与自动化指南》中讨论了特定脚本编写选项及其在 SPSS Modeler 内的使用，该文档在产品下载过程中以 PDF 文件形式提供。

保存和加载超节点

超节点的其中一个优点为：可将它们保存下来并在其他流中复用。 保存和装入超节点时，请注意它们使用 `.slb` 扩展。

保存超节点

1. 放大超节点。
2. 从“超节点”菜单中，选择**保存超节点**。
3. 在对话框中指定文件名和目录。
4. 选择是否要将已保存的超节点添加到当前项目中。
5. 单击**保存**。

加载超节点

1. 从 IBM SPSS Modeler 窗口中的“插入”菜单中，选择**超节点**。
2. 从当前目录中选择超节点文件 (`.slb`) 或浏览至另一个文件。
3. 单击**装入**。

注: 所导入的超节点具有其所有参数的缺省值。 要更改参数，请双击流工作区上的超节点。

注意事项

本信息是为在美国提供的产品和服务编写的。 IBM 可能会提供其他语言形式的本资料。 但是，您可能需要拥有该语言的产品副本或产品版本，才能对其进行访问。

IBM 可能在其他国家或地区不提供本文档中讨论的产品、服务或功能。 有关您所在区域当前可获得的产品和服务的信息，请向您当地的 IBM 代表咨询。 任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。 只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。 但是，评估和验证任何非 IBM 产品、程序或服务，则由用户自行负责。

IBM 可能已拥有或正在申请与本文档内容有关的各项专利。 提供本文档并不意味着授予用户使用这些专利的任何许可。 您可以以书面形式将许可查询寄往：

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

有关双字节 (DBCS) 信息的许可查询，请与您所在国家或地区的 IBM 知识产权部门联系，或以书面形式将查询寄往：

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION“按现状”提供本出版物，不附有任何种类的（无论是明示的还是暗含的）保证，包括但不限于暗含的有关非侵权、适销和适用于某种特定用途的保证。 某些管辖区域在某些交易中不允许免除明示或暗含的保证。因此本条款可能不适用于您。

本信息中可能包含技术方面不够准确的地方或印刷错误。 此处的信息将定期更改；这些更改将编入本资料的新版本中。 IBM 可以随时对本出版物中描述的产品和/或程序进行改进和/或更改，而不另行通知。

本信息中对非 IBM Web 站点的任何引用都只是为了方便起见才提供的，不以任何方式充当对那些 Web 站点的保证。 那些 Web 站点中的资料不是本 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的被许可方如果要了解有关程序的信息以达到如下目的：(i) 允许在独立创建的程序和其他程序（包括本程序）之间进行信息交换，以及 (ii) 允许对已经交换的信息进行相互使用，请与下列地址联系：

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

只要遵守适当的条件和条款，包括某些情形下的一定数量的付费，都可获得这方面的信息。

本文档中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际程序许可协议或任何同等协议中的条款提供。

所引用的性能数据和客户示例仅作说明用途。 实际的性能结果可能会因特定的配置和运营条件而异。

涉及非 IBM 产品的信息是从这些产品的供应商、已出版说明或其他可公开获得的资料中获取。 IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。 有关非 IBM 产品性能的问题应当向这些产品的供应商提出。

关于 IBM 未来方向或意向的声明都可随时更改或收回，而不另行通知，它们仅仅表示了目标和意愿而已。

本信息包含在日常业务操作中使用的数据和报告的示例。 为了尽可能完整地说明这些示例，示例中可能会包括个人、公司、品牌和产品的名称。 所有这些名字都是虚构的，若与实际个人或业务企业相似，纯属巧合。

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp., 在全球许多管辖区域注册的商标或注册商标。 其他产品和服务名称可能是 IBM 或其他公司的商标。 IBM 商标的最新列表可从 Web 上的“Copyright and trademark information”处获得，网址为：www.ibm.com/legal/copytrade.shtml。

Adobe、Adobe 徽标、PostScript 以及 PostScript 徽标是 Adobe Systems Incorporated 在美国和／或其他国家或地区的注册商标或商标。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国或其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国和\或其他国家或地区的注册商标。

Microsoft、Windows、Windows NT 和 Windows 徽标是 Microsoft Corporation 在美国和/或其他国家或地区的商标。

UNIX 是 The Open Group 在美国和其他国家或地区的注册商标。

Java 和所有基于 Java 的商标和徽标是 Oracle 和/或其子公司的商标或注册商标。

产品文档的条款和条件

根据以下条款和条件授予这些出版物的使用许可权。

适用性

这些条款和条件是对 IBM Web 站点的任何使用条款的补充。

个人使用

您可以复制这些出版物供个人非商业性使用，但前提是保留所有专有权声明。 未经 IBM 明确同意，您不可以分发、展示或制作这些出版物或其中任何部分的演绎作品。

商业性使用

您仅可在贵公司内部复制、分发和显示这些出版物，但前提是保留所有专有权声明。 未经 IBM 明确同意，您不可以制作这些出版物的演绎作品，或者在您的企业外部复制、分发或展示这些出版物或其中的任何部分。

权利

除非本许可权中明确授予，否则不得授予对这些出版物或其中包含的任何信息、数据、软件或其他知识产权的任何许可权、许可证或权利，无论明示的还是暗含的。

只要 IBM 认为这些出版物的使用会损害其利益或者 IBM 判定未正确遵守上述指示信息，IBM 将有权撤销本文授予的许可权。

只有您完全遵循所有适用的法律和法规，包括所有的美国出口法律和法规，您才可以下载、出口或再出口该信息。

IBM 对这些出版物的内容不作任何保证。 这些出版物“按现状”提供，不附有任何种类的（无论是明示的还是暗含的）保证，包括但不限于暗含的有关适销性、非侵权和适用于某种特定用途的保证。

词汇表

C

协方差 *(Covariance)*

两个变量间关联性的非标准化测量值，等于叉积偏差除以 N-1。

K

峰度 *(Kurtosis)*

存在离群值的程度的测量。 对于正态分布，峰度统计量的值为零。 正峰度值表示数据呈现比正态分布更极端的离群值。 负峰度值表示数据呈现比正态分布极端程度较低的离群值。

M

Maximum

数值变量的最大值。

平均值

集中趋势的测量。 算术平均值，等于总和除以观测值数。

中位数

大于或小于一半观测值的值，即 50th 个百分位。 如果有偶数个观测值，则中位数为它们以升序或降序排列时两个中间观测值的平均值。 中位数是集中趋势的一种测量，对离群值不敏感（与平均值不同，平均值会受部分极高或极低值的影响）。

Minimum

数值变量的最小值。

众数 *(Mode)*

最常出现的值。 如果多个值共享最大出现频率，则每个值都是一个众数。

R

范围

数字变量的最大值与最小值的差值就是用最大值减最小值后得出的值。

S

偏度 *(Skewness)*

分布不对称性的测量。 正态分布是一种对称性分布，其偏度值为 0。 具有显著性正偏度的分布右侧尾部较长。 具有显著负偏态的分布具有向左延伸的长尾。 提示：取大于其标准误差两倍的偏度值指示离开对称的距离。

标准差

围绕平均值的离差的测量，等于方差的平方根。 以和原始变量相同的单位度量标准差。

标准差

对围绕平均值的离差的测量。 在正态分布中，68% 的观测值落入与均值相距不到一个标准差的范围内，95% 落入两个标准差的范围内。 例如，如果平均年龄值 45，标准差为 10，则 95% 的观测值将介于正态分布的 25 到 65 之间。

标准错误 *(Standard Error)*

检验统计量值因样本而异的测量。 这是统计量抽样分布的标准差。 例如，均数的标准误差是样本均数的标准差。

峰度标准误差 *(Standard Error of Kurtosis)*

峰度与其标准误差的比率可用作正态性检验 (即，如果比率小于 -2 或大于 + 2 ，那么可以拒绝正态性)。 峰度较大的正值表示该分布的尾部比正态分布的尾部长；峰度的负值表示较短的尾部（与箱形均匀分布的尾部变得相似）。

均值标准误差 *(Standard Error of Mean)*

对取自相同分布的样本之间的平均值可能有多大差异的测量。 用于粗略将观测到的均数与假设值对比（即，如果差异与标准误差的比率小于 -2 或大于 +2，则可以得出此均数与假设值不同的结论）。

偏度标准误差 *(Standard Error of Skewness)*

偏度与其标准误差的比率可用作正态性检验 (即，如果该比率小于 -2 或大于 + 2 ，那么可以拒绝正态性)。 偏度正值越大表示长尾向右越长；负极值表示向左的长尾。

Sum

所有带有非缺失值的观测值的值的合计或总计。

U

UNIQUE

同步评估所有效应，同时为任意类型的所有其他效应调整每一个效应。

V

有效

有效观测值既不包含系统缺失值，也不包含定义为用户缺失的值。

偏差

对围绕平均值的离差的测量，值等于与平均值的差的平方和除以个案数减一。 方差按单元计量，即变量自身单元数的平方。

索引

Special Characters

Numerics

T

U

V

W

X