

IBM SPSS Modeler 16 源、过程和输出节点

IBM

注释

在使用本资料及其支持的产品之前，请阅读第 301 页的『声明』中的信息。

产品信息

此版本适用于 IBM(r) SPSS(r) Modeler V16R0M0 以及所有后续发行版和修改版，直到在新版本中另有声明为止。

目录

前言	vii	导入 Cognos 数据	29
第 1 章 关于 IBM SPSS Modeler	1	导入 Cognos 报告	30
IBM SPSS Modeler 产品	1	Cognos 连接	30
IBM SPSS Modeler	1	Cognos 位置选择	30
IBM SPSS Modeler Server	1	指定数据或报告参数	31
IBM SPSS Modeler Administration Console	2	IBM Cognos TM1 源节点	31
IBM SPSS Modeler Batch	2	导入 Cognos TM1 数据	31
IBM SPSS Modeler Solution Publisher	2	SAS 源节点	32
IBM SPSS Collaboration and Deployment Services 的 IBM SPSS Modeler Server 适配器	2	为 SAS 源节点设置选项	32
IBM SPSS Modeler 版本	2	Excel 源节点	32
IBM SPSS Modeler 文档	3	XML 源节点	33
SPSS Modeler Professional 文档	3	从多个根元素中选择	34
SPSS Modeler Premium 文档	4	从 XML 源数据中除去不需要的空格	34
应用程序示例	4	“模拟生成”节点	35
Demos 文件夹	4	为“模拟生成”节点设置选项	36
第 2 章 源节点	5	克隆字段	40
概述	5	拟合详细信息	40
企业视图节点	6	指定参数	41
为企业视图节点设置选项	7	分布	43
Enterprise View 连接	7	用户输入节点	45
选择 DPD	8	为用户输入节点设置选项	45
选择表	8	公共源节点选项卡	48
“数据视图”节点	9	在源节点中设置测量级别	48
为“数据视图”节点设置选项	9	从源节点中过滤字段	49
数据库源节点	10	第 3 章 记录操作节点	51
设置数据库节点选项	11	记录操作概述	51
添加数据库连接	11	选择节点	52
为数据库连接指定预设值	12	样本节点	53
选择数据库表	14	样本节点选项	53
查询数据库	15	聚类和分层设置	55
变量文件节点	16	层的样本大小	56
为变量文件节点设置选项	16	平衡节点	56
固定文件节点	18	为平衡节点设置选项	57
为固定文件节点设置选项	18	汇总节点	57
设置字段存储类型和格式	19	为汇总节点设置选项	58
Data Collection 节点	20	RFM 汇总节点	59
Data Collection 导入文件选项	21	为 RFM 汇总节点设置选项	60
IBM SPSS Data Collection 导入元数据属性	23	排序节点	60
数据库连接字符串	23	排序优化设置	61
高级属性	23	合并节点	61
导入多重响应集	23	连接类型	61
IBM SPSS Data Collection 列导入说明	24	指定合并方法和关键字	63
Analytic Server 源	24	选择用于部分连接的数据	64
Analytic Server 选择数据源	24	指定合并的条件	64
Analytic Server 凭证	25	过滤合并节点中的字段	64
受支持的节点	25	设置输入顺序和标记	65
IBM Cognos BI 源节点	28	合并优化设置	65
Cognos 对象图标	28	追加节点	66
		设置追加选项	66
		区分节点	67

区分优化设置	68	固定宽度分级	115
区分组合设置	69	分位数 (相等计数或总和)	116
“流式 TS”节点	71	个案排秩	117
流式 TS 字段选项	71	均数/标准差	117
流式 TS 模型选项	72	最优分级	118
时间序列专家建模器标准	73	预览生成的分级	119
时间序列指数平滑法标准	73	RFM 分析节点	119
时间序列 ARIMA 标准	74	RFM 分析节点设置	119
转换函数	75	RFM 分析节点分级	120
处理离群值	75	分区节点	121
流式 TS 部署选项	76	分区节点选项	121
“空间时间限制”节点	76	设为标志节点	122
定义空间时间限制密度	77	为设为标志节点设置选项	122
第 4 章 字段操作节点 79		重新结构化节点	122
字段操作概述	79	为重新结构化节点设置选项	123
自动准备数据	81	转置节点	124
“字段”选项卡	82	为转置节点设置选项	124
“设置”选项卡	82	时间区间节点	125
“分析”选项卡	86	指定时间间隔	126
生成派生节点	92	时间间隔构建选项	126
类型节点	92	估计期	127
测量级别	93	预测值(S)	127
转换连续数据	95	支持的间隔	128
什么是实例化?	95	历史记录节点	134
数据值	96	为历史记录节点设置选项	135
定义缺失值	98	字段重排节点	135
检查类型值	98	设置字段重排选项	135
设置字段角色	99	第 5 章 图形节点 137	
复制类型属性	99	通用图形节点功能	137
字段格式设置选项卡	100	审美原则、交叠、面板和动画	138
过滤或重命名字段	101	使用“输出”选项卡	139
设置过滤选项	101	使用“注释”选项卡	140
整体节点	104	3D 图形	140
整体节点设置	104	图形板节点	141
导出节点	105	图形板基本选项卡	141
为导出节点设置基本选项	106	图形板 详细选项卡	144
导出多个字段	107	可用的内置图形板可视化类型	145
设置导出公式选项	107	创建地图直观表示	150
设置导出标志选项	107	图形板示例	151
设置派生名义选项	108	“图形板外观”选项卡	161
设置导出状态选项	108	设置模板、样式表和地图的位置	162
设置导出计数选项	109	管理模板、样式表和地图文件	163
设置导出条件选项	109	转换和分发地图 Shapefile	164
使用导出节点对值进行重新编码	109	地图的关键概念	164
填充节点	109	使用地图转换实用程序	165
使用填充节点进行存储类型转换	110	分发地图文件	169
匿名化节点	110	散点图节点	169
为匿名化节点设置选项	111	散点图节点选项卡	172
匿名化字段值	112	散点图选项选项卡	173
重新分类节点	112	散点图外观选项卡	174
为重新分类节点设置选项	113	使用散点图	174
对多个字段进行重新分类	114	分布节点	175
重新分类字段的存储类型和测量级别	114	分布图选项卡	175
分级节点	114	分布外观选项卡	176
为分箱节点设置选项	115	使用分布节点	176

直方图节点	179
直方图选项卡	179
直方图选项选项卡	179
直方图外观选项卡	179
使用直方图	180
收集节点	180
收集散点图选项卡	181
收集选项选项卡	181
收集外观选项卡	181
使用集合图	182
多重散点图节点	183
多重散点图选项卡	183
多重散点图外观选项卡	185
使用多重散点图	185
网络节点	186
网络散点图选项卡	187
网络选项选项卡	188
网络外观选项卡	189
使用网络图形	190
时间散点图节点	193
时间散点图选项卡	193
时间散点图外观选项卡	194
使用时间散点图	194
评估节点	195
评估散点图选项卡	199
评估选项选项卡	200
评估外观选项卡	201
读取模型评估结果	201
使用评估图表	202
探索图形	203
使用带状区域	204
使用区域	208
使用标记后的元素	210
从图形中生成节点	211
编辑直观表示	213
编辑直观表示的一般规则	214
编辑和格式化文本	215
更改颜色、模式、划线和透明度	216
旋转并更改点元素的形状和高宽比	216
更改图形元素的大小	217
指定边距和填充	217
设置数字格式	217
更改轴和刻度设置	218
编辑类别	219
更改方向面板	221
转换坐标系	221
更改统计量和图形元素	221
更改图例的位置	224
复制直观表示和直观表示数据	224
键盘快捷键	224
添加标题和脚注	224
使用图形样式表	225
打印、保存、复制和导出图形	226

第 6 章 输出节点 227

输出节点概述	227
--------	-----

管理输出	228
查看输出	228
发布到 Web	229
在 HTML 浏览器中查看输出	230
导出输出	230
选择单元格和列	231
表节点	231
表节点的“设置”选项卡	231
表节点的“格式”选项卡	231
输出节点的“输出”选项卡	231
表格浏览器	232
矩阵节点	233
矩阵节点的设置选项卡	233
矩阵节点的外观选项卡	234
矩阵节点的输出浏览器	234
分析节点	235
分析节点的分析选项卡	235
分析输出浏览器	236
数据审核节点	237
数据审核节点的设置选项卡	238
数据审核的质量选项卡	238
数据审核输出浏览器	239
变换节点	243
变换节点的选项选项卡	243
变换节点的输出选项卡	244
变换节点的输出查看器	244
统计量节点	246
统计量节点的设置选项卡	246
统计量输出浏览器	246
平均值节点	247
比较独立组的平均值	248
在成对字段之间比较平均值	248
平均值节点选项	248
平均值节点输出浏览器	248
报告节点	250
报告节点的模板选项卡	250
报告节点输出浏览器	251
设置全局量节点	251
设置全局量节点的设置选项卡	251
“模拟拟合”节点	252
分布拟合	252
“模拟拟合”节点的“设置”选项卡	253
“模拟评估”节点	254
“模拟评估”节点的“设置”选项卡	254
“模拟评估”节点的输出	256
IBM SPSS Statistics 帮助应用程序	259

第 7 章 导出节点 261

导出节点概述	261
数据库导出节点	262
数据库节点的“导出”选项卡	262
数据库导出合并选项	263
数据库导出模式选项	264
数据库导出索引选项	266
数据库导出高级选项	267
批量加载程序设计	268

平面文件导出节点	274
“平面文件导出”选项卡	274
IBM SPSS Data Collection 导出节点	275
Analytic Server 导出	276
IBM Cognos BI 导出节点	276
Cognos 连接	276
ODBC 连接	277
IBM Cognos TM1 导出节点	278
连接到 Cognos TM1 多维数据集可导出数据	278
映射可供导出的 Cognos TM1 数据	278
SAS 导出节点	279
SAS 导出节点“导出”选项卡	279
Excel 导出节点	279
Excel 节点“导出”选项卡	279
XML 导出节点	280
写入 XML 数据	280
XML 映射记录选项	281
XML 映射字段选项	281
XML 映射预览	281
第 8 章 IBM SPSS Statistics 个节点 283	
IBM SPSS Statistics 节点 - 概述	283
Statistics 文件节点	284
Statistics 转换节点	285
Statistics 转换节点 - “语法”选项卡	285
允许的语法	285
Statistics 模型节点	287
Statistics 模型节点 - “模型”选项卡	287
Statistics 模型节点 - 模型块汇总	288
Statistics 输出节点	288

Statistics 输出节点 - “语法”选项卡	288
Statistics 输出节点 - “输出”选项卡	289
Statistics 导出节点	290
Statistics 导出节点 - “导出”选项卡	290
重命名或过滤 IBM SPSS Statistics 的字段	291

第 9 章 超节点 **293**

超节点概述	293
超节点的类型	293
源超节点	293
过程超节点	293
终端超节点	294
创建超节点	294
嵌套超节点	295
锁定超节点	295
锁定和解锁超节点	295
编辑锁定的超节点	295
编辑超节点	296
修改超节点类型	296
添加注解和重命名超节点	296
超节点参数	297
超节点和缓存	298
超节点和脚本编写	299
保存和加载超节点	299

声明 **301**

商标	302
--------------	-----

索引 **305**

前言

IBM® SPSS® Modeler 是 IBM Corp. 企业级数据挖掘工作平台。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 Corp. 产品使用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 Corp. 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，也可以作为客户机与 SPSS 建模器服务器一起使用。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<http://www.ibm.com/software/analytics/spss/products/modeler/>。

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
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Collaboration and Deployment Services 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。请参阅主题第 2 页的『IBM SPSS Modeler 版本』，了解更多信息。

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，因而使数据集的传输速度大大加快。

SPSS 建模器服务器 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 建模器服务器 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 是一个图形应用程序，用于管理多个 SPSS 建模器服务器 配置选项，这些选项还可以通过选项文件进行配置。此应用程序提供了一个用于监视和配置 SPSS 建模器服务器 安装的控制台用户界面，并且可供当前 SPSS 建模器服务器 用户免费使用。应用程序只能安装在 Windows 计算机上；但是它可以在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。使用 SPSS Modeler Batch 需要获得 SPSS 建模器服务器 许可。

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher 是一个支持创建 SPSS Modeler 流的打包版本的工具，该版本的流可以由外部运行时引擎运行或嵌入到外部应用程序中。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 Collaboration and Deployment Services - Scoring 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。

IBM SPSS Collaboration and Deployment Services 的 IBM SPSS Modeler Server 适配器

IBM SPSS Collaboration and Deployment Services 的一些适配器使 SPSS Modeler 和 SPSS 建模器服务器 能够与 IBM SPSS Collaboration and Deployment Services 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，或者从瘦客户机应用程序 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（例如用于实体分析或社交网络的数据）和非结构化文本数据。SPSS Modeler Premium 由下列组件组成。

IBM SPSS Modeler Entity Analytics 将额外维度添加到 IBM SPSS Modeler 预测性分析中。预测性分析尝试根据过去的行为预测未来的行为，而实体分析侧重于通过在记录本身中解决身份冲突改进当前数据的连贯性和一致性。身份可以指个人、组织、对象或可能不确定的任何其他实体的身份。解析身份在多个领域十分关键，包括客户关系管理、内部欺诈检测、反洗钱以及国家和国际安保。

IBM SPSS Modeler Social Network Analysis 将关于关系的信息转换为字段，这些字段可描述个人和组社交行为的特征。使用介绍社交网络之下关系的数据，IBM SPSS Modeler Social Network Analysis 可识别影响

网络中他人行为的社交领导。另外，您可以确定哪些人员受其他网络参与者的影响最大。通过这些结果与其他度量结合，您可以创建作为预测模型基础的复杂个人档案。包含此社交信息的模型比不包含此社交信息的模型执行地更好。

IBM SPSS Modeler Text Analytics 采用了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和现有结构化数据中进行组合（例如人口统计学），并且可用于借助 IBM SPSS Modeler 的一整套数据挖掘工具来进行建模，以此实现更好更集中的决策。

IBM SPSS Modeler 文档

可以从 SPSS Modeler 的帮助菜单中获取在线帮助格式的文档。此文档包括 SPSS Modeler、SPSS 建模器服务器和 SPSS Modeler Solution Publisher 的文档以及《应用程序指南》和其他支持材料。

每个产品 DVD 的 \Documentation 文件夹下都提供了该产品的 PDF 格式的完整文档（包括安装指示信息）。安装文档也可从 <http://www-01.ibm.com/support/docview.wss?uid=swg27038316> 站点上下载。

还可以从 SPSS Modeler 信息中心 (<http://publib.boulder.ibm.com/infocenter/spssmodl/v16r0m0/>) 获取这两种格式的文档。

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**。使用 SPSS Modeler 的一般使用介绍，包括如何构建数据流、处理缺失值、生成 CLEM 表达式、处理项目和报告以及将用于部署的流打包为 IBM SPSS Collaboration and Deployment Services、预测应用程序或 IBM SPSS Modeler Advantage。
- **IBM SPSS Modeler 源、过程和输出节点**。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 建模节点**。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 Algorithms Guide**。描述 IBM SPSS Modeler 中使用的建模方法的数学基础。本指南仅以 PDF 格式提供。
- **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。还可以在“帮助”菜单中查阅本指南的在线版本。请参阅主题第 4 页的『应用程序示例』，了解更多信息。
- **IBM SPSS Modeler 脚本编制和自动化**。通过编写脚本实现系统自动化的相关信息，包括用于操作节点和流的属性信息。
- **IBM SPSS Modeler 部署指南**。有关在 IBM SPSS Collaboration and Deployment Services Deployment Manager 中以处理作业的步骤形式运行 IBM SPSS Modeler 流和方案的信息。
- **IBM SPSS Modeler CLEF 开发者指南**。CLEF 提供了将第三程序（例如，数据处理例程或建模算法）作为节点集成到 IBM SPSS Modeler 的功能。
- **IBM SPSS Modeler 数据库内数据挖掘指南**。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 管理与性能指南**。提供有关如何配置和管理 IBM SPSS Modeler Server 的信息。
- **IBM SPSS Modeler Administration Console User Guide**。有关安装和使用控制台用户界面以监视和配置 IBM SPSS Modeler Server 的信息。控制台实现为部署管理器应用程序的插件。
- **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。

- **IBM SPSS Modeler Entity Analytics 用户指南**。提供有关将实体分析与 SPSS Modeler 配合使用的信息，包括存储库安装和配置、实体分析节点和管理任务。
- **IBM SPSS Modeler Social Network Analysis 用户指南**。本指南指导使用 SPSS Modeler 执行社交网络分析，包括组分析和传播分析。
- **SPSS Modeler Text Analytics 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。
- **IBM SPSS Modeler Text Analytics Administration Console User Guide**。提供有关安装和使用控制台用户界面以监视和配置 IBM SPSS Modeler Server，以便与 SPSS Modeler Text Analytics 配合使用的信息。控制台实现为 部署管理器 应用程序的插件。

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要小得多，但涉及的概念和方法应可扩展到实际的应用程序。

可以通过在 SPSS Modeler 中的“帮助”菜单中单击**应用程序示例**来访问示例。数据文件和样本流安装在产品安装目录下的 *Demos* 文件夹中。请参阅主题『[Demos 文件夹](#)』，了解更多信息。

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

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

Demos 文件夹

与应用程序示例一起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。可以从 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组访问此文件夹，也可以在“打开文件”对话框中的最近目录的列表中单击 *Demos* 来进行访问。

第 2 章 源节点

概述

使用源节点能够导入以多种格式存储的数据，这些格式包括平面文件、IBM SPSS Statistics (.sav)、SAS、Microsoft Excel 和 ODBC 兼容关系数据库。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：



Enterprise View 节点用于创建指向 IBM SPSS Collaboration and Deployment Services Repository 的连接，使您可以将 Enterprise View 数据读入流中，并将模型打包装入其他用户可通过存储库访问的方案。请参阅主题第 6 页的『企业视图节点』，了解更多信息。



“数据视图”节点可用于访问 IBM SPSS Collaboration and Deployment Services 分析数据视图中定义的数据源。分析数据视图定义了一个用于访问数据的标准界面，并将多个物理数据源与该界面进行关联。请参阅主题第 9 页的『“数据视图”节点』，了解更多信息。



“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。请参阅主题第 10 页的『数据库源节点』，了解更多信息。



“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。请参阅主题第 16 页的『变量文件节点』，了解更多信息。



“固定文件”节点会从固定字段文本文件（即文件字段不定界，而是从同一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。请参阅主题第 18 页的『固定文件节点』，了解更多信息。



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 中的高速缓存文件（也使用同一格式）读取数据。



IBM SPSS Data Collection 节点从符合 IBM SPSS Data Collection 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 IBM SPSS Data Collection Developer Library 才可使用此节点。请参阅主题第 20 页的『Data Collection 节点』，了解更多信息。



IBM Cognos BI 源节点从 Cognos BI 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。请参阅主题第 32 页的『SAS 源节点』，了解更多信息。



Excel 节点可以从任何版本的 Microsoft Excel 中导入数据。不要求指定 ODBC 数据源。请参阅主题第 32 页的『Excel 源节点』，了解更多信息。



XML 源节点将 XML 格式的数据导入到流中。可以导入某个目录中的单个文件或所有文件。还可选择指定模式文件，以从中读取 XML 结构。



“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。对于模型输入中存在不确定性的情况，此节点在对预测模型的结果进行评估时非常有用。



“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。请参阅主题第 45 页的『用户输入节点』，了解更多信息。

要开始执行流，可将源节点添加到流画布中。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。

企业视图节点

使用 Enterprise View 节点，可以在共享的 IBM SPSS Collaboration and Deployment Services Repository 中创建并维护 IBM SPSS Modeler 会话和 Enterprise View 之间的连接。这样，您可以将 Enterprise View 中的数据读入 IBM SPSS Modeler 流，还可以将 IBM SPSS Modeler 模型打包至方案（共享存储库的其他用户可访问该方案）中。

方案是包含具有特定节点、模型和其他属性的 IBM SPSS Modeler 流的文件，它可以部署到 IBM SPSS Collaboration and Deployment Services Repository 以进行评分或自动模型刷新。将“企业视图”节点用于方案可确保多用户情况下的所有用户使用的是同一数据。**连接**是从 IBM SPSS Modeler 会话到 IBM SPSS Collaboration and Deployment Services Repository 中的 Enterprise View 的链接。

Enterprise View是属于某个组织的数据全集（不考虑数据的物理位置）。每个连接都包括以下各项的特定选择内容：一个 **Application View**（为特定应用而裁剪后的 Enterprise View 的子集）、一个**数据提供者定义**（DPD

- 将逻辑 Application View 图表和列链接到物理数据源) 和一个环境 (确定应与定义的商业段关联的特定列)。尽管实际数据位于一个或多个数据库或其他外部来源中, 但企业视图、应用程序视图和 DPD 的定义均在存储库中存储和描述。

建立连接后, 便可以指定要在 IBM SPSS Modeler 中使用的 Application View 表。在应用程序视图中, 表是一种逻辑视图, 其中包含来自一个或多个物理数据库中的一个或多个物理表的某些或所有列。因此, 通过 Enterprise View 节点可以将来自多个数据库表的记录视为 IBM SPSS Modeler 中的一个表。

要求

- 要使用 Enterprise View 节点, 必须首先在本地安装并配置 IBM SPSS Collaboration and Deployment Services Repository, 并定义 Enterprise View、Application View 和 DPD。

注意: 访问 IBM SPSS Collaboration and Deployment Services 存储库需要单独许可证。有关更多信息, 请参阅<http://www.ibm.com/software/analytics/spss/products/deployment/cds/>。

- 此外, 必须在用于修改或运行流的每台计算机上安装 IBM SPSS Collaboration and Deployment Services Enterprise View Driver。对于 Windows, 只需在已安装 IBM SPSS Modeler 或 IBM SPSS Modeler Server 的计算机上安装该驱动程序即可, 无需进行任何进一步的配置。在 UNIX 中, 则必须在启动脚本中添加对 `pev.sh` 脚本的引用。有关安装 IBM SPSS Collaboration and Deployment Services Enterprise View Driver 的详细信息, 请与本地管理员联系。
- DPD 是根据特定的 ODBC 数据源定义的。要通过 IBM SPSS Modeler 使用 DPD, 必须在某个 IBM SPSS Modeler 服务器主机上定义同名的 ODBC 数据源, 该主机还必须连接到 DPD 中引用的数据存储。

为企业视图节点设置选项

可以使用“企业视图”对话框的“数据”选项卡中的选项进行下列操作:

- 选择现有的存储库连接
- 编辑现有的存储库连接
- 新建存储库连接
- 选择应用程序视图表

有关使用存储库的详细信息, 请参阅 *IBM SPSS Collaboration and Deployment Services* 管理者指南。

连接。根据下拉列表提供的选项可以选择现有的存储库连接、编辑现有连接或添加连接。如果已通过 IBM SPSS Modeler 登录到存储库, 选择**添加/编辑连接**选项将显示“Enterprise View 连接”对话框, 您可在其中定义或编辑当前连接所需的详细信息。如果您未登录, 那么此选项将显示存储库的“登录”对话框。

有关登录到存储库的信息, 请参阅 *IBM SPSS Modeler 用户指南*。

建立指向存储库的连接后, 该连接将保持不变, 直到您从 IBM SPSS Modeler 中退出。同一流中的其他节点可以共享一个连接, 但必须为每个新流创建新连接。

成功登录后, 将显示“企业视图连接”对话框。

表名称。此字段最初是空白字段, 并且在创建连接之后才可被填充。如果知道要访问的应用程序视图表的名, 可在表名字段中输入此名称。否则, 可单击**选择按钮**打并列出了可用的 Application View 表的对话框。

Enterprise View 连接

通过此对话框, 可以定义或编辑存储库连接所需的详细信息。可指定下列内容:

- 应用程序视图和版本

- 环境
- 数据提供者定义 (DPD)
- 连接说明

连接。列出现有存储库连接。

- **添加新连接。**显示“检索对象”对话框，您可在其中搜索和选择存储库中的 Application View。
- **复制选定的连接。**生成所选连接的副本，使您无需再次浏览同一 Application View。
- **删除选定的连接。**从列表中删除选定的连接。

连接详细信息。针对“连接”面板中当前选定的连接，显示 Application View、版本标签、环境、DPD 和说明文本。

- **应用程序视图。**下拉列表将显示所选的应用程序视图（如果有）。如果已连接到当前会话中的其他应用程序视图，那么这些视图也将显示在下拉列表中。单击旁边的“浏览”按钮可搜索存储库中的其他应用程序视图。
- **版本标签。**在下拉字段中列出了为指定的应用程序视图定义的所有版本标签。版本标签可帮助识别特定的存储库对象版本。例如，某个特定的应用程序视图可能有两个版本。通过使用标签，可以为在开发环境中使用的版本指定标签 **TEST**，为在生产环境中使用的版本指定标签 **PRODUCTION**。选择合适的标签。

注意：标签不得包含“[”字符，否则，“企业视图”对话框的“数据”选项卡上将不会显示表名称。

- **环境。**下拉字段中列出了所有的有效环境。环境设置决定了可用的 DPD，从而指定应与已定义的商业段关联的特定列。例如，选中**分析**时，仅返回那些定义为**分析**的 Application View 图列。默认的环境是 **分析**；也可以选择 **操作**。
- **数据提供者。**下拉列表将显示所选应用程序视图的数据提供者定义（最多十个）的名称。仅显示引用所选应用程序视图的 DPD。单击旁边的“浏览”按钮可查看与当前应用程序视图相关的所有 DPD 的名称和路径。
- **描述。**有关存储库连接的描述性文本。此文本将用于连接的名称 - 单击**确定**将使此文本显示在“连接”下拉列表、“Enterprise View”对话框的标题栏以及工作区上的 Enterprise View 节点标签中。

选择 DPD

“选择数据提供者”对话框将显示引用当前应用程序视图的所有 DPD 的名称和路径。

应用程序视图可以有多个 DPD，以支持工程的不同阶段。例如，用于构建模型的历史记录数据可能来自一个数据库，而操作数据来自另一个数据库。

DPD 是根据特定的 ODBC 数据源定义的。要通过 IBM SPSS Modeler 使用 DPD，必须在某个 IBM SPSS Modeler 服务器主机上定义同名的 ODBC 数据源，该主机还必须连接到 DPD 中引用的数据存储。

要选择使用的 DPD，请在列表中选择其名称，然后单击 **确定**。

选择表

“选择表”对话框将列出当前应用程序视图中引用的所有表。如果未建立指向 IBM SPSS Collaboration and Deployment Services Repository 的连接，该对话框将为空。

要选择使用的表，请在列表中选择其名称，然后单击 **确定**。

“数据视图”节点

使用“数据视图”节点可以将 IBM SPSS Collaboration and Deployment Services 分析数据视图中定义的数据包括在流中。分析数据视图定义用于访问数据的结构，此结构描述了预测模型和业务规则中使用的实体。该视图将数据结构与物理数据源关联以进行分析。

预测性分析需要在表中对数据进行组织，表中的每行对应于一个对其进行预测的实体。表中的每列表示实体的一个可度量属性。某些属性可能通过汇总另一个属性的值进行派生。例如，表中的行可以表示客户，而列对应于客户名称、性别、邮政编码以及客户在过去一年内购买费用超出 \$500 的次数。最后一列派生自客户订单历史记录，该历史记录通常存储在一个或多个相关表中。

预测分析过程涉及在模型的整个生命周期内使用不同的数据集。在预测模型的初始开发过程中，可以使用历史数据，此数据通常已经知道要预测的事件的结果。要对模型的有效性和准确性进行评估，可以根据不同的数据验证候选模型。验证模型之后，将它部署到生产使用，以便为批处理过程中的多个实体或实时过程中的单个实体生成得分。如果您在决策管理过程中将模型与业务规则进行组合，那么可以使用模拟数据来验证该组合的结果。但是，尽管整个模型开发过程阶段所使用的数据不同，每个数据集都必须为模型提供同一属性集。此属性集将保持不变；要分析的数据记录将发生更改。

分析数据视图包含以下要素，这些要素能够满足预测性分析的特殊需求：

- 数据视图模式或数据模型，定义了一个接口，用于将数据作为组织到相关表中的一组数据进行访问。该模型中的属性可从其他属性派生而来。
- 一个或多个数据访问方案，这些方案提供了具有物理值的数据模型属性。通过指定哪些数据访问方案对于特定应用程序处于活动状态，您可以控制可用于数据模型的数据。

要点：要使用“数据视图”节点，必须先在其站点安装并配置 IBM SPSS Collaboration and Deployment Services Repository。此节点引用的分析数据视图通常是使用 IBM SPSS Collaboration and Deployment Services Deployment Manager 在存储库中创建和存储的。

为“数据视图”节点设置选项

使用“数据视图”节点对话框的**数据**选项卡上的选项可以为从 IBM SPSS Collaboration and Deployment Services Repository 中选择的分析数据视图指定数据设置。

分析数据视图。单击省略号按钮 (...) 可选择分析数据视图。如果您当前未连接到存储库服务器，请在“存储库：服务器”对话框中指定服务器的 URL，并单击**确定**，然后在“存储库：凭证”对话框中指定连接凭证。有关登录存储库并检索对象的更多信息，请参阅《IBM SPSS Modeler 用户指南》。

表名称。请从分析数据视图中的数据模型中选择一个表。数据模型中的每个表都表示预测分析过程中涉及的一个概念或实体。表的字段对应于这些表所表示的实体的属性。例如，如果要对客户订单进行分析，那么数据模型可以包括一个客户表和一个订单表。客户表可能包含客户标识、年龄、性别、婚姻状况和居住国家或地区等属性。订单表可能包含订单标识、订单中的商品数、总开销以及下单的客户的标识等属性。客户标识属性可用于将客户表中的客户与订单表中相应客户的订单进行关联。

数据访问方案。请从分析数据视图中选择一个数据访问方案。数据访问方案会将分析数据视图中的数据模型表与物理数据源进行关联。分析数据视图通常包含多个数据访问方案。更改使用中的数据访问方案时，您可以更改流所使用的数据。例如，如果分析数据视图包含用于训练模型的数据访问方案以及用于检验模型的方案，那么您可以通过更改使用中的数据访问方案从训练数据切换至检验数据。

可选属性。如果使用分析数据视图的应用程序不需要某个特定属性，那么您可以将此属性标记为可选。与必需属性不同，可选属性可以包含空值。您可能需要对应用程序进行调整，以包括可选属性的空值处理。例如，调用 IBM Operational Decision Manager 中创建的业务规则时，IBM Analytical Decision Management 将查询规

则服务以确定必需的输入。对于任何规则服务的必填字段，如果要进行评分的记录包含空值，那么将不会调用该规则，并且将使用缺省值填充该规则的输出字段。如果可选字段包含空值，那么将调用该规则。可以对该规则进行空值检查以控制处理。

要将属性指定为可选属性，请单击**可选属性**，然后选择可选属性。

在字段中包含 XML 数据。选择此选项可创建一个包含每行数据的可执行对象模型 XML 数据的字段。如果数据将与 IBM Operational Decision Manager 配合使用，那么此信息是必需的。请为此新字段指定名称。

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、Oracle 等。

要对数据库执行读取或写入，您必须为相关数据库安装并配置 ODBC 数据源，并根据需要配置读许可权或写许可权。IBM SPSS Data Access Pack 包括一组用于此用途的 ODBC 驱动程序，您可以从 IBM SPSS Data Access Pack DVD 或下载站点获取这些驱动程序。如果您有关于创建或设置 ODBC 数据源权限的问题，请联系您的数据库管理员。

支持的 ODBC 驱动程序

有关使用 IBM SPSS Modeler 16 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>)。

在哪里安装驱动程序

请注意，必须在每台可能进行处理的计算机上安装并配置 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，必须在两个计算机上都安装 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。

要点：如果您在 UNIX 上使用 IBM SPSS Modeler Server 访问 Teradata 数据库，并且使用的是 Teradata 提供的 ODBC 驱动程序，那么必须使用随此驱动程序一起安装的 ODBC 驱动程序管理器。（注：如果使用的是 SDAP Teradata 驱动程序，那么无需执行这些更改。）为了对 IBM SPSS Modeler Server 进行此更改，请在靠近 `modelersrv.sh` 脚本的顶部、注释所指示的地方为 `ODBC_DRIVER_MANAGER_PATH` 指定一个值。必须将此环境变量设置为 Teradata ODBC 驱动程序随附的 ODBC 驱动程序管理器所在的位置（在 Teradata ODBC 驱动程序缺省安装中，此位置为 `/usr/odbc/lib`）。您必须重新启动 IBM SPSS Modeler Server 以使所做更改生效。有关为 Teradata 访问提供支持的 IBM SPSS Modeler Server 平台以及支持的 Teradata ODBC 驱动程序版本的详细信息，请访问公司支持站点 <http://www.ibm.com/support>。

在 UNIX 系统中配置 ODBC 驱动程序

默认情况下，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
ln -s libspssodbc_datadirect.so libspssodbc.so
```

此命令可删除默认链接并新建至 DataDirect 驱动程序管理器的链接。

使用下列一般步骤访问数据库中的数据:

1. 为要使用的数据库安装 ODBC 驱动程序并配置数据源。
2. 在数据库节点对话框中, 使用表模式或 SQL 查询模式连接到数据库。
3. 从数据库中选择表。
4. 使用数据库节点对话框中的选项卡, 可以更改使用类型和过滤数据字段。

在下面的几个主题中将对这些步骤进行更详细地说明。

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。请参阅主题第 15 页的『查询数据库』, 了解更多信息。

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。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。

给表名和列名加上引号。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。
- 如果从不想给表名和字段名加引号, 则选中**从不**。
- 如果想给所有表名和字段名加引号, 则选中**始终**。

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。

注意。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 中读取空字符串。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 将空字符串值等同于空值进行处理并存储。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, 首先必须选择要连接到的数据源。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。

注意: 要获取打开此对话框的替代方法, 请从此主菜单中选择:

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。

用户名。如果数据源受密码保护, 请输入用户名。

密码。如果数据源受密码保护, 请输入密码。

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”。请参阅主题『为数据库连接指定预设值』, 了解更多信息。

要删除连接, 可从列表选择一个连接, 然后单击 **删除**。

完成选择后, 请单击**确定**。

- 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, 必须在两个计算机上都安装 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。

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。

- 在 DB2 9.1 或更高版本上运行的 IBM InfoSphere Warehouse。请参阅主题『用于 IBM DB2 InfoSphere Warehouse 的设置』, 了解更多信息。
- SQL Server 2008 或更高的 Enterprise 与 Developer 版本。请参阅主题第 13 页的『用于 SQL Server 的设置』, 了解更多信息。
- Oracle 10g 与 11gR1 或更高的 Enterprise 或 Personal 版本。请参阅主题第 13 页的『用于 Oracle 的设置』, 了解更多信息。
- 所有以类似方式连接到数据库或模式的 IBM Netezza、IBM DB2 for z/OS 和 Teradata。请参阅主题第 14 页的『用于 IBM Netezza、IBM DB2 for z/OS、IBM DB2 LUW 和 Teradata 的设置』, 了解更多信息。

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 IBM DB2 InfoSphere Warehouse 的设置

这些设置针对 DB2 9.1 或更高版本上运行的 IBM InfoSphere Warehouse 显示。

表空间。用于导出的表空间。数据库管理员可以创建表空间或将表空间配置为分区。建议选择其中一个表空间（而不是缺省表空间）以用于数据库导出。

使用压缩。如果选中此选项，那么将使用压缩创建要导出的表格（例如，相当于 SQL 中的 CREATE TABLE MYTABLE(...) COMPRESS YES;）。

请勿记录更新。如果选中了此选项，请避免在创建表和插入数据（相当于 SQL 中的 CREATE TABLE MYTABLE(...) NOT LOGGED INITIALLY;）时进行记录。

用于 SQL Server 的设置

这些设置针对 SQL Server 2008 或更高版本的 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。

压缩支持。选择压缩层级。

- **行。**启用低级别压缩（例如，相当于 SQL 中的 CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW);）。
- **页面。**启用页面级别压缩（例如，SQL 中的 CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);）。

用于 Oracle 的设置

Oracle 10g 设置

这些设置针对 Oracle 10g Enterprise 或 Personal 版本显示。

使用压缩。如选中，将使用压缩为导出创建表格。对于此版本的数据库，仅可使用基本压缩（例如，SQL 中的 CREATE TABLE MYTABLE(...) COMPRESS;）。

Oracle 11gR1 设置

这些设置针对 Oracle 11g R1 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。

压缩支持。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**直接加载操作**选项的效果相同。
- **直接加载操作。**仅对批量（直接路径）插入操作启用压缩（例如，SQL 中的 CREATE TABLE MYTABLE(...) COMPRESS FOR DIRECT_LOAD OPERATIONS;）。
- **所有操作。**对所有操作启用压缩（例如，SQL 中的 CREATE TABLE MYTABLE(...) COMPRESS FOR ALL OPERATIONS;）。

Oracle 11gR2 设置 -“基本”选项

这些设置针对使用“基本”选项的 Oracle 11g R2 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。

压缩支持。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 11gR2 设置 -“高级”选项

这些设置针对使用“高级”选项的 Oracle 11g R2 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。

压缩支持。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 Netezza、IBM DB2 for z/OS、IBM DB2 LUW 和 Teradata 的设置

指定 IBM Netezza、IBM DB2 for z/OS、IBM DB2 LUW 或 Teradata 的预设值时，系统将提示您选择以下选项：

使用服务器评分适配器数据库/模式。如果选择此项，会启用服务器评分适配器数据库/模式选项。

服务器评分适配器数据库/模式。从下拉列表中，选择您需要的连接。

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。

注意：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 32 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。

设为默认值。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

已连接到数据源后，可以选择使用 SQL 查询导入字段。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。

如果指定了多个 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）。请参阅主题『在 SQL 查询中使用流参数』，了解更多信息。

加载查询。单击此选项可打开文件浏览器，使用该浏览器可加载先前保存的查询。

保存查询。单击此选项可打开“保存查询”对话框，使用该对话框可保存当前的查询。

导入缺省值。单击此选项可导入使用对话框中选择的表和列自动构建的示例 SQL SELECT 语句。

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 会尝试对查询进行解析并确定是否存在应该逐条发送给数据库的 SQL 语句。

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 1. 流参数示例。

流参数名称（示例）	存储	流参数值	解释为
PString	字符串	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	时间戳	2011-03-02 23:05:01	ts{'2011-03-02 23:05:01'}
PColumn	未知	IntValue	IntValue

在 SQL 查询中，可以采用与 CLEM 表达式中相同的方式来指定流参数，即 '\$P-<parameter_name>'，其中 <parameter_name> 是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';
```

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。

读取分隔的文本数据时的注意事项

- 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)。也可以选择通过节点来删除空格。
- 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。如果做不到这点, 可以用双引号括起所有的文本字段, 前提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了双引号, 那么可以用单引号括起文本字段, 前提是字段值内均不包含单引号。如果单引号和双引号都不能使用, 就需要修改文本值以删除或代替分隔符或单/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, 则取消 **读取文件中的字段名** 为每个字段分配一个一般名称, 例如 *Field1*、*Field2*, 依此类推。
- 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(美语为 period 或英语为 full-stop)仅能在适当的情况下使用。
- 日期值和时间值应该采用“流选项”对话框中可识别的的格式之一, 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都采用同一格式, 并且包含日期的任意字段内的所有值必须使用同一格式。

为变量文件节点设置选项

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。

可复制显示自数据源的示例文本并将其粘贴到下列控件中: EOL 注解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。

读取文件中的字段名。此选项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 注解字符。指定字符（例如 # 或 !）表示数据中的注解。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。

注意。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。您可以选择系统默认值、流默认值或 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。**流默认值**是从流属性对话框的“选项”选项卡中选择的符号。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。

行分隔符为换行符。要将换行符用作行分隔符，而非字段分隔符，可选择此选项。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**。

请注意，如果选择此选项，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“;|”。默认的定界符是逗号。

注意：如果还将逗号定义为小数分隔符，那么此处的缺省设置将不起作用。如果逗号既是字段定界符又是小数分隔符，则可在定界符列表中选择**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。

列和类型的扫描行数。指定要扫描特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。

注意：为避免当使用来自先前 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 13 之前的版本中保存的信息将不选中此复选框。

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。

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。

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。

为固定文件节点设置选项

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。

- 每个断点线会自动将一个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。

开始。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。

注意。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**可从数据输入中除去无效字符。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。您可以选择系统默认值、流默认值或 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。**流默认值**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。

注意：为避免当使用来自先前 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 13 之前的版本中保存的信息将不选中此复选框。

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。

设置字段存储类型和格式

使用固定文件节点、变量文件节点、XML 源节点和用户输入节点的“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。

对于从其他源中读取的数据，存储类型自动确定，但可使用转换函数（例如 `to_integer`）在过滤节点或导出节点中对其进行更改。

字段。使用字段列可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。请参阅主题第 93 页的『测量级别』，了解更多信息。

可用存储类型有：

- **字符串。**用于包含非数字数据（也称作字母数字数据）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。**标准格式表示的日期值，比如年月日（例如 26.09.07）。具体格式在“流属性”对话框中指定。
- **时间。**指的是持续时间。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部分的值，例如 2007-09-26 09:04:00，具体取决于“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。）

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`。请参阅主题第 110 页的『使用填充节点进行存储类型转换』，了解更多信息。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，如果想将值为 *Jan 2003*、*Feb 2003* 等字符串字段转换为日期存储格式，请选择 **MON YYYY** 作为流的

默认日期格式。“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。请参阅主题第 109 页的『使用导出节点对值进行重新编码』，了解更多信息。

读取混合数据。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，从不同的区域设置中合并数据时，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，这些选项可用于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。请参阅主题第 100 页的『字段格式设置选项卡』，了解更多信息。

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。请参阅主题第 100 页的『设置字段格式选项』，了解更多信息。

对于所有存储类型，也可以选择 **流默认值** 以便导入时使用流默认设置。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。

Data Collection 节点

Data Collection 源节点根据 IBM Corp. 的市场调查软件所使用的 IBM SPSS Data Collection Survey Reporter Developer Kit 导入调查数据。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）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。

注意：此节点需要 IBM SPSS Data Collection Survey Reporter Developer Kit，它已随 IBM Corp. 的 IBM SPSS Data Collection 软件产品一起发行。有关详细信息，请参阅 IBM SPSS Data Collection Web 页面 ()。除安装 Developer Kit 外，不需要其他配置。

注释(O)

- 可从平面、表格 VDATA 格式或层级 HDATA 格式中的源（在这些源包括元数据源的情况下）读取调查数据（要求 IBM SPSS Data Collection 4.5 或更高的版本）。
- 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 IBM 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。

Data Collection 导入文件选项

使用 Data Collection 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

注意: 要查看可用提供程序文件类型的完整列表, 您需要安装随 IBM SPSS Data Collection 软件提供的 IBM SPSS Data Collection Survey Reporter Developer Kit。有关更多信息, 请参阅 IBM SPSS Data Collection 网页: <http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/>。

元数据提供者。可从 IBM SPSS Data Collection Survey Reporter Developer Kit 软件所支持的多种格式中导入调查数据。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 IBM SPSS Data Collection 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 IBM SPSS Data Collection 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*。
- **统计量文件。**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。请参阅主题第 23 页的『IBM SPSS Data Collection 导入元数据属性』, 了解更多信息。

观测值数据设置

注意: 要查看可用提供程序文件类型的完整列表, 您需要安装随 IBM SPSS Data Collection 软件提供的 IBM SPSS Data Collection Survey Reporter Developer Kit。有关更多信息, 请参阅 IBM SPSS Data Collection 网页: <http://www.ibm.com/software/analytics/spss/products/data-collection/survey-reporter-dev-kit/>。

获取观测值数据设置。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。

观测值数据提供者。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。请参阅主题第 23 页的『数据库连接字符串』，了解更多信息。此组件的内部名称是 *mrADODsc*。
- **分隔的文本文件 (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。**从本机的 IBM SPSS Data Collection 数据格式文件中读取观测值数据 (IBM SPSS Data Collection 4.5 之前的版本)。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 IBM SPSS Data Collection 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*。
- **Quantum 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。请参阅主题第 23 页的『数据库连接字符串』，了解更多信息。内部名称是 *mrRdbDsc2*。
- **统计量文件。**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。默认情况下，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 IBM SPSS Data Collection 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*。

观测值数据类型。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。请参阅主题第 23 页的『数据库连接字符串』，了解更多信息。

观测值数据工程。从 IBM SPSS Data Collection 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。

多响应变量导入方式。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。请参阅主题第 23 页的『导入多重响应集』，了解更多信息。

IBM SPSS Data Collection导入元数据属性

导入 IBM SPSS Data Collection 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。默认类型为 **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 IBM SPSS Data Collection 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 IBM SPSS Data Collection 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 ID 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 IBM SPSS Data Collection 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。请参阅主题第 103 页的『**编辑多重响应集**』，了解更多信息。

将多个响应导入单个字段（适用于在以前版本中创建的流）

在旧版本的 IBM 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。

IBM SPSS Data Collection列导入说明

IBM SPSS Data Collection 数据中的列按照在下表中汇总的方式读入 IBM SPSS Modeler。

表 2. IBM SPSS Data Collection 列名导入摘要

IBM SPSS Data Collection 列类型	IBM SPSS Modeler 存储器	测量级别
布尔标志 (yes/no)	字符串	标记 (值为 0 和 1)
分类	字符串	名义
日期或时间戳	时间戳	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)	字符串	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 IBM SPSS Modeler	
对象 (二进制数据, 例如显示不规则文字的传真或声音记录)	不导入到 IBM SPSS Modeler	
无 (未知类型)	不导入到 IBM 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*。

Analytic Server 源

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 建模器服务器 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。

数据源。假定 SPSS 建模器服务器 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。请参阅主题『Analytic Server 选择数据源』, 了解更多信息。

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...**

Analytic 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。可以输入字符串、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。

Analytic Server 凭证

如果用于访问 Analytic Server 的凭证与用于访问 SPSS 建模器服务器的凭证不同，那么在 Analytic Server 上运行时，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 上执行很多 SPSS Modeler 节点，但是某些节点的执行过程可能存在一些差异，而且某些节点当前不受支持。本主题详细介绍了当前支持级别。

一般

- Analytic Server 将不接受加引号的建模器字段名称中通常可接受的某些字符。
- 要在 Analytic Server 中运行“建模器”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”节点结束；允许合并，但不允许派生。
- 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，而得分的输出数据模型将遵循目标存储。因此，如果连续目标具有整数存储，那么写入值与得分的数据模型将出现不匹配，并且此不匹配将在您尝试读取评分数据时引起错误。

源

- 将在本地运行开始于 Analytic Server 源节点以外的任何内容的流。

记录操作

支持所有记录操作。以下是有关这些节点的功能的进一步说明。

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。
- 不支持复杂的抽样方法。

汇总

- 不支持连续关键字。
- 不支持订单统计信息（中位数、第一四分位数和第三四分位数）。

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，这些操作用于保持“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。
- 对于根据派生的随机数键进行排序以随机地重新排列记录，从而避免原始记录有序时造成模型构建算法产生偏差而言，依次使用“排序”和具有针对超大数据集进行优化目标的建模节点（神经网络、线性、C&R 树、Quest 或 CHAID）是一种很有用的模式。

通常，应该将“排序”节点放在尽可能靠近需要已排序记录的操作的位置。

合并

- 不支持按顺序合并。

- 不支持按条件合并。
- 不支持“优化”选项卡。
- 目前，不支持将“样本”节点或模型块放在 Analytic Server“源”节点与“合并”节点之间。通常，可以指定“选择”节点来代替“样本”节点的功能。
- Analytic Server 无法根据空字符串键进行连接；即，如果要作为合并依据的其中一个键包含空字符串，那么将从合并后的输出中删除所有包含空字符串的记录。
- 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。

字段操作

支持下列节点：自动数据准备、类型、过滤、派生、整体、填充、重新分类、分级、RFM 分析、分区、设为标志、重构和字段重排。以下是有关这些节点的功能的进一步说明。

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。

类型

- 不支持“检查”列。
- 不支持“格式”选项卡。

导出

- 支持除序列函数以外的所有派生函数。
- 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。
- 不能在比较中单独使用标志字段；即，`if (flagField) then ... endif` 将产生错误；变通方法是使用 `if (flagField=trueValue) then ... endif`
- 建议在使用 `**` 运算符时将指数指定为实数（例如 `x**2.0`，而不是 `x**2`），以匹配 Modeler 中的结果

填充

- 支持派生节点所支持的一组功能。

分级 不支持以下功能。

- 最优分级
- 排序
- 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：超出 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“保流在当前分级中”选项。RFM 近度、频率和值度得分并非始终与 Modeler 根据同一数据计算得出的得分相匹配。得分范围将相同，但得分分配（分级数字）可能会相差 1。

图形 支持所有“图形”节点。

建模 支持有限数目的“建模”节点，其中包括线性、神经网络、C&RT、Chaid 和 Quest。以下是有关这些节点的功能的进一步说明。

线性 根据大型数据构建模型时，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。如果输入记录按某种违背了所使用的整体模型构建算法基于的随机性假定的系统化方式进行排序，那么您也可能会遇到问题。
- 不支持自动数据准备。在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。
- 未计算分割模型的准确性统计信息。

神经网络

根据大型数据构建模型时，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。如果输入记录按某种违背了所使用的整体模型构建算法基于的随机性假定的系统化方式进行排序，那么您也可能会遇到问题。
- 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。
- 未计算分割模型的准确性统计信息。

C&R 树、CHAID 和 Quest

根据大型数据构建模型时，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。如果输入记录按某种违背了所使用的整体模型构建算法基于的随机性假定的系统化方式进行排序，那么您也可能会遇到问题。
- 不支持交互式会话。
- 未计算分割模型的准确性统计信息。

模型评分

支持对下列模型块进行评分：C&RT、Quest、CHAID、线性、回归、神经网络、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 均数、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`
- 对模型进行评分时，Analytic Server 不会通过检查来确认模型中使用的所有字段是否存在于数据集中，因此请先确保满足这一条件，然后再在 Analytic Server 中运行

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、统计信息和均数。

导出 流可以开始于 Analytic Server 源节点，并以 Analytic Server 导出节点以外的导出节点结束，但数据将从 HDFS 移动到 SPSS 建模器服务器，并最终移动至导出位置。

IBM Cognos BI 源节点

通过 IBM Cognos BI 源节点可将 Cognos BI 数据库数据或单一列表报告导入到数据挖掘会话中。这样，可将 Cognos 的商务智能功能与 IBM SPSS Modeler 的预测分析能力融为一体。可以导入关系数据、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）。请参阅主题『Cognos 对象图标』，了解更多信息。

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。注：要导入的数据必须使用 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。注意：只能导入单一列表报告；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。

Cognos 对象图标

可从 Cognos BI 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 3. Cognos 对象图标.










图标	对象
	数据包
	命名空间
	查询对象
	查询项目
	测量维度

表 3. Cognos 对象图标 (续).

图标	对象
	测量
	维度
	层级层次
	水平
	过滤器
	报告
	独立计算

导入 Cognos 数据

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

连接。单击**编辑**按钮将显示一个对话框，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。请参阅主题第 30 页的『Cognos 连接』，了解更多信息。

位置。建立 Cognos 服务器连接后，单击此字段旁边的**编辑**按钮可显示从中导入内容的可用数据包列表。请参阅主题第 30 页的『Cognos 位置选择』，了解更多信息。

内容。显示选中数据包的名称，以及和该数据包关联的名称空间。双击名称空间可显示能够导入的对象。各个对象类型通过不同的图标表示。请参阅主题第 28 页的『Cognos 对象图标』，了解更多信息。

在选择要导入的对象时，应选中对象，并单击两个右箭头的上部，以便将对象移动到**待导入字段**窗格中。选择查询主体将导入它的所有查询项。双击查询主体将展开该主体，以便您可以选择它的一个或多个查询项。通过按住 Ctrl 并单击（选择个别项）、按住 Shift 并单击（选择一块项）以及按住 Ctrl 和 A（选择所有项），可以执行多种选择。

在选择要应用的过滤器时（如果数据包定义了过滤器），应在“内容”窗格中导航至过滤器，选中过滤器并单击两个右箭头的下部，以便将过滤器移动到**需应用的过滤器**窗格中。通过按住 Ctrl 并单击（选择个别过滤器）和按住 Shift 并单击（选择一块过滤器），可以执行多种选择。

待导入字段。列出您已选定要导入 IBM SPSS Modeler 以供处理的数据库对象。如果不再需要一个特定对象，请将其选中并单击左箭头以将其移回到**内容**窗格中。可采用和**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之前应用的过滤器。如果不再需要一个特定过滤器，请将其选中并单击左箭头以将其移回到**内容**窗格中。可采用和**内容**相同的方式执行多选。

参数。如果此按钮处于启用状态，那么选定对象定义了参数。您可以使用参数来进行调整（例如，执行参数化计算），然后再导入数据。如果定义了参数但未提供缺省值，那么此按钮将显示一个警告三角形。单击此按钮可显示参数并（可选）对其进行编辑。如果此按钮处于禁用状态，那么报告未定义任何参数。

在导入前汇总数据。如果要导入汇总数据而不是原始数据，选中此复选框。

导入 Cognos 报告

要从 IBM Cognos BI 数据库导入预定义报告，请确保 IBM Cognos BI 对话框的“数据”选项卡上的方式设置为报告。注意：只能导入单一列表报告；不支持导入多个列表。

连接。单击编辑按钮将显示一个对话框，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。请参阅主题『Cognos 连接』，了解更多信息。

位置。建立 Cognos 服务器连接后，单击此字段旁边的编辑按钮可显示从中导入内容的可用数据包列表。请参阅主题『Cognos 位置选择』，了解更多信息。

内容。显示包含报告的选定数据包或文件夹的名称。导航到相应的报告，选中它并单击右箭头以将其移动到待导入报告字段中。

要导入的报告。指示您已选定要导入 IBM SPSS Modeler 的报告。如果不再需要该报告，请将其选中并单击左箭头以将其移回到内容窗格中，或者将其他报告移动到该字段中。

参数。如果此按钮处于启用状态，那么选定报告定义了参数。可以先使用参数来进行调整，然后再导入报告（例如，指定报告数据的开始日期和结束日期）。如果定义了参数但未提供缺省值，那么此按钮将显示一个警告三角形。单击此按钮可显示参数并（可选）对其进行编辑。如果此按钮处于禁用状态，那么报告未定义任何参数。

Cognos 连接

使用“Cognos 连接”对话框可以选择在其中导入或导出数据库对象的 Cognos BI 服务器。

Cognos 服务器 URL。输入要从中导入或导出数据对象的 Cognos BI 服务器的 URL。这是 Cognos BI 服务器上 IBM Cognos 配置的“外部分派器 URI”环境属性的值。如果不确定要使用哪个 URL，请与 Cognos 系统管理员联系。

方式。如果希望使用特定的 Cognos 名称空间、用户名和密码（例如作为管理员）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证书，在该情况您不需填写其他字段。

名称空间。指定用于登录服务器的 Cognos 安全认证提供者。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。

用户名。输入用于登录到服务器的 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。单击此按钮存储这些设置作为您的默认值，避免每次打开节点时都重新输入它们。

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 BI 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**。虽然报告文件夹可以包含其他报告文件夹和各个报告，但是对于每个 Cognos BI 源节点，只能选择一个数据包或报告文件夹。

指定数据或报告参数

如果在 Cognos BI 中定义了参数，那么对于数据对象或报告，您可以先指定这些参数的值，然后再导入对象或报告。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。参数名称，与在 Cognos BI 数据库中指定的名称相同。

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。

自动除去表中的无效参数。默认选择此选项，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。

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

从 Cognos PM 主数据中心连接，首先选择从中导入数据的 TM1 服务器，即包含一个或多个 TM1 多维数据集的服务器。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：必须将以下三个进程（*ExportToSPSS.pro*、*ImportFromSPSS.pro* 和 *SPSSCreateNewMeasures.pro*）从 Modeler 复制到 TM1 服务器，才能使用 SPSS Modeler 中的 TM1 源或导出节点。要将这些进程添加到 TM1 服务器，必须将这些文件复制到 TM1 服务器的数据目录，然后重新启动 TM1 服务器。以下目录中提供了这些文件：<Modeler 安装目录>/ext/bin/pasw.tm1/scripts。

导入 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“导入”选项卡上，选择相关 TM1 PM 主数据中心及关联的服务器、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。

PM 系统。输入主数据中心的 URL，该主数据中心包含要与其进行连接的 TM1 服务器。

TM1 服务器。建立 Cognos 主数据中心连接后，请选择包含要导入的数据的服务器，然后单击**登录**。如果您先前未连接到此服务器，那么系统将提示您输入**用户名和密码**；或者，您可以选择其他服务器。

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

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**窗格中。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。

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。

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。

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。

SAS 源节点

注意: 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

导入。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按钮 (...) 来浏览文件的位置。

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按钮 (...) 来浏览文件的位置。

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

使用 Excel 源节点可以从任何版本的 Microsoft Excel 中导入数据。

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。

使用指定的范围。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。

在空行中。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。

第一行存在列名称。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。

字段存储和测量级别

从 Excel 中读取值时, 默认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)。请参阅主题第 19 页的『设置字段存储类型和格式』, 了解更多信息。

默认情况下, 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。要避免此问题, 可以在 Excel 电子表格中手动将单元格格式设置为 **文本**, 这样将按字符串读取所有的值 (包括数字)。

XML 源节点

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

XML 源节点允许您将 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/>)。

读取单一文件。默认情况下, IBM 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 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，您可以在该对话框中选择要使用的根元素。请参阅主题『从多个根元素中选择』，了解更多信息。

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，并单击右方向箭头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，然后单击右箭头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箭头按钮。

- **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。

选择要显示的根元素。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 源数据中除去不需要的空格

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”节点除去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。

“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 4. “模拟拟合”节点的更新结果

“模拟生成”节点中的字段	“模拟拟合”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。

表 4. “模拟拟合”节点的更新结果 (续)

	“模拟拟合”节点 中的字段
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。

除去指向“模拟拟合”节点的更新链接

通过完成下列步骤, 可以除去“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**除去更新链接**。将除去此链接。

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。

选择项。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
- 缺失
- 检查
- 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:

- 添加和除去字段。
- 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”选用板添加到流画布中, 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。
- **存储类型**。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，您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。带符号的整数表示自 1970 年 1 月 1 日午夜以来（或在此之前）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。请参阅主题第 40 页的『拟合详细信息』，了解更多信息。



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。请参阅主题第 43 页的『分布』，了解更多信息。

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。请参阅主题第 42 页的『迭代』，了解更多信息。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。请参阅主题第 41 页的『指定参数』，了解更多信息。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。请参阅主题第 41 页的『指定参数』，了解更多信息。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。请参阅主题第 40 页的『拟合详细信息』，了解更多信息。

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 5. “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。请参阅主题第 41 页的『指定参数』，了解更多信息。
	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。请参阅主题第 40 页的『克隆字段』，了解更多信息。
	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。
- **0。**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。

显示方式。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。

将已拟合的多路列联表用于具有分类分布的输入。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，那么将从列联表中除去该字段。

“高级选项”视图

要模拟的观测值数。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。

- **最大观测值数。**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。
- **迭代数。**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。**只有在迭代数大于 1 的情况下，才会启用此选项。将使用所显示的公式自动计算此数字，并且此数字无法进行编辑。
- **创建迭代字段。**只有在迭代数大于 1 的情况下，才会启用此选项。选择此选项后，将启用名称字段。请参阅主题第 42 页的『迭代』，了解更多信息。
- **名称。**只有在选中了创建迭代字段复选框，并且迭代数大于 1 的情况下，才会启用此选项。可以通过在此文本字段中输入内容来编辑迭代字段的名称。请参阅主题第 42 页的『迭代』，了解更多信息。

随机种子。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**。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。

要创建的副本数。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。

初始副本编号。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。

副本编号步进值。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。

处理方法（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。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 43 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，您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。
- **拒绝小于最小值的值。**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。
- **拒绝大于最大值的值。**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：

- **迭代固定字段。**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

- **迭代分布参数。**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

- 迭代分类分布或范围分布的分布参数。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，所需要的试验次数的分布。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。此分布包含一组区间，系统为每个区间都指定了一个概率，以使所有区间的概率总和等于 1。指定区间内的值来自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 6. 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	$0 \leq Probability \leq 1$	
贝塔	整数、实数和日期时间	形状 1 形状 2 最小值 最大值	≥ 0 ≥ 0 $< Maximum$ $> Minimum$	最小值和最大值是可选的。
二项	整数、实数和日期时间	试验次数 (n) 概率 最小值 最大值	> 0 , 整数 $0 \leq Probability \leq 1$ $< Maximum$ $> Minimum$	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)	$0 \leq Value \leq 1$	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	面数	$2 \leq Sides \leq 20$	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。

表 6. 可用于定制拟合的分布 (续)

分布	定制拟合支持的存储类型	参数	参数限制	注释
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	刻度 最小值 最大值	> 0 $< \textit{Maximum}$ $> \textit{Minimum}$	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 刻度 最小值 最大值	≥ 0 ≥ 0 $< \textit{Maximum}$ $> \textit{Minimum}$	最小值和最大值是可选的。
对数正态	整数、实数和日期时间	形状 1 形状 2 最小值 最大值	≥ 0 ≥ 0 $< \textit{Maximum}$ $> \textit{Minimum}$	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	阈值 概率 最小值 最大值	≥ 0 $0 \leq \textit{Probability} \leq 1$ $< \textit{Maximum}$ $> \textit{Minimum}$	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	阈值 概率 最小值 最大值	≥ 0 $0 \leq \textit{Probability} \leq 1$ $< \textit{Maximum}$ $> \textit{Minimum}$	最小值和最大值是可选的。
正态	整数、实数和日期时间	均数 标准差 最小值 最大值	≥ 0 > 0 $< \textit{Maximum}$ $> \textit{Minimum}$	最小值和最大值是可选的。
泊松	整数、实数和日期时间	均数 最小值 最大值	≥ 0 $< \textit{Maximum}$ $> \textit{Minimum}$	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)	$0 \leq \textit{Value} \leq 1$	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	$\textit{Minimum} \leq \textit{Value} \leq \textit{Maximum}$ $< \textit{Maximum}$ $> \textit{Minimum}$	

表 6. 可用于定制拟合的分布 (续)

分布	定制拟合支持的存储类型	参数	参数限制	注释
均匀	整数、实数和日期时间	最小值 最大值	< <i>Maximum</i> > <i>Minimum</i>	
韦伯	整数、实数和日期时间	速率 级别 位置 最小值 最大值	> 0 > 0 ≥ 0 < <i>Maximum</i> > <i>Minimum</i>	位置、最大值和最小值是可选的。

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”选项板中添加的节点，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。也可以选择将字段留为空 - 留为空的字段将填入系统空值 (\$null\$)。

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"

对于数值字段，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默认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。** LOW
- **胆固醇。** NORMAL HIGH
- **药品。** (留空)

表 7. 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 8. 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更改。请参阅主题第 93 页的『测量级别』，了解更多信息。

可用存储类型有：

- **字符串。**用于包含非数字数据（也称作字母数字数据）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。**标准格式表示的日期值，比如年月日（例如 26.09.07）。具体格式在“流属性”对话框中指定。
- **时间。**指的是持续时间。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部分的值，例如 2007-09-26 09:04:00，具体取决于“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。）

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`。请参阅主题第 110 页的『使用填充节点进行存储类型转换』，了解更多信息。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，如果想将值为 *Jan 2003*、*Feb 2003* 等字符串字段转换为日期存储格式，请选择 **MON YYYY** 作为流的默认日期格式。“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。请参阅主题第 109 页的『使用导出节点对值进行重新编码』，了解更多信息。

读取混合数据。 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

对于数值字段，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

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。请参阅主题第 101 页的『设置过滤选项』，了解更多信息。
- **“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 中数据的值和字段标签。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，用于描述某个给定字段中数据的特征。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**。请参阅主题第 93 页的『测量级别』，了解更多信息。

注意：字段的测量级别与字段的存储类型不同，后者表明数据是作为字符串、整数、实数、日期、时间还是时间戳进行存储。

- **值**。此列允许您指定从数据集读取数据值的选项，或使用**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。请参阅主题第 96 页的『数据值』，了解更多信息。
- **缺失**。用于指定字段缺失值的处理方法。请参阅主题第 98 页的『定义缺失值』，了解更多信息。
- **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。请参阅主题第 98 页的『检查类型值』，了解更多信息。
- **角色**。用于告知建模节点字段将成为用于某个机器学习过程的**输入**（预测变量字段）还是**目标**（预测字段）。**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。请参阅主题第 99 页的『设置字段角色』，了解更多信息。

请参阅主题第 92 页的『类型节点』，了解更多信息。

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，则在源节点上进行实例化是最方便的方法。

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。请参阅主题第 101 页的『设置过滤选项』，了解更多信息。

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 ID*）的值合并记录。您可能还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”节点可基于特定条件从数据流中选择或废弃记录子集。例如，可以选择有关特定销售区域的记录。



“样本”节点选择记录的子集。受支持的样本类型有许多，其中包括分层、聚类和非随机（结构化）样本。取样对于提高性能和选择相关记录组或交易组用于分析会很有用。



“平衡”节点纠正数据集中的不平衡度，因而它遵循指定的条件。平衡指令调整根据指定系数条件为真的记录的比例。



“合计”节点用汇总和合计的输出记录替代一系列输入记录。



使用“近因、频率和货币 (RFM) 汇总”节点，您可以采用客户的历史记录事务处理数据，删除所有无用数据以及将所有他们保留的事务处理数据组合成一行，且该行中列出了他们与您上次谈业务的时间、所完成的交易量以及这些交易的总货币价值。



“排序”节点可根据一个或多个字段的值将记录按升序或降序排序。



“合并”节点获取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”节点用于连接各组记录。也可以用于合并结构类似但数据不同的数据集。



“区分”节点将除去重复的记录，方法是将第一个可区分记录遍历到数据流，或者废弃第一个记录而将任何重复记录遍历到数据流。



“流式 TS”节点在某个步骤中构建时间序列模型并对其进行评估，而不需要“时间区间”节点。

记录操作选项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。

- 包括。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。受支持的样本类型有许多，其中包括分层、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。请参阅主题第 121 页的『分区节点』，了解更多信息。

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。

系统化或 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。

抽样加权。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默认模式，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：

- **从第一条记录开始连续抽取。**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。**选择此选项将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。

使用块级别抽样（仅限数据库中）。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。

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 **包含** 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。请参阅主题第 55 页的『聚类和分层设置』，了解更多信息。

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。

样本单元。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。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。

注意：为从数据库中读取的记录选择**设置随机种子**选项时，在抽样前需要使用 **Sort** 节点以确保每次执行节点时能得到同一结果。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。请参阅主题第 60 页的『排序节点』，了解更多信息。

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。

集群。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。

分层方式。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。

使用输入加权。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(O)

- 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系列检验主项，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，这样您就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。请参阅主题第 55 页的『聚类和分层设置』，了解更多信息。
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。请参阅主题第 95 页的『什么是实例化？』，了解更多信息。
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**，此操作通过上下箭头按钮完成。

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得分时，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

汇总是一个经常用于减小数据集大小的数据准备任务。继续执行汇总之前，应该花一些时间来清理数据，尤其要关注缺失值。一旦完成汇总，或许会丢失可能有用的缺失值信息。

您可以使用 **Aggregate** 节点将一个输入记录序列替换为汇总，即经过汇总的输出记录。例如，您可能有一系列输入销售记录，例如下表中所显示的记录。

表 9. 销售记录输入示例

年龄	性别	地区	分支地点	销售
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	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* 作为键字段汇总这些记录。然后选择使用**平均值**模式汇总 *Age*，使用**合计**模式汇总 *Sales*。在“Aggregate 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以下表中。

表 10. 汇总的记录示例

年龄（平均值）	性别	地区	销售（合计）	记录计数
35.5	F	N	25	4
29	F	S	6	1

表 10. 汇总的记录示例 (续)

年龄 (平均值)	性别	地区	销售 (合计)	记录计数
34.5	M	N	20	2
33.75	M	S	20	4

例如，您可从中知道，北部地区的四名女性销售员工的平均年龄为 35.5 岁，其销售合计为 25 件产品。

注意如果不指定汇总方式，那么将自动废弃 *Branch* 之类的字段。

为汇总节点设置选项

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）以用于每个汇总字段的输出

您还可以指定用于新增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，对于新增的性能，启用并行处理可能会对汇总操作有好处。

关键字段。列出可用作汇总类别的字段。连续（数字）字段和类别字段都可用作关键字段。如果您选择多个关键字段，这些值将进行合并，从而生成用于汇总记录的关键值。对于每个唯一关键字段，都会生成一个经过汇总的记录。例如，如果 *Sex* 和 *Region* 是关键字段，*M* 和 *F* 与区域 *N* 和 *S* 的每个唯一性组合（四个唯一性组合）都将具有一个经过汇总的记录。要添加关键字段，请使用窗口右侧的字段选择器按钮。

关键字是连续的。如果您知道在输入中具有相同关键值的所有记录被分成了一组，那么可以选择此选项（例如，如果对关键字段上的输入进行了排序）。这样做有助于提高性能。

对话框的剩余部分被分为两个主要区域 - 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，以及选择的汇总方式。要向此列表添加字段，请使用右侧的“字段选择器”按钮。可用的汇总模式如下。

注意：某些方式不适用于非数字字段（例如，合计不适用于日期/时间字段）。不能用于选定汇总字段的模式将被禁用。

- **合计。**选择此选项可返回每个关键字段组合的合计值。合计是指所有含非缺失值的观测值的总计值。
- **平均值。**选择此选项可返回每个关键字段组合的平均值。该平均值是对集中趋势的测量，它是算术平均值（总和除以观测值个数）。
- **最小值。**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，均值容易受到少数非常大或非常小的值的影响）。又称为第 50 个百分位或第二个四分位。
- **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个数减一。
- **第一个四分位**。选择此选项可返回每个关键字段组合的第一个四分位（第 25 个百分位）值。
- **第三个四分位**。选择此选项可返回每个关键字段组合的第三个四分位（第 75 个百分位）值。

注意：运行包含 Aggregate 节点的流时，将 SQL 推送回 Oracle 数据库时针对第一个四分位和第三个四分位所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“全部应用”按钮，将选择的方式应用于上面列出的所有字段。

新的字段名称扩展。选择此选项可添加后缀或前缀（如“1”或“new”）以便复制汇总的字段。例如，如果您选择了后缀选项，并将“1”指定为扩展时，针对字段 Age 的最小值汇总结果将产生一个 Age_Min_1 字段名。注意：汇总扩展名（例如 _Min 或 Max_）将自动添加到新字段，以指示所执行的汇总的类型。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。选择此选项可在每个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。在编辑字段中键入内容，可以为此字段创建定制名称。

注意：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 Filler 节点将空白值替换为空值。您还可以使用选择节点删除空白值。

汇总表达式

表达式类似于使用值、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个记录，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。

请注意，**关键字段**和创建的汇总表达式之间存在关联，因为汇总表达式根据关键字段进行分组。

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其进行管理的规则：

- 您可以使用标量函数将多个汇总函数相合并以生成一个汇总结果。例如：

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

- 汇总函数可对多个标量函数的结果进行操作。例如：

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

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 ID 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。请参阅主题第 119 页的『RFM 分析节点』，了解更多信息。

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）。请参阅主题第 61 页的『合并节点』，了解更多信息。

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。

计算相对于以下内容的近因。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。

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 (ID)：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，如果某个字符串字段具有 *Jan 2007*、*Feb 2007* 等类似值，则可以使用填充节点或 `to_date()` 函数将其转换为日期字段。请参阅主题第 110 页的『使用填充节点进行存储类型转换』，了解更多信息。

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。

新的字段名称扩展。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。

只包含最近交易。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：

- **以下日期后的交易日期。**指定交易日期以在分析时包含其之后的记录。
- **最近的交易。**指定从**计算相对于以下内容的近因**日期字段所返回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。

保存第二个最近交易的日期。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。

- 通过使用右侧的“字段选择器”按钮可向列表**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**。
- **删除字段**，此操作通过使用红色删除按钮完成。
- **对指令排序**，此操作通过上下箭头按钮完成。

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。

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。

数据预先经过排序。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。

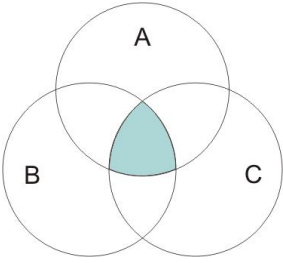
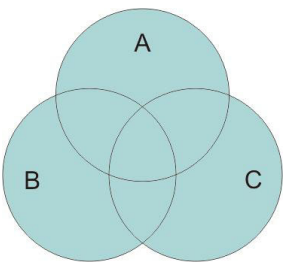
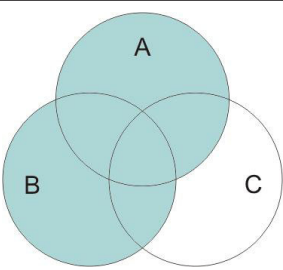
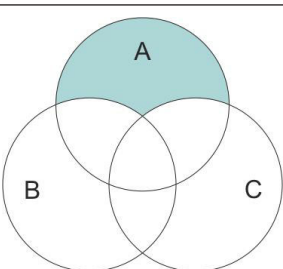
- **按顺序合并**会按照输入的顺序连接来自所有源的相应记录，直到穷尽最小的数据源。如果使用此选项，务必确保已使用排序节点完成了对数据的排序。
- **使用关键字段合并**（如客户标识）可指定如何将来自一个数据源的记录与来自其他数据源的记录相匹配。连接的类型有许多，其中包括内部连接、完全外部连接、部分外部连接和反连接。请参阅主题『[连接类型](#)』，了解更多信息。
- **按条件合并**，可指定要执行合并必须满足的条件。可以在节点中直接指定条件，也可以使用“表达式构建器”构建条件。

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 ID*）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。

	<p>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。</p>
	<p>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。</p>
	<p>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。</p>
	<p>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。</p>

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...	74703.1C
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 ID 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。选择**顺序**或**关键字**来指定合并记录的方法。选择**关键字**可激活该对话框的下半部分。

- **顺序。**按照顺序合并记录，以便将每个输入中的第 n 个记录合并在一起，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。这就意味着，所创建记录的数目是最小数据集中的记录数。
- **关键字。**使用关键字字段（如交易标识）合并关键字字段中具有同一值的记录。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，如果具有相同关键字段值 A 的记录的其他字段中包含有不同的值： B 、 C 和 D ，则合并后的字段将对于 A 与值 B 、 A 与值 C 以及 A 与值 D 的每个组合都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。**使用此选项来指定合并的条件。请参阅主题第 64 页的『指定合并的条件』，了解更多信息。

可能的关键字。仅列出那些在所有输入数据源中字段名称完全匹配的字段。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。可以使用多个关键字段。可以通过“过滤”节点或源节点上的“过滤”选项卡来重命名不匹配的输入字段。

用于合并的关键字。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。要从列表中除去关键字段，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。

合并重复的关键字段。如果上面选择了多个关键字段，那么此选项将确保只有一个具有该名称的输出字段。默认情况下，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。如果禁用此选项，则必须使用合并节点对话框中的“过滤器”选项卡重命名或排除重复的关键字段。

仅包含匹配的记录(内部连接)。选择此选项将只合并完整的记录。

包含匹配和不匹配的记录(完全外部连接)。选择此选项将执行“完全外部连接”。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。

包含匹配的记录和选定的未匹配记录(部分外部连接)。选择此选项会对在子对话框中选择的表执行“部分外部连接”。单击 **选择** 可指定将在合并中为其保留不完整记录的表。

包含第一个数据集中不与任何其他记录相匹配的记录(反连接)。选择此选项将执行“反连接”类型，其中只有第一个数据集中的不匹配记录会传递到下游。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。请参阅主题第 61 页的『连接类型』，了解更多信息。

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。请参阅主题第 61 页的『连接类型』，了解更多信息。

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。

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。请参阅主题第 101 页的『过滤或重命名字段』，了解更多信息。

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，可以更改默认标记。

- **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，也可以选择将其除去。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。

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 11. 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(Y)
Partial	如果较大数据集中没有不完整记录, 是。
全功能	否(N)
反连接	如果较大数据集是第一个输入, 是。

所有输入已按关键字段排序。 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关键字段(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业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 **Append** 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

匹配字段方法。 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。

字段来源。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。

生成新字段，显示记录的来源数据集。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。该默认字段名为**输入**。

区分节点

必须除去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”节点来查找或除去数据集中的重复记录，或者根据各条记录创建单条组合记录。

通过使用“区分”节点，您可以将第一个可区分记录遍历到数据流，以删除重复的记录，或者废弃第一个记录而将任何重复记录遍历到数据流，以查找重复的记录。

另外，可以在每个区分关键字值中为返回结果定义一种排序顺序。如果您需要为每个区分关键字返回特定行，则必须在区分节点中对记录排序，而不能使用上游排序节点（请参阅以下的“排序区分节点中的记录”）。

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。**为您提供了一种对非数字字段进行汇总的方式；选择此选项将使“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。请参阅第 69 页的『区分组合设置』以获取更多信息。
- **仅包括每个组中的第一个记录。**在数据流中包括第一个区分记录，并除去任何重复记录。
- **仅废弃每个组中的第一个记录。**丢弃发现的第一个区分记录，而将任何重复记录传递到数据流。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。可以进行下列操作：

- 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。列出用于确定记录在每个区分关键字值中的排序方式以及记录是按升序还是降序排序的字段。可以进行下列操作：

- 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。

缺省排序顺序。默认情况下，指定是否按照**升序**或**降序**顺序排列记录。

排序区分节点中的记录

在区分节点中使用**组内记录的排序依据**选项，可以为每个区分关键字返回特定行，并且不需要在前面使用排序节点。例如，假定我们具有以下处方药品用户年龄的数据。

表 12. 处方药用户数据.

年龄	Drug
50	药品 A
71	药品 B
44	药品 A

表 12. 处方药用户数据 (续).

年龄	Drug
65	药品 X
39	药品 A
75	药品 C
72	药品 Y
57	药品 X
79	药品 Y
69	药品 C
74	药品 B
85	药品 Y
69	药品 X

要找到每种药品年龄最大的用户，我们可将模式设为仅包括每个组中的第一个记录，将“药品”用作关键字段，将“年龄”用作排序方式字段，设置降序排序。输入顺序不影响结果，因为排序选择指定应返回给定药品的哪些行，最终数据输出应如下：

表 13. 排序后的处方药用户数据.

年龄	Drug
50	药品 A
74	药品 B
75	药品 C
69	药品 X
85	药品 Y

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。

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

已经按“设置”选项卡上的**分组字段和排序字段对输入数据集进行排序**。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。如果已安装 IBM SPSS Modeler Entity Analytics，那么还可以使用此工具组合（或序列化）从 SPSS Entity Analytics 输出的重复记录。

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。

例如，假设 SPSS Entity Analytics 将三条记录标记为同一实体，如下表中所示。

表 14. 同一实体的多条记录示例.

SEA-ID	姓名	年龄	银行	最高学历	总债务
0003	Bob Jones	27	K	学士	27000
0003	Robert Jones	35	N	硕士	42000
0003	Robbie Jones	27	D	博士	7000

我们希望将这三条记录汇总为随后在下游使用的单条记录。可以使用“汇总”节点将总债务相加并计算平均年龄；但是，无法对姓名和银行等详细信息求平均值。如果指定了使用哪些详细信息来创建组合记录，那么可以派生单条记录。

通过选择下列详细信息，可以根据该表创建组合记录。

- 对于**姓名**，使用第一条记录
- 对于**年龄**，采用最大值
- 对于**银行**，连接所有值（不带定界符）
- 对于**最高学历**，采用列表（博士 硕士 学士）中的第一项
- 对于**债务**，采用总计

通过组合（或汇总）这些详细信息，最终获得包含下列详细信息的单条组合记录。

- 姓名: Bob Jones
- 年龄: 35
- 银行: KND
- 最高学历: 博士
- 债务: 76000

这使我们能够最好地了解 Bob Jones: 博士学历、至少 35 岁，至少有三个已知的银行账户，并且有大笔总债务。

为“组合”选项卡设置选项

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；如果未连接节点，那么将不显示任何字段。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
- 总数
- 均数
- 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
 - 最近
 - 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。请参阅『区分组合 - 定制选项卡』以获取更多信息。

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。

- **Use。**选择“频率最高”或“频率最低”。
- **结。**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。
- **包括值 (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

- **连接值。**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 TS”节点

可以使用“流式时间序列”(TS) 节点在一个步骤内对时间序列模型进行构建和评分，并且无论是否有上游“时间区间”节点，您都可以使用此节点。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。

并非始终需要将“时间区间”节点与“流式 TS”节点配合使用，前提是满足以下几点：

- 数据等间隔分布。
- 数据具有时间序列标签。
- 不需要添加其他记录。
- 数据具有标准周期性，例如每周五天或每天八小时。

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。如果您的数据尚不满足此要求，那么上游“时间区间”节点可根据需要变换值。请参阅主题第 125 页的『时间区间节点』，了解更多信息。

以下是与“时间区间”节点相关的其他注意事项：

- 字段必须是数字。
- 不得将日期字段用作输入。
- 忽略分区。

“流式 TS”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 (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进行预测。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。通常，有关时间序列模型的更多信息，请参阅《IBM SPSS Modeler 16 建模节点》文档中的主题“时间序列模型”。

通过 IBM SPSS Modeler Solution Publisher、使用 IBM SPSS Collaboration and Deployment Services 评分服务或 IBM InfoSphere Warehouse，支持在流式部署环境中使用“流式 TS”节点。

流式 TS 字段选项

在“字段”选项卡中，您可以指定要在构建模型时使用的字段。在构建模型之前，需要指定要将哪些字段用作目标和输入。通常，“流式 TS”节点使用来自上游“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。这是默认选项。

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，请指定下面的字段。请注意，不接受存储为日期的字段作为目标字段或输入字段。

- **目标。**选择一个或多个目标字段。此操作类似于在“类型”节点中将字段角色设置为目标。时间序列模型的目标字段的测量级别必须为连续。为每个目标字段创建了单独的模型。目标字段将除其自身以外的所有已指定的输入字段视为可能的输入。因此，可以将同一字段包含在两个列表中；此类字段将用作所有模型（除此字段作为目标的模型以外）的可能输入。
- **输入。**选择输入字段。此操作类似于在“类型”节点中将字段角色设置为输入。时间序列模型的输入字段必须是数字。

流式 TS 模型选项

方法。您可以选择专家建模器、指数平滑法或 ARIMA。选择**标准**可指定所选方法的选项。

- **专家建模器。**选择此选项可使用专家建模器，此方法用于自动查找每个序列的最佳拟合模型。
- **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。

置信限制上限和下限的派生字段。选择此选项可针对模型预测和残差自相关的置信区间生成字段。选择此选项后，将启用**置信限制宽度 (%)**控件。

置信限制宽度 (%)。指定所计算的置信区间。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。

时间区间选项

在对话框的此部分中，您可以选择使用来自上游“时间区间”节点的估计和预测规范，也可以在“流式 TS”节点上指定这些设置。

使用“时间区间”节点中的设置。选择此选项可使用有关在上游“时间区间”节点中设置的估计和预测规范的信息。

信息的第一行指示将从模型中排除所有记录还是将这些记录用作维持记录。

第二行提供有关在“时间区间”节点上指定的所有预测期的信息。

如果第一行为**未定义时间区间**，那么这指示未连接“时间区间”节点。尝试运行流时，此情况将导致发生错误。如果选择使用**“时间区间”节点中的设置**，那么必须包括来自“流式 TS”节点的上游“时间区间”节点。

指定设置。在“流式 TS”节点中不存在上游“时间区间”节点的情况下，选择此选项可指定估计和预测设置。

向上估计到。选择**最后一个有效观测值**可使用最后一个有效观测值之前的所有记录来估计模型。用于预测的记录开始于第一个无效的观测值。有效观测值是所有目标字段具有非空值的观测值。选择**从最后一个记录计数**可预留一些最新记录以用于预测。最新记录可以具有目标字段的值，尽管这不是必需的。如果选择了**从最后一个记录计数**，那么将启用**偏移量**控件，您可以在此控件中指定仅用于预测的记录数。缺省记录数为 4，最大记录数为 99。

注：将用于预测的记录必须具有所有输入字段的值。

时间区间。指定用于估计和预测的时间区间。可用的选项如下：年、季度、月、周、天、小时、分钟和秒。

时间序列专家建模器标准

模型类型。可用的选项如下：

- **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的情况下，才会启用此选项。如果选择此选项，那么专家建模器将同时考虑季节性模型和非季节性模型。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。

事件和干预。允许您指定一些输入字段作为事件字段或干预字段。执行此操作会将字段标识为包含事件（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。对于标识为事件或干预字段的输入，专家建模器不会考虑任何转换函数。

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。

离群值

自动检测离群值。缺省情况下，不会执行离群值的自动检测。选择此选项以执行离群值的自动检测，然后选择所需的离群值类型。请参阅主题第 75 页的『处理离群值』，了解更多信息。

时间序列指数平滑法标准

模型类型。指数平滑法模型分类为季节性或非季节性¹。只有在周期为季节性时，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。

- **简单。**此模型适用于不具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。
- **霍特线性趋势。**此模型适用于具有线性趋势但没有季节性的序列。其相关平滑参数为水平和趋势，并且在此模型中，这两个参数不受彼此的值限制。霍特模型比布朗模型更加常用，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。
- **布朗线性趋势。**此模型适用于具有线性趋势但没有季节性的序列。其相关平滑参数为水平和趋势，但在此模型中，假定这两个参数等效。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。
- **阻尼趋势。**此模型适用于具有逐渐消失的线性趋势但没有季节性的序列。其相关平滑参数为水平、趋势和阻尼趋势。阻尼指数平滑法非常类似于自回归阶数为 1、差分阶数为 1 且移动平均值阶数为 2 的 ARIMA 模型。
- **简单季节性。**此模型适用于没有趋势且季节效应不随时间变化的序列。其相关平滑参数为水平和季节。季节指数平滑模型非常类似于自回归阶数为 0、差分阶数为 1、季节性差分阶数为 1，以及移动平均值阶数为 1、 p 和 $p+1$ （其中 p 是季节区间中的周期数）的 ARIMA 模型。对于每月数据， $p = 12$ 。
- **温特斯加性。**此模型适用于具有线性趋势且季节效应不随时间变化的序列。其相关平滑参数为水平、趋势和季节。温特斯加性指数平滑法非常类似于自回归阶数为 0、差分阶数为 1、季节性差分阶数为 1，以及移动平均值阶数为 $p+1$ （其中 p 是季节区间中的周期数）的 ARIMA 模型。对于每月数据， $p = 12$ 。

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4, 1-28.

- **温特斯乘法。**此模型适用于具有线性趋势且季节效应随序列量级发生变化的序列。其相关平滑参数为水平、趋势和季节。温特斯的可乘指数平滑法与任何 ARIMA 模型都不相似。

目标变换。可以指定对每个从属变量进行建模之前要对其执行的变换。请参阅《IBM SPSS Modeler 16 建模节点》文档中的主题“序列变换”以获取更多信息。

- **无。**未执行变换。
- **平方根。**将执行平方根变换。
- **自然对数。**将执行自然对数变换。

时间序列 ARIMA 标准

“时间序列”节点使您可以构建定制非季节性或季节性 ARIMA 模型（也称为 Box-Jenkins 模型），无论是否具有固定的一组输入（预测变量）变量²都是如此。您可以为任何或所有输入变量定义转换函数，并指定自动检测离群值或显式离群值集。

所有指定的输入变量都显式地包括在模型中。这与使用专家建模器相反。使用专家建模器时，只有在输入变量与目标变量具有统计学意义的显著关系时，才会包括这些输入变量。

模型

“模型”选项卡使您可以指定定制 ARIMA 模型的结构。

ARIMA 阶数。可以在“结构”网格的相应单元格中输入模型的各个 ARIMA 部分的值。所有值都必须是非负整数。对于自回归和移动平均值部分，值表示最大阶数。所有值为正的较低阶数都将包括在模型中。例如，如果指定 2，那么模型将包括阶数 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“季节性”列中的单元格。

- **自回归 (p)。**模型中自回归阶数的值。自回归阶数指定序列中的哪些先前值用于预测当前值。例如，自回归阶数 2 指定序列中过去两个周期的值将用于预测当前值。
- **差分 (d)。**指定估计模型之前应用于序列的差分阶数。在存在趋势的情况下，差分是必需的（存在趋势的序列通常不稳定，并且 ARIMA 建模假定平稳性），且用于消除趋势的影响。差分阶数对应于序列趋势的程度 - 第一阶差分考虑线性趋势，第二阶差分考虑二次趋势，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先前值的序列均值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，需要考虑过去两个周期中与每个周期的序列均值的偏差。

季节性阶。季节性自回归、移动平均值和差分部分与其非季节性对等项起着相同的作用。但是，对于季节性阶，当前序列值受到以一个或多个季节性周期分隔的先前序列值的影响。例如，对于每月数据（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的 12 个周期的序列值的影响。于是，对于每月数据，季节性阶 1 与指定非季节性阶 12 的效果相同。

目标变换。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。请参阅《IBM SPSS Modeler 16 建模节点》文档中的主题“序列变换”以获取更多信息。

- **无。**未执行变换。
- **平方根。**将执行平方根变换。
- **自然对数。**将执行自然对数变换。

2. Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. *Time series analysis: Forecasting and control*, 3rd ed. Englewood Cliffs, N.J.: Prentice Hall.

将常量包括在模型中。除非您确定序列总均数为 0，否则包括常量是标准操作。应用差分时，建议您排除常量。

转换函数

通过“转换函数”选项卡，您可以为任何或所有输入字段定义转换函数。使用转换函数可以指定如何使用这些字段过去的值来预测目标序列未来的值。

只有在指定了输入字段（角色设置为输入）的情况下才会显示此选项卡，输入字段在“类型”节点或“流式 TS”节点的“字段”选项卡上指定（选择**使用定制设置 - 输入**）。

顶部列表将显示所有输入字段。此对话框中剩余的信息特定于在列表中所选的输入字段。

转换函数阶数。可以在“结构”网格的相应单元格中输入转换函数的各个组成部分的值。所有值都必须是非负整数。对于分子和分母部分，值表示最大阶数。所有值为正的较低阶数都将包括在模型中。另外，对于分子部分，将始终包括阶数 0。例如，如果对分子指定 2，那么模型将包括阶数 2、1 和 0。如果对分母指定 3，那么模型将包括阶数 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“季节性”列中的单元格。

分子。转换函数的分子阶数指定使用所选独立（预测变量）序列中的哪些先前值来预测从属序列的当前值。例如，分子阶数 1 指定使用过去一个时间周期的独立序列的值（以及独立序列的当前值）来预测每个从属序列的当前值。

分母。转换函数的分母阶数指定如何使用与所选独立（预测变量）序列先前值的序列均数的偏差来预测从属序列的当前值。例如，分母阶数 1 指定在预测每个从属序列的当前值时，需要考虑与过去一个时间周期的独立序列的均值的偏差。

差分。指定在估计模型之前应用于所选独立（预测变量）序列的差分阶数。在存在趋势的情况下，差分是必需的，并用于消除趋势的影响。

季节性阶。季节性分子、分母和差分部分与其非季节性对等项起着相同的作用。但是，对于季节性阶，当前序列值受到以一个或多个季节性周期分隔的先前序列值的影响。例如，对于每月数据（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的 12 个周期的序列值的影响。于是，对于每月数据，季节性阶 1 与指定非季节性阶 12 的效果相同。

延迟。设置延迟将导致输入字段的影响延迟指定的时间间隔数。例如，如果将延迟设置为 5，那么直到经过 5 个周期 ($t + 5$) 之后，在 t 时间处的输入字段的值才会影响预测。

变换。一组独立变量的转换函数规范还包括将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。将执行自然对数变换。

处理离群值

“离群值”选项卡提供了一些用于处理数据中的离群值的选项³。

请不要检测离群值或对其进行建模。缺省情况下，不会检测离群值，也不会对其进行建模。选择此选项可禁用对离群值执行的任何检测或建模操作。

自动检测离群值。选择此选项以执行离群值的自动检测，并选择一个或多个所显示的离群值类型。

3. Pena, D., G. C. Tiao, and R. S. Tsay, eds. 2001. *A course in time series analysis*. New York: John Wiley and Sons.

要检测的离群值类型。选择要检测的离群值类型。受支持的类型为:

- 加性 (缺省值)
- 水平变动 (缺省值)
- 创新
- 瞬态
- 季节性加性
- 局部趋势
- 加性修补

请参阅《IBM SPSS Modeler 16 建模节点》文档中的主题“离群值”以获取更多信息。

流式 TS 部署选项

如果您正在流式部署环境中通过 IBM SPSS Modeler Solution Publisher 使用“流式 TS”节点,那么缺省情况下,每次使用时间序列模型对记录进行评分时,都将重建模型并进行新的预测。您可以使用“部署”选项卡上的选项来更改此缺省行为,以指定时间序列模型的重建时间。

注: 在 IBM SPSS Modeler 中,将忽略“部署”选项卡上的设置以正常使用“流式 TS”节点。

在需要时强制完成重建 (可能会更改模型类型)。如果您希望仅在指定数目的新记录的情况下才重建时间序列模型,请选择此选项。针对每三条或每四条新记录 (而不是每条新记录) 重建模型可以提高性能。选择此选项将启用需要进行重建前的最小新记录数控件。

需要进行重建前的最小新记录数。您可以根据新记录的数目选择时间序列模型的重建时间。请选择下列其中一个选项:

- 如果选择了**使用记录数目的百分比 (%)**,那么将在新记录数达到记录总数的指定百分比时重建时间序列模型。缺省百分比为 10%。
- 如果选择了**使用记录数目的绝对值**,那么将在新记录数达到指定数目时重建时间序列模型。可以为要重建的模型指定的新记录的最小数目为 1,最大数目为 100。缺省记录数为 4。

用于标识新记录的字段。选择在有新记录时进行观察以检测新记录的字段。所选择的字段必须是用于对时间序列中的记录进行排序的字段。只能选择连续字段。

“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:

- geohash - 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。选择用于标识纬度的字段。

经度字段。选择用于标识经度的字段。

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。请参阅『定义空间时间限制密度』以获取更多信息。

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。

STB 密度。选择空间和时间密度，将根据所选内容派生新字段。请参阅『定义空间时间限制密度』以获取更多信息。

实体标识字段。选择要用作逗留标识的实体。

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。

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。

允许逗留跨越 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，那么选择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。如果未选择此选项，那么所有逗留时间都必须在单一空间时间限制之内。

满足时间限制的事件的最小比例 (%)。只有在选择了**允许逗留跨越 STB 边界**的情况下，此选项才可用。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。例如，如果设置为 25%，并且事件的比例为 26%，那么这符合逗留条件。

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。

时间区间。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。

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。您可以完全自动化地使用节点，允许节点选择并应用修正，或者也可在修正前预览更改，按照需要接受、拒绝或修改。



“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空值的选项、设置用于建模的字段的角色、指定字段和值标签，以及为字段指定值。



“过滤”节点用于过滤（废弃）字段、重命名字段并将字段从一个源节点映射到另一源节点。



“派生”节点将修改数据值或根据一个或多个现有字段创建新字段。它可创建的字段类型包括公式、标志、名义、状态、计数和条件。



“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。



填充节点会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”节点经常结合“类型”节点使用以替换缺失值。



“匿名化”节点用于变换字段名和字段值的下游表示方式，从而掩饰了原始数据。如果要允许其他用户使用敏感数据（如客户名称或其他详细信息）构建模型，这种节点将十分有用。



“重新分类”节点可将一组分类值转换为另一组值。对于压缩类别或为分析而进行的数据重新分组，重新分类非常有用。



“分级”节点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（集合）字段。例如，用户可将连续收入字段转换为一个包含各组收入的新的分类字段，作为其与平均值之间的偏差。一旦创建新字段分级后，即可根据分割点创建“派生”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。



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”节点根据为一个或多个名义字段定义的分类值获取多个标志字段。



“重构”节点可将一个名义字段或标志字段转换为一组字段（该字段组由已成为另一字段的值填充）。例如，给定一个名为 支付类型 的字段，其值为 贷方、现金和借方，则将创建三个新字段（贷方、现金、借方），每个字段可能包含实际支付的值。



转置节点交换行和列中的数据，以便记录变成字段，字段变成记录。



时间区间节点指定区间，创建用于对时间序列数据进行建模的标签（如果需要）。如果值的间隔不是均匀的，那么此节点会根据需要填充值或将值集中起来以生成均匀的记录区间。



“历史记录”节点将创建新字段，其中包含之前记录中的字段数据。“历史记录”节点最常用于顺序数据，如时间序列数据。使用“历史记录”节点前，您可能希望使用“排序”节点对此数据进行排序。



“字段重排”节点定义了用于显示下游字段的自然顺序。此顺序将影响字段在多个位置的显示方式，如表格、列表和字段选择器。处理大型数据集时，此操作有助于使所需字段更为直观。

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。请参阅主题第 243 页的『生成其他用于数据准备的节点』，了解更多信息。

自动准备数据

准备数据以进行分析是任何项目中最重要的一步，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，您可以快速、轻松地准备数据以供建模，而无需具备相关统计概念的预备知识。您可以更快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。

注意：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。旧字段不用于进一步分析，其角色被设置为“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次序。

- **均衡速度和精确度。**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。
- **优化速度。**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。
- **优化准确性。**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。**培训** 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，就不需要再次培训。这允许您在典型数据子集上培训该节点，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得分字段。例如，如果要针对不同输入生成得分字段图表，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。请参阅主题第 92 页的『生成派生节点』，了解更多信息。

- **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。

“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。这是默认选项。

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默认设置进行了更改，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。

何处理建模中排除的字段。 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。

注意： 此操作还将应用到已变换的目标。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。

如果传入字段与现有分析不匹配。指定在您执行经过培训的 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。

准备日期和时间以供建模。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。

计算到参考日期为止已过去的时间。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。
- 日期持续时间的单位。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。

计算到参考时间为止已过去的时间。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。
- 持续时间的单位。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。

抽取循环时间元素。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。

- 从日期提取。对于任何日期输入，请指定是否要提取年、月、日或任意组合。
- 从时间提取。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。

排除低质量的输入字段。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。

排除唯一类别过多的名义字段。除去类别超过个数的字段，而不会用于进一步分析。请指定一个正整数。默认值为 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。除去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。

注意：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。

对输入和目标字段进行建模准备。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值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。
- **重新排序名义字段。**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：用均值替换缺失值。**选择本选项以替换连续（范围）特征的缺失值。
- **名义字段：用众数替换缺失值。**选择本选项以替换名义（集合）特征的缺失值。
- **有序字段：用中位数替换缺失值。**选择本选项以替换有序（有序集合）特征的缺失值。

有序字段值的最大数目。指定重新定义有序（有序集合）字段为连续（范围）的阈值。默认值为 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）。

连续字段值的最小数目。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。默认值为 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）。

离群值分界值。指定离群值截断标准（采用标准差测量），默认值为 3。

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置于常用尺度上。要标准化连续输入字段，选中本复选框并选择正态化方法。默认为 **z 得分转换**，其中您可以指定**最终均值**（默认值为 0），以及**最终标准差**（默认值为 1）。或者也可以选择使用**最小/最大转换**并指定最小值和最大值，默认值分别为 0 和 100。

当您在“构建&选择特征”面板上选择**执行特征构建**时，本字段特别有用。

重新调整具有 Box-Cox 转换的连续目标。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**默认值为 0，同时**最终标准差**默认值为 1。

注意：如果您选择标准化目标，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。请参阅主题第 92 页的『生成派生节点』，了解更多信息。

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。

注意：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。

变换、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。

合并松散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 的默认概率值。

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。

没有目标时，基于计数合并稀疏类别。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个案或记录的最小百分比，默认值为 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个案百分比均不少于指定最小个案百分比，合并将停止。

在保留预测能力的同时分级连续字段。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 的默认齐次子集概率值。

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。

注意：ADP 中的分级与 IBM SPSS Modeler 其他部分的最佳分级不同。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。选择此选项将除去相关系数低的特征。如果需要，更改 0.05 的默认概率值。

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。

执行特征构建。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。

已变换和构造的字段。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名称扩展设置为“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得出的持续时间。指定要应用到从日期和时间计算的持续时间的名称扩展。

从日期和时间抽取的周期元素。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得分的建议。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）。请参阅主题『字段处理概要』，了解更多信息。
- 字段。请参阅主题第 87 页的『字段』，了解更多信息。
- 操作摘要。请参阅主题第 88 页的『操作摘要』，了解更多信息。

有四个链接/辅助视图：

- 预测能力（缺省视图）。请参阅主题第 88 页的『预测能力』，了解更多信息。
- 字段表。请参阅主题第 88 页的『字段表』，了解更多信息。
- 字段详细信息。请参阅主题第 89 页的『字段详细信息』，了解更多信息。
- 操作详细信息。请参阅主题第 89 页的『操作详细信息』，了解更多信息。

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

“字段处理概要”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。请参阅主题第 88 页的『**字段表**』以获取更多信息。**建议在分析中使用的特征**在“预测能力”链接视图中显示。请参阅主题第 88 页的『**预测能力**』，了解更多信息。

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。默认情况下，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**。

注意：对于同时具有原始和已变换版本的预测变量，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。请参阅主题第 89 页的『字段详细信息』，了解更多信息。
- **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**带有下列线，单击可在链接视图中显示有关所执行操作的更多信息。请参阅主题第 89 页的『操作详细信息』，了解更多信息。

注意：只会显示每个字段的原始和最终变换版本，而不会显示在分析过程中使用的任何中间版本。

预测能力

在首次运行分析时默认显示，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“字段处理概要”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。

注意：如果字段因为类别太多而被排除，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。

注意：如果未定义目标，或者在“主视图”面板中单击了目标，那么将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序列列出。不过，对于某些步骤，可能对特定字段执行了多个操作。

注意：对于未变换的字段，将不显示此表格。

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。请参阅主题『操作详细信息』，了解更多信息。
- **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， $0.06*\text{age} + 1.21*\text{height}$ 。

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。

注意：如果关闭了某个操作或者未指定目标，那么在“操作摘要”主视图中单击此操作时将显示一条错误消息，而不是显示操作详细信息。

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、指定的平均值和标准偏差。
- 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得分字段。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”节点生成的得分字段名称。如果尺度（范围）目标已转换，则得分字段以转换后的单位进行显示，例如，以 $\log(\$)$ 代替 $\$$ 。要解释和使用结果，您必须将预测值转换回原始尺度。

注意：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。

注意：如果未使用自动建模节点或整体节点，那么您将需要编辑“派生”节点以便为模型转换正确的得分字段。

标准化连续目标

缺省情况下，如果您在“准备输入和目标”面板上选中了使用 **Box-Cox 转换重新标度连续目标** 复选框，那么这将转换目标，并且您所创建的新字段将成为模型构建的目标。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，因此您无法看到得分和解释。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。

派生节点会将得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得分字段。如果构建单独的模型，那么需要编辑“派生”节点，以便从实际得分字段中派生。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”节点除去原始目标字段（如果仍然存在），然后重新命名目标和得分字段。

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 中数据的值和字段标签。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，用于描述某个给定字段中数据的特征。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**。请参阅主题第 93 页的『测量级别』，了解更多信息。

注意：字段的测量级别与字段的存储类型不同，后者表明数据是作为字符串、整数、实数、日期、时间还是时间戳进行存储。

- **值**。此列允许您指定从数据集读取数据值的选项，或使用**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。请参阅主题第 96 页的『数据值』，了解更多信息。
- **缺失**。用于指定字段缺失值的处理方法。请参阅主题第 98 页的『定义缺失值』，了解更多信息。
- **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。请参阅主题第 98 页的『检查类型值』，了解更多信息。
- **角色**。用于告知建模节点字段将成为用于某个机器学习过程的**输入**（预测变量字段）还是**目标**（预测字段）。**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。请参阅主题第 99 页的『设置字段角色』，了解更多信息。

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**。请参阅主题第 95 页的『转换连续数据』，了解更多信息。
- 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”。默认情况下，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。请参阅主题第 99 页的『复制类型属性』，了解更多信息。
- 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）。请参阅主题第 19 页的『设置字段存储类型和格式』，了解更多信息。

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 15. 测量级别图标

图标	测量级别
	缺省
	连续
	分类
	标记
	名义
	有序
	无类型

可以使用以下测量级别:

- **缺省值。**具有未知存储类型和值的数据（例如，由于其尚未被读取）将显示为<默认值>。
- **连续。**用于描述数值，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。**用于字符串值（可区分值的确切数目未知时）。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“流属性”对话框中指定的名义字段成员最大数。
- **标志。**用于具有两个不同值以指示特征存在与否（例如 true 与 false、Yes 与 No 或 0 与 1）的数据。所用值可能有所不同，但其中总会有个值代表“真”值，另一个代表“假”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，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但值不同的数据。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 是某个整数集合的默认排列顺序，而 HIGH, LOW, NORMAL（按字母升序）是某个字符串集合的顺序。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。**用于不属于任何上述类型的数据，具有单个值的字段，或集合的成员数超过定义的最大值的名义数据。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。

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。请参阅主题第 95 页的『转换连续数据』，了解更多信息。

使用自动归类

1. 在类型节点中或源节点的“类型”选项卡中，将所需字段的值列设置为 **<读取>**。此操作将使元数据可用于所有下游节点。可以使用对话框中的墨镜按钮将所有字段快速设置为 **<读取>** 或 **<传递>**。
2. 单击 **读取值** 可立即读取数据源中的值。

要手动设置字段测量级别

1. 选择表中的某个字段。
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。

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。为实现系统资源最优化，实例化是一种用户导向过程 - 由用户通过在源节点的“类型”选项卡中指定选项或通过类型节点运行数据来指导软件读取值。

- 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**<默认>**。
- 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。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。所有数据通过节点后，它们将变为完全实例化数据，但设置为 **<传递>** 的值除外。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更改或更新值，需要在“类型”选项卡中对其进行编辑，或将字段的值设置为 **<读取>** 或 **<读取 +>**。

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。

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 16. 自动输入的指示信息

选项	功能
<读取>	将在执行节点时读取数据
<读取+>	将读取数据并将其附加到当前数据（如果存在）。
<遍历>	未读取数据。
<当前>	保留当前数据值。
指定...	启动单独的对话框，用于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”选项或通过单击字段列中的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，可以使用对话框工具栏中的下列按钮重置值信息：

- 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测量级别十分有用。

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。在此列表上选择指定选项将打开一个单独的对话框，您可在其中为所选字段设置读取、指定、标注和处理值的选项。

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。显示当前选定的测量级别。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。

存储类型。显示已知的存储类型。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所生成的字段，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和建模时字段的角色（预测值、概率和倾向等）。

值。选择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。读取值的选项包括：

- **读取数据。**选择此选项将在执行节点时读取数据。此选项与 **<读取>** 相同。
- **遍历。**选择此选项将不为当前字段读取数据。此选项与 **<传递>** 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。将此选项与值检查结合使用，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。注意：不能为测量级别为无类型或 **<缺省>** 的字段指定值或标签。
- **从数据扩展值。**选择此选项可将在此输入的值追加到当前数据。例如，如果 *field_1* 的范围为 (0,10)，您输入的值范围为 (8,16)，则会通过添加 16 来扩展范围，而不删除原始最小值。新的范围将是 (0,16)。选择此选项会自动将自动归类选项设置为 **<读取+>**。

检查值。选择强制转换值以符合指定的连续、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。将值检查与“指定值”选项结合使用，可以使数据中的值符合预期值。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。选择此选项可激活下面的控件，这些控件可用于声明数据中的缺失值或空白值。

- **缺失值表。**可以在此将特定值（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。如果将某个限制值留为空白，那么范围将不受限制；例如，如果仅指定下限为 100 而没有指定上限，那么会将所有大于或等于 100 的值定义为缺失值。限制值包括在内；例如，下限为 5、上限为 10 的范围的定义中将包括 5 和 10。可以为任意存储类型定义缺失值范围，包括日期/时间和字符串（这时将采用字母排列顺序确定某个值是否在范围内）。
- **空值/空白。**您还可以将系统空值（在数据中显示为 \$null\$）和空白（没有可见字符的字符串值）指定为空白值。注意：出于分析的需要，“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，并且在某些情况下会以不同方式进行处理。

注意：要将空白值编码为未定义值或 \$null\$，应使用“过滤”节点。

描述。使用此文本框可指定字段标签。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(D)

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。使用表中的值列，可以根据您对当前字段的了解指定值。使用此表，可以输入字段的预期值，并使用“检查值”下拉列表检查数据集与这些值的一致性。使用箭头键和删除键可以修改现有值，以及对值进行重新排序或删除值。

标签。使用标签列可以为集合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。

指定标志的值

标志字段用于显示具有两个不同值的数据。标志的存储类型可以是字符串、整数、实数或日期/时间。

真。指定条件成立时字段的标志值。

假。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。

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：

On (*)。指示已为此字段定义缺失值处理。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。

Off。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。要为所有字段设置“检查”设置，请单击字段列并按 **Ctrl+A**。然后使用检查列中任意字段的下拉列表。

可用的“检查”设置如下：

无。将遍历值而不进行检查。这是默认设置。

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。 如果找到非法值，那么将废弃整条记录。

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）。

注意：“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。请参阅主题第 121 页的『分区节点』，了解更多信息。

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**。注意：可以采用按住 Ctrl 并单击的方法或使用上下文菜单中的**选择字段**选项选择多个字段。

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。
- **检查。**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。请参阅主题『设置字段格式选项』，了解更多信息。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：Statistics 导出节点和 Statistics 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。默认设置为 **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** 来覆盖默认设置。

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。

- **全部选中。**选中所有字段。
- **全部不选。**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**。请参阅主题第 93 页的『测量级别』，了解更多信息。
- **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。

小数符号。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“定义的区域设置”（在该情况下将采用当前区域设置的默认设置）。

小数位（标准、科学表示法、货币、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。

对齐。指定列中的值应采用的对齐方式。默认设置为 **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”来覆盖默认设置。

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，作为医学研究者，您可能并不关心患者（记录级数据）的钾等级（字段级数据）；因此，可以将 *K*（钾）字段过滤掉。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、Statistics 文件和 XML）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 Statistics 导出、Statistics 变换、Statistics 模型和 Statistics 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。请参阅主题第 291 页的『重命名或过滤 IBM SPSS Statistics 的字段』，了解更多信息。
- 可以使用上述任意节点中的“过滤”选项卡来定义或编辑多重响应集。请参阅主题第 103 页的『编辑多重响应集』，了解更多信息。
- 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。此选项默认处于选中状态，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。注意：选择此选项将除去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。请参阅主题第 291 页的『重命名或过滤 IBM SPSS Statistics 的字段』，了解更多信息。
- 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。请参阅主题第 103 页的『编辑多重响应集』，了解更多信息。
- 设置默认过滤状态。

您还可以使用对话框顶部的箭头切换按钮指定要在默认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，这种方法十分有用。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。

数字位数。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 17. 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

通过单击左上角的过滤按钮菜单并选择**匿名化字段名**，可匿名化包含“过滤”选项卡的任意节点中的字段名。匿名化的字段名包含一个字符串前缀加一个基于数字的唯一值。

对名称进行匿名化。选择**仅选定字段**将仅对“过滤”选项卡中已选定的字段的名称进行匿名化。默认设置为**所有字段**，选择该选项将对所有字段名进行匿名化。

字段名称前缀。匿名化字段名的默认前缀为 **anon_**；如果要使用其他前缀，请选择**自定义**并键入自己的前缀。

对多个响应集进行匿名化。匿名化多重响应集名称的方法与匿名化字段名的方法相同。请参阅主题『编辑多重响应集』，了解更多信息。

要恢复原始字段名，请在过滤按钮菜单中选择 **使用输入字段名**。

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 Data Collection 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删除字段。

注释(O)

- 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除去通过“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。

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，从而使整体准确性更高。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得分的字段。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 18. 整体方法字段创建。

整体方法	字段名称(F)
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得分。

过滤出整体模型生成的字段。从输出中除去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得分，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得分的准确性与各个输入模型得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，将会对其得分求平均值。这是求综合得分唯一可用的方法。

求得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。默认对新模型生成标准误差差计算；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，若五个模型中有三个都预测为 **是**，其他两个预测为 **否**，则是以 3 比 2 的投票结果取胜。另外也可以根据每个预测的置信度或倾向值进行 **加权** 投票。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。
- **原始或调整后的倾向（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：

$\text{abs}(0.5 - \text{propensity}) * 2$

或者，对于调整后的倾向，绝对倾向计算方法如下：

$\text{abs}(0.5 - \text{adjusted propensity}) * 2$

导出节点

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”节点将修改数据值或根据一个或多个现有字段创建新字段。它可创建的字段类型包括公式、标志、名义、状态、计数和条件。



“重新分类”节点可将一组分类值转换为另一组值。对于压缩类别或为分析而进行的数据重新分组，重新分类非常有用。



“分级”节点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（集合）字段。例如，用户可将连续收入字段转换为一个包含各组收入的新的分类字段，作为其与平均值之间的偏差。一旦创建新字段分级后，即可根据分割点创建“派生”节点。



“设为标志”节点根据为一个或多个名义字段定义的分类值获取多个标志字段。



“重构”节点可将一个名义字段或标志字段转换为一组字段（该字段组由已成为另一字段的值填充）。例如，给定一个名为 **支付类型** 的字段，其值为 **贷方**、**现金** 和 **借方**，则将创建三个新字段（**贷方**、**现金**、**借方**），每个字段可能包含实际支付的值。



“历史记录”节点将创建新字段，其中包含之前记录中的字段数据。“历史记录”节点最常用于顺序数据，如时间序列数据。使用“历史记录”节点前，您可能希望使用“排序”节点对此数据进行排序。

使用导出节点

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。

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。默认名称为 **DeriveN**，其中 **N** 是截止到目前由当前会话所创建的派生节点数。

派生为。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。

注意：对于同时派生多个字段，@FIELD 函数是一种重要的工具。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，用于对多个字段应用对数转换的 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）。默认名称为 *DeriveN*，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。请记住，在引用多个字段的操作时可使用 @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

派生公式节点用于根据 CLEM 表达式的结果为数据集中的每条记录创建新字段。请注意，此表达式不能是条件表达式。要根据条件表达式导出值，请使用标志或条件类型的导出节点。

公式。使用 CLEM 语言指定用于为新字段派生值的公式。

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。

真值。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。默认值为 T。

假值。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。默认值为 F。

以下情况时为真。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
```

要准备分析该数据，可以使用 `hassubstring` 函数为每个响应生成单独的标志字段，所使用的表达式如下：

```
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。

若此条件为真。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，则会使用默认值。

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。

以下情况时切换“On”。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。

以下情况时切换“Off”。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。

以下情况时增加。 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。

增加的值。 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。

以下情况时重置。 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. 在“派生为”中选择适当的字段类型 (“名义”、“标志”等)。
2. 指定对值进行重新编码的条件。例如, 可以设置为: 如果 Drug='drugA' 则值为 1, 如果 Drug='drugB' 则值为 2, 依此类推。

填充节点

填充节点用于替换字段值和更改存储类型。您可以选择基于指定的 CLEM 条件 (如 @BLANK(FIELD)) 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, 可以通过指定表达式 (如 @GLOBAL_MEAN) 为空白值填充字段的均值。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。

填入字段。 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。

注意: 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。采用标准条件 `@BLANK(@FIELD)` 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。采用标准条件 `@NULL(@FIELD)` 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。

注意：如果所选字段为字符串，那么应将其替换为字符串值。使用默认值 0 或其他数字值作为字符串字段的替换值将产生错误。

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** 作为流的默认日期格式。

“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。请参阅主题第 109 页的『使用导出节点对值进行重新编码』，了解更多信息。

匿名化节点

处理要在节点下游模型中包含的数据时，可使用 `Anonymize` 节点对字段名、字段值或这两者进行掩饰。这样，便可以随意分发生成的模型（例如分发至技术支持部门），而无需担心非授权用户能够查看机密数据（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，这具体取决于 Anonymize 节点在流中的放置位置。例如，如果通过选择节点在上游插入一个匿名化节点，那么当该选择节点中的选择标准作用于现已匿名化的值时，这些标准将需要更改。

用于匿名化的方法取决于多种因素。对于字段名以及除“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默认字符串 *anon_*，*n* 是整数，该值从 0 开始，并在遇到每个唯一值时递增（例如，*anon_S0*、*anon_S1* 等）。

类型为“连续”的字段值必须进行转换，因为数值范围处理的是整数或实数，而不是字符串。因此，只能通过将范围变换为不同范围对其进行匿名化，从而掩饰原始数据。范围内的值 *x* 的转换按以下方法执行：

$A*(x + B)$

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

例如，现有字段 *AGE*，其中尺度因子 *A* 设为 7，转换偏移量 *B* 设为 3，*AGE* 的值将转换为：

$7*(AGE + 3)$

为匿名化节点设置选项

您可在选择要在下游对其值进行掩饰的字段。

请注意，必须先匿名化节点的上游对数据字段进行实例化，之后才能执行匿名化操作。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。如果有任何字段名已匿名化，则会在此显示匿名化的名称。

测量。字段的测量级别。

对值进行匿名化。选择一个或多个字段，单击此列，然后选择**是**将使用默认前缀 *anon_* 对字段值进行匿名化；选择**指定**将显示一个对话框，您可在其中输入自己的前缀，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，不能在单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进行此操作。

查看当前字段。选择此选项可查看当前连接到 Anonymize 节点的数据集的字段。默认情况下，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，此选项将十分有用。

指定如何对字段值进行匿名化

使用“替换值”对话框，可以选择要对匿名化字段值使用默认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“匿名化值”设置更改为**是**。

字段值前缀。匿名化字段值的默认前缀为 **anon_**；如果要使用其他前缀，请选择**自定义**并输入自己的前缀。

“转换值”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转换要采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。选择此选项可为转换指定您自己的值。

- **乘数。**字段值在转换中的乘数。最小值为 1；最大值通常为 10，但有时需要降低该值以避免溢出。
- **增加的数字。**在转换中将为字段值增加的数字。最小值为 0；最大值通常为 1000，但有时需要降低该值以避免溢出。

匿名化字段值

已在“设置”选项卡上选中进行匿名化的字段，其值将在以下情况被匿名化：

- 当您运行包含匿名化节点的流时
- 当您预览值时

要预览值，请单击“匿名化值”选项卡中的**匿名化值**按钮。下面，从下拉列表中选择一个字段名。

如果测量级别为“连续”，则会显示以下内容：

- 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”，屏幕将显示该字段的原始值和匿名化值。

如果显示黄色背景，那么表示所选字段的设置自上次匿名化值以来已发生更改，或已对 **Anonymize** 节点的上游数据进行更改，使匿名化值不再正确。此时将显示当前值的集合；再次单击 **匿名化值** 按钮可根据当前设置生成一组新的值。

对值进行匿名化。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”的字段采用随机播种，每次单击此按钮都会创建一组新的值。

清除值。将原始值和匿名化值从表中清除。

重新分类节点

“重新分类”节点可实现一组分类值到另一组分类值的转换。对于压缩类别或为分析而进行的数据重新分组，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。请参阅主题第 176 页的『使用分布节点』，了解更多信息。

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。请参阅主题第 114 页的『分级节点』，了解更多信息。

- 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。请参阅主题第 109 页的『使用导出节点对值进行重新编码』，了解更多信息。

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。

重新分类为。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。

重新分类字段。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。请参阅主题第 114 页的『对多个字段进行重新分类』，了解更多信息。

重新分类值。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。**此列列出所选字段的现有值。
 - **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区域设置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。

用于未指定的值。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**），也可以指定默认值。

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。指定扩展部分（如 `_recode`），然后选择要将其附加到原始字段名的后面还是前面。

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默认值，存储类型将按以下方式决定：检查新值和默认值，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，使用重新分类 $4 \Rightarrow 0, 5 \Rightarrow 0$ 对整数集合 { 1, 2, 3, 4, 5 } 进行重新分类将生成新的整数集合 { 1, 2, 3, 0 }，而使用 $4 \Rightarrow \text{“Over 3”}, 5 \Rightarrow \text{“Over 3”}$ 则会生成字符串集合 { “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。

分级节点

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。

何时使用分箱节点

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。请参阅主题第 105 页的『导出节点』，了解更多信息。
- 要为现有集合创建新类别，请使用重新分类节点。请参阅主题第 112 页的『重新分类节点』，了解更多信息。

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\$)。**在分级转换过程中,空值将被忽略,并在转换之后保持空值。

“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,可以同时多个字段进行分级。使用右侧的按钮可添加或删除字段。

分级方法。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名。*_BIN* 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_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,可以使用默认值 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 19. 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 至 <23	>=23 至 <33	>=33 至 <43	>=43 至 <53	>=53 至 <63	>=63 至 <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: $18 [最低数据值] - 5 [0.5 \times (分级宽度 10)] = 13$ 。

分级数。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。请参阅主题第 119 页的『预览生成的分级』，了解更多信息。

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。默认扩展名为 `_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。默认值为 `_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。尽管每个十分位数包含的底方数大致相同，但各人在其中拥有的底方数并不相同，所开底方最多的个人集中在十分位数 10 中。请注意，此方法会假定所有值均大于零，如果实际情况不是这样则可能产生意外结果。

结。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。

- **添加到下一个分级。**选择此选项可将结值上移至下一个分级。
- **保留在当前分级中。**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 20. 按记录计数分位的示例.

值(V)	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:

值的总数/分位数

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** 模式下, 会将其添加到分级 2 中。在 **保留在当前分级中** 模式下, 会将其留在分级 1 中, 从而将分级 4 的值范围不在现有数据值的范围内。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 21. 分级示例结果.

分级	下限	上限
1	>=10	<15
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。

个案排秩

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。

排序顺序。选择**升序** (将最低值标记为 1) 或**降序** (将最高值标记为 1)。

排序。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。

分数排序值。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。

分数排序百分比值。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。

扩展名。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 标准差。**选择此选项将生成三个分级。
- **+/- 2 标准差。**选择此选项将生成五个分级。
- **+/- 3 标准差。**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 22. 标准差分级示例.

分级 1	分级 2	分级 3
$x < (\text{Mean} - \text{Std. Dev})$	$(\text{Mean} - \text{Std. Dev}) \leq x \leq (\text{Mean} + \text{Std. Dev})$	$x > (\text{Mean} + \text{Std. Dev})$

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 *取消赎回权百分比* 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，以及要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 *pastdue_OPTIMAL* 的新字段，以及另一个名为 *inforeclosure_OPTIMAL* 的新字段。

主管字段。用于构造分级的分类字段。

预分级字段以提高大型数据集的性能。表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。

分级端点。可以指定应包括（下限 $\leq x$ ）还是排除（下限 $< x$ ）端点下限或上限。

第一个分级和最后一个分级。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。

分位数。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，会分配给客户一个得分；然后将这些得分组合在一起，从而得到 RFM 的总分值。此得分表示为每个 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 得分并对这些得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：不应忽视得分低的客户，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“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。请参阅主题第 59 页的『RFM 汇总节点』，了解更多信息。

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，它们分别接近因、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。

间隔数。 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 得分的计算方法如下： $(\text{近因分值} \times \text{近因权重}) + (\text{频率分值} \times \text{频率权重}) + (\text{货币分值} \times \text{货币权重})$ 。

结。 指定如何分级相同的得分。选项为：

- **添加到下一个分级。** 选择此选项可将结值上移至下一个分级。
- **保留在当前分级中。** 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。（这是默认值。）

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：

- 近因得分。近因排序（分级值）
- 频率得分。频率排序（分级值）
- 消费金额得分。货币排序（分级值）
- RFM 得分。近因、频率和货币得分的加权和。

将离群值添加到最终分级。 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。

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。请参阅主题第 99 页的『设置字段角色』，了解更多信息。

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。请参阅主题第 105 页的『导出节点』，了解更多信息。

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。

分区节点选项

分区字段。指定由该节点创建的字段名称。

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区域设置之间移动，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。
- **向系统定义的值追加标签。**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区域设置。
- **将标签用作值。**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，这也使数据特定于区域设置，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。

设置随机种子。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。

注意：为从数据库中读取的记录选择**设置随机种子**选项时，在抽样前需要使用 Sort 节点以确保每次执行节点时能得到同一结果。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。请参阅主题第 60 页的『排序节点』，了解更多信息。

启用 SQL 以分配记录到分区。（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**唯一字段**下拉列表中，选择具有唯一值的字段（例如 ID 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。请参阅主题第 10 页的『数据库源节点』，了解更多信息。

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

集合字段。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）。请参阅主题第 92 页的『类型节点』，了解更多信息。

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。默认情况下，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。

真值。指定设置标志时节点所用的真值。默认情况下，此值为 **T**。

假值。指定设置标志时节点所用的假值。默认情况下，此值为 **F**。

汇总关键字。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 23. 重新结构化之前的数据示例.

CustID	Account	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 24. 重新结构化之后的数据示例.

CustID	Account	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 25. 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）。请参阅主题第 92 页的『类型节点』，了解更多信息。

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）。

创建重构的字段。此处列出新创建的重构字段。默认情况下，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*。

包括字段名称。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。

使用来自其他字段的值。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

默认情况下，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。

为转置节点设置选项

新字段名

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。**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。注意：此选项仅可用于数字字段。

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）。

时间区间节点

通过“时间区间”节点，您可以为“时间序列”建模节点、“流式 TS”节点或“时间图”节点中用于估计或预测的时间序列数据指定间隔并生成标签。支持全部范围的时间间隔，从秒到年。例如，如果某个序列的日测量开始于 2005 年 1 月 3 日，那么可以为从该日期开始的记录添加标签，第二行为 1 月 4 日，依此类推。您也可以指定周期性 - 例如，每周五天或每天八小时。

此外，还可以指定要用于估计的记录的范围。可以选择是否要排除序列中最早的记录，以及是否要指定 hold-out。这样，您可以保留时间序列数据中最近的记录，以便对其已知值与这些周期的估计值进行比较，从而测试模型。

您还可以指定要对未来多少个时间周期进行预测，并且可以指定下游“时间序列”建模节点或“流式 TS”节点进行预测时使用的未来值。

时间区间节点会生成格式适用于指定的间隔和周期的 *TimeLabel* 字段，以及用于为每条记录分配唯一整数的 *TimeIndex* 字段。此外，还可以生成若干其他字段，具体取决于选择的间隔或周期（如测量所在的分钟或秒钟）。

您可根据需要填充或汇总值，以确保测量等度间隔。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。如果数据未达到这种要求，节点可对其转换以符合要求。

注释(O)

- 周期间隔可能不符合实际时间。例如，以标准的一周五个工作日为基准的序列会将周五与周一之间的间隔视为一天。
- “时间区间”节点假定每个序列各占一个字段或列，每个测量各占一行。如有必要，可以将数据转置以满足此要求。请参阅主题第 124 页的『转置节点』，了解更多信息。
- 对于不是等度间隔的序列，可以指定一个字段，用于标识每个测量的日期或时间。请注意，这需要具有适当格式的日期、时间或时间戳记字段，以用作输入。如有必要，可以使用“填充”节点将现有字段（如字符串标签字段）转换为这种格式。请参阅主题第 110 页的『使用填充节点进行存储类型转换』，了解更多信息。
- 查看生成的标签和索引字段的详细信息时，打开值标签的显示通常很有帮助。例如，查看包含用于月数据的生成值的表时，可以单击工具栏中的值标签图标，以查看 *January, February, March*, 等，而不是 1、2、3 等。



图 6. 值标签图标

指定时间间隔

使用“间隔”选项卡，可以指定用于构建或标注序列的间隔和周期性。特定设置取决于所选间隔。例如，如果选择 **小时（每天）**，则可以指定每周的天数、每周的起始日期、每天的小时数，以及每天的起始时间（小时）。请参阅主题第 128 页的『支持的间隔』，了解更多信息。

标注或构建序列

可以连续标注记录，也可以根据指定的日期、时间戳或时间字段构建序列。

- **从第一个记录开始标注。**指定标注连续记录的起始日期和/或时间。例如，如果标注每天的各个小时，那么应指定序列开始的日期和时间（小时），之后的每个小时使用一条记录。除添加标签以外，此方法不会更改原始数据。而是，它会假定记录已等度间隔，每个测量之间的间隔一致。在数据中，必须使用空行表示所有缺失的测量。
- **根据数据构建。**对于不是等度间隔的序列，可以指定一个字段，用于标识每个测量的日期或时间。请注意，这需要具有适当格式的日期、时间或时间戳记字段，以用作输入。例如，如果某个字符串字段具有 *Jan 2000*、*Feb 2000* 等类似值，则可以使用填充节点将其转换为日期字段。请参阅主题第 110 页的『使用填充节点进行存储类型转换』，了解更多信息。**根据数据构建**选项也会根据需要通过填充或汇总记录对数据进行转换以满足指定的间隔，例如，通过将周“累计”为月，或通过将缺失记录替换为空白值或外推值。可以在“构建”选项卡中指定用于填充或汇总记录的函数。请参阅主题『时间间隔构建选项』，了解更多信息。

新的字段名称扩展。可用于指定应用于节点所生成的所有字段的前缀或后缀。例如，使用缺省前缀 *\$TI_* 创建的字段的名称将为 *\$TI_TimeIndex* 和 *\$TI_TimeLabel*，依此类推。

日期格式。指定节点创建的 *TimeLabel* 字段的格式，该格式适用于当前间隔。此选项的可用性取决于当前选择内容。

时间格式。指定节点创建的 *TimeLabel* 字段的格式，该格式适用于当前间隔。此选项的可用性取决于当前选择内容。

时间间隔构建选项

使用时间区间节点中的“构建”选项卡，可以指定用于汇总或填充字段以满足指定间隔的选项。仅当在“间隔”选项卡中选择了**根据数据构建**选项时，这些设置才适用。例如，如果有以周和月为单位的混合数据，则可以对周值进行汇总或“累计”，以获得均匀的月间隔。或者，也可以将间隔设置为周，并通过为所有缺失周插入空白值或使用指定的填充函数外推缺失值来填充序列。

在填充或汇总数据时，现有的所有日期或时间戳记字段实际上都会被生成的 *TimeLabel* 和 *TimeIndex* 字段代替，并从输出中删除。无类型字段也会被删除。对于以持续时间的形式测量时间的字段（如测量服务电话时长而不是电话起始时间的字段），只要它们在内部存储为时间字段而不是时间戳记，便会被保留。请参阅主题第 19 页的『设置字段存储类型和格式』，了解更多信息。其他字段以“构建”选项卡中指定的选项为基础进行汇总。

- **使用缺省字段和函数。**根据需要指定应对所有字段进行汇总或填充，如上所述，日期、时间戳和无类型字段除外。缺省函数将根据测量级别进行应用，例如，连续字段使用均值函数进行汇总，名义字段则使用众数函数。您可以在对话框的下半部分更改一个或多个测量级别的缺省设置。
- **指定字段和函数。**可用于指定要填充或汇总的字段，以及分别使用的函数。所有未选中的字段都将从输出中删除。使用右侧的图标在表中添加或删除字段，或单击相应列中的单元格以更改用于该字段的汇总或填充函数，以覆盖缺省设置。无类型字段将从列表中排除，且不能添加到表中。

缺省值。指定缺省情况下用于不同类型字段的汇总和填充函数。选择 **使用默认设置** 时，将应用这些默认函数，它们还将应用为添加到表中的所有新字段的初始默认设置。（更改默认设置不会更改表中的任何现有设置，但会应用于之后添加的所有字段。）

汇总函数。可用的汇总函数如下：

- **连续**。用于连续字段的可用函数包括**均值、合计、众数、最小值和最大值**。
- **名义**。选项包括**众数、第一个和最后一个**。“第一个”表示汇总组中的第一个非空值（按日期排序）；“最后一个”表示组中的最后一个非空值。
- **标志**。选项包括**如果任何值都为真，则为真、众数、第一个和最后一个**。

填充函数。可用的填充函数如下：

- **连续**。选项包括**空白值和“最近点的平均值”**，后者表示将要创建的时间周期之前的三个最近非空值的均值。如果没有三个值，那么新值将是空白值。最近值仅包括实际值；在非空值的搜索中不会考虑先前创建的填充值。
- **名义**。**空白值和最近值**。“最近”是指将要创建的时间周期之前的最近非空值。同样，最近值的搜索中仅考虑实际值。
- **标志**。选项包括**空白值、真和假**。

生成的数据集中的最大记录数。指定所创建记录数的上限，如果不指定，该数字可能相当大，在时间间隔设置为秒（无论是否有意如此设置）时尤其如此。例如，如果填充至秒钟，那么只有两个值（2000 年 1 月 1 日和 2001 年 1 月 1 日）的序列将生成 31,536,000 条记录（60 秒 x 60 分钟 x 24 小时 x 365 天）。如果超出指定的最大值，系统将停止处理并显示警告消息。

计数字段

汇总或填充值时，会创建一个新的 **计数** 字段，用于指示确定新记录的过程中涉及的记录数。例如，如果四个周值汇总至一个月中，计数将是 4。对于填充记录，计数为 0。字段的名称为 *Count* 加上在“间隔”选项卡中指定的前缀或后缀。

估计期

使用时间区间节点的“估计”选项卡，可以指定模型估计中使用的记录范围，以及任何 holdout。可以根据需要在下游建模过程中覆盖这些设置，但在此进行指定比分别针对每个节点进行指定更为方便。

开始估计。可以在数据起点开始估计期，或排除对于预测而言用处不大的旧值。根据具体数据，您会发现缩短估计期可以提高性能（并减少用于数据准备的时间），而不会明显降低预测精度。

结束估计。可以使用截至数据末尾的所有记录估计模型，也可以“保留”最近的记录以评估模型。在后一种情况下，您有效地“预测”已知值，可以比较观测值和预测值以测量模型的有效性。

预测值(S)

通过“时间区间”节点的“预测”选项卡，您可以指定要预测的记录数，并指定下游“时间序列”建模节点或“流式 TS”节点进行预测时使用的未来值。可以根据需要在下游节点中覆盖这些设置，但在此处指定这些设置比分别针对每个节点指定设置更为方便。

将记录扩展到未来。指定估计期以外的预测记录数。请注意，这些记录可以是“预测”，也可以不是，具体取决于“估计”选项卡中指定的 holdout 数。

未来指示字段。已生成字段的标签，用于指示记录是否包含预测数据。标签的默认值为 *\$TI_Future* 。

要在预测中使用的未来值。对于要预测的每条记录（holdout 值除外），如果使用预测变量字段（角色设置为输入），则必须为每个预测变量的预测周期指定估计值。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”选项卡中选项的约束。例如，如果在“构建”选项卡中选择了**指定字段和函数**，则会将未汇总或填充的所有字段从流中删除，且不能将其用于建模。

注意：如果为流中不再可用（因为已将其删除或由于“构建”选项卡中的选项更新）的字段指定了未来值，那么此字段在“预测”选项卡中将显示为红色。

- **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 26. 可用于测量级别的函数。

测量级别	函数(F)
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。

未来值

可以在此处指定下游“时间序列”建模节点或“流式 TS”节点进行预测时使用的未来值。可以根据需要在下游节点中覆盖这些设置，但在此处指定这些设置比分别针对每个节点指定设置更为方便。

可以手动输入值，也可以单击对话框右边的选择器按钮，从为当前字段定义的值列表中选择值。

可以指定的未来值数量与扩展时间序列的记录数对应。

支持的间隔

“时间区间”节点支持全部范围的间隔（从秒到年）以及循环（例如，季节性循环）和非循环周期。在“间隔”选项卡的“时间间隔”字段中指定间隔。

周期

选择 **周期** 可以为不符合任何其他指定间隔的现有非循环序列添加标签。该序列必须已具有正确顺序，每个测量之间的间隔一致。选择这种间隔时，**根据数据构建** 选项将不可用。

样本输出

基于指定的起始值（*Period 1*、*Period 2*，依此类推）为记录添加递增标签。创建了如下表中所示的新字段。

表 27. 新字段创建示例

\$TI_TimeIndex (整数)	\$TI_TimeLabel (字符串)	\$TI_Period (整数)
1	Period 1	1
2	Period 2	2
3	Period 3	3
4	Period 4	4
5	Period 5	5

循环周期

选择 **循环周期** 可以为具有不符合某种标准间隔的重复循环的现有序列添加标签。例如，如果您的财政年度中只有 10 个月，那么可以使用此选项。该序列必须已具有正确顺序，每个测量之间的间隔一致。（选择这种间隔时，**根据数据构建** 选项将不可用。）

样本输出

基于指定的起始循环和周期（*Cycle 1*、*Period 1*、*Cycle 1*、*Period 2*，依此类推）为记录添加递增标签。例如，如果每个循环的周期数设为 3，那么将创建如下表中所示的新字段。

表 28. 新字段创建示例

\$TI_TimeIndex (整数)	\$TI_TimeLabel (字符串)	\$TI_Cycle (整数)	\$TI_Period (整数)
1	Cycle 1, Period 1	1	1
2	Cycle 1, Period 2	1	2
3	Cycle 1, Period 3	1	3
4	Cycle 2, Period 1	2	1
5	Cycle 2, Period 2	2	2

年

对于年，可以指定标注连续记录的起始年，也可以选择 **根据数据构建**，以指定用于标识每条记录的年度的时间戳或日期字段。

样本输出

创建了如下表中所示的新字段。

表 29. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (字符串)	\$TI-Year (整数)
1	2000	2000
2	2001	2001
3	2002	2002
4	2003	2003
5	2004	2004

季

对于季度序列，可以指定财政年度的起始月份。此外，可以指定标注连续记录的起始季度和年（例如，Q1 2000），也可以选择 **根据数据构建**，以选择用于标识每条记录的季度和年度的时间戳或日期字段。

样本输出

对于从一月开始的财政年度，新字段的创建和填充如下表所示：

表 30. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (字符串)	\$TI-Year (整数)	\$TI-Quarter (带标签的整数)
1	Q1 2000	2000	1 (Q1)
2	Q2 2000	2000	2 (Q2)
3	Q3 2000	2000	3 (Q3)
4	Q4 2000	2000	4 (Q4)
5	Q1 2001	2001	1 (Q1)

如果年度开始于一月以外的月份，那么新字段将如下表中所示（假定财政年度从七月开始）。要查看用于标识每个季度月份的标签，请通过单击工具栏图标打开值标签的显示。



图 7. 值标签图标

表 31. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (字符串)	\$TI-Year (整数)	\$TI-Quarter (带标签的整数)
1	Q1 2000/2001	1	1 (Q1 7-9 月)
2	Q2 2000/2001	1	2 (Q2 10-12 月)
3	Q3 2000/2001	1	3 (Q3 1-3 月)
4	Q4 2000/2001	1	4 (Q4 4-6 月)
5	Q1 2001/2002	2	1 (Q1 7-9 月)

月

可以选择标注连续记录的起始年和月，也可以选择 **根据数据构建**，以选择用于标识每条记录月份的时间戳或日期字段。

样本输出

创建了如下表中所示的新字段。

表 32. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (日期)	\$TI-Year (整数)	\$TI-Months (带标签的整数)
1	2000 年 1 月	2000	1 (一月)
2	2000 年 2 月	2000	2 (二月)
3	2000 年 3 月	2000	3 (三月)
4	2000 年 4 月	2000	4 (四月)
5	2000 年 5 月	2000	5 (五月)

周（非周期性）

对于以周为单位的序列，可以选择一周内循环开始的日期。

请注意，周只能是非周期性的，因为不同的月、季度甚至年不一定具有相同的周数。但对于非周期性模型，可以将时间戳数据轻松汇总或填充至周级别。

样本输出

创建了如下表中所示的新字段。

表 33. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (日期)	\$TI-Week (整数)
1	1999-12-27	1
2	2000-01-03	2
3	2000-01-10	3
4	2000-01-17	4
5	2000-01-24	5

一周的 *\$TI-TimeLabel* 字段将显示该周的第一天。在前面的表中，用户从 2000 年 1 月 1 日开始标注。但一周从周一开始，而 2000 年 1 月 1 日是周六。因此，包含 1 月 1 日的一周将从 1999 年 12 月 27 日开始，并且作为第一个点的标签。

日期格式将决定为 *\$TI-TimeLabel* 字段生成的字符串。

天（每周）

对于周循环内的日测量，可以指定每周的天数和每周的起始日期。可以指定标注连续记录的起始日期，也可以选择 **根据数据构建**，以选择用于标识每条记录日期的时间戳或日期字段。

样本输出

创建了如下表中所示的新字段。

表 34. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (日期)	\$TI-Week (整数)	\$TI-Day (带标签的整数)
1	2005 年 1 月 5 日	1	3 (周三)
2	2005 年 1 月 6 日	1	4 (周四)
3	2005 年 1 月 7 日	1	5 (周五)
4	2005 年 1 月 10 日	2	1 (周一)
5	2005 年 1 月 11 日	2	2 (周二)

注意：对于第一个时间周期，一周总是从 1 开始，而且不会基于日历进行循环。因此，第 52 周后面是第 53、54 周，依此类推。一周并不反映一年中的该周，只是序列中每周增加的数字而已。

天（非周期性）

如果有不符合常规周循环的日测量，那么可选择以天为单位的非周期性序列。可以指定标注连续记录的起始日期，也可以选择 **根据数据构建**，以选择用于标识每条记录日期的时间戳或日期字段。

样本输出

创建了如下表中所示的新字段。

表 35. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (日期)
1	2005 年 1 月 5 日
2	2005 年 1 月 6 日
3	2005 年 1 月 7 日
4	2005 年 1 月 8 日
5	2005 年 1 月 9 日

小时 (每天)

对于日循环内的时测量，可以指定每周的天数、一天中的小时数（如八小时工作日）、一周的起始日期，以及每天的起始时间（小时）。可以基于 24 小时时钟（例如，14:05 = 2:05 p.m.）指定小时，最小单位为分钟。

可以指定标注连续记录的起始日期和时间，也可以选择 **根据数据构建**，以选择用于标识每条记录的日期和时间的戳字段。

样本输出

创建了如下表中所示的新字段。

表 36. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (时间戳)	\$TI-Day (带标签的整数)	\$TI-Hour (带标签的整数)
1	2005 年 1 月 5 日 8:00	3 (周三)	8 (8:00)
2	2005 年 1 月 5 日 9:00	3 (周三)	9 (9:00)
3	2005 年 1 月 5 日 10:00	3 (周三)	10 (10:00)
4	2005 年 1 月 5 日 11:00	3 (周三)	11 (11:00)
5	2005 年 1 月 5 日 12:00	3 (周三)	12 (12:00)

小时 (非周期性)

如果有不符合常规日循环的时测量，请选择此选项。可以指定标注连续记录的起始时间，也可以选择 **根据数据构建**，以选择用于标识每条记录时间的时间戳或时间字段。

小时以 24 小时时钟为基准 (13:00 = 1:00 p.m.)，且不会重叠循环（第 25 小时接着第 24 小时）。

样本输出

创建了如下表中所示的新字段。

表 37. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (字符串)	\$TI-Hour (带标签的整数)
1	8:00	8 (8:00)
2	9:00	9 (9:00)
3	10:00	10 (10:00)
4	11:00	11 (11:00)
5	12:00	12 (12:00)

分钟（每天）

对于日循环内以分钟为单位的测量，可以指定每周的天数、一周的起始日期、一天中的小时数，以及一天的起始时间。基于 24 小时时钟指定小时，可以使用冒号进一步指定至分钟和秒钟（例如，2:05:17 p.m. = 14:05:17）。此外，也可以指定递增的分钟数（每分钟、每两分钟，依此类推，这里的增量必须是可整除 60 的值）。

可以指定标注连续记录的起始日期和时间，也可以选择 **根据数据构建**，以选择用于标识每条记录的日期和时间的时间戳字段。

样本输出

创建了如下表中所示的新字段。

表 38. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (时间戳)	\$TI-Minute
1	2005-01-05 08:00:00	0
2	2005-01-05 08:01:00	1
3	2005-01-05 08:02:00	2
4	2005-01-05 08:03:00	3
5	2005-01-05 08:04:00	4

分钟（非周期性）

如果有不符合常规日循环的以分为单位的测量，请选择此选项。此外，也可以指定递增的分钟数（每分钟、每两分钟，依此类推，这里的指定值必须是可整除 60 的数值）。

可以指定标注连续记录的起始时间，也可以选择 **根据数据构建**，以选择用于标识每条记录时间的时间戳或时间字段。

样本输出

创建了如下表中所示的新字段。

表 39. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (字符串)	\$TI-Minute
1	8:00	0
2	8:01	1
3	8:02	2
4	8:03	3
5	8:04	4

- 创建 *TimeLabel* 字符串时，会在小时与分钟之间使用冒号。小时不会重叠循环 - 第 25 小时接着第 24 小时。
- 分钟按对话框中指定的值递增。例如，如果增量为 2，*TimeLabel* 将是 8:00、8:02，依此类推；分钟将是 0、2，依此类推。

秒（每天）

对于日循环内以秒钟为单位的间隔，可以指定每周的天数、一周的起始日期、一天中的小时数，以及一天的起始时间。基于 24 小时时钟指定小时，可以使用冒号进一步指定至分钟和秒钟（例如，2:05:17 p.m. = 14:05:17）。此外，也可以指定递增的秒数（每秒钟、每两秒钟，依此类推，这里的指定值必须是可整除 60 的数值）。

可以指定标注连续记录的起始日期和时间，也可以选择 **根据数据构建**，以选择用于指定每条记录的日期和时间的时间戳字段。

样本输出

创建了如下表中所示的新字段。

表 40. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (时间戳)	\$TI-Minute	\$TI-Second
1	2005-01-05 08:00:00	0	0
2	2005-01-05 08:00:01	0	1
3	2005-01-05 08:00:02	0	2
4	2005-01-05 08:00:03	0	3
5	2005-01-05 08:00:04	0	4

秒（非周期性）

如果有不符合常规日循环的以秒钟为单位的测量，请选择此选项。此外，也可以指定递增的秒数（每秒钟、每两秒钟，依此类推，这里的指定值必须是可整除 60 的数值）。

指定标注连续记录的起始时间，或选择 **根据数据构建**，以选择用于标识每条记录时间的时间戳或时间字段。

样本输出

创建了如下表中所示的新字段。

表 41. 新字段创建示例

\$TI-TimeIndex (整数)	\$TI-TimeLabel (字符串)	\$TI-Minute	\$TI-Second
1	8:00:00	0	0
2	8:00:01	0	1
3	8:00:02	0	2
4	8:00:03	0	3
5	8:00:04	0	4

- 创建 *TimeLabel* 字符串时，会在小时与分钟、分钟与秒钟之间使用冒号。小时不会重叠循环 - 第 24 小时之后是第 25 小时。
- 秒钟按指定为增量的数值递增。如果增量为 2，*TimeLabel* 将是 8:00:00、8:00:02，依此类推；秒钟将是 0、2，依此类推。

历史记录节点

“历史记录”节点最常用于顺序数据，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字段的新字段。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。这表示，当节点处理记录 10 时，将从记录 7 至记录 3 添加字段。使用箭头可调整范围值。

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。历史记录字段将填入未定义值，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。默认的替换值为系统空值 `undef`。使用字符串 `$null$` 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多个位置的显示方式，如表格、列表和字段选择器。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. 选择表中的某个字段。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。

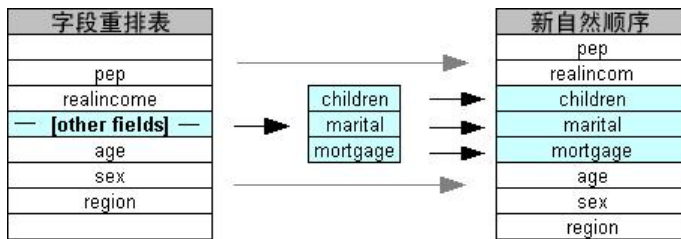


图 8. 说明“其他字段”在新字段顺序中的结合方式的图表

- 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。

排序方式。 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
- 类型
- 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。

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浏览导入到 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”节点可在一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值字段间的关系。可通过使用点（散点）或线创建散点图。



“分布”节点显示了标志（类别）值（例如抵押类型或性别）的出现次数。通常，可以使用“分布”节点来显示数据中的不平衡度，然后在创建模型前使用“平衡”节点来纠正此类不平衡度。



“直方图”节点显示了数值字段的值的出现次数。它经常用来在数据操作和模型构建之前探索数据。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度。



“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。



使用多重散点图节点可创建在一个 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。要研究几个变量随时间的变化情况时，多重散点图非常有用。



Web 节点说明了两个或多个符号（分类）字段值之间关系的强度。该图使用不同粗细的线来表示关系强度。例如，您可以使用 Web 节点来研究电子商务网站上一系列商品的购买之间的关系。



“时间散点图”节点显示一个或多个时间序列数据集。通常情况下，您首先要使用时间区间节点创建一个 *TimeLabel* 字段，该字段用于为 *x* 轴设置标签。



“评估”节点有助于评估和比较预测模型。评估图表显示了模型对特定结果的预测优劣。它根据预测值和预测置信度来对记录进行排序。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。

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。以下小节含有这些通用选项的详细信息。

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和数据区域生成“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、交叠、面板和动画

交叠和审美原则

外观（和重叠）向直观表示添加维数。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量。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。

应该尝试使用审美原则和交叠，以找出完全满足需求的效果。以下说明可能会有助于您选出合适的效果。

注意：并非所有外观或重叠都可以用于所有可视化类型。

- **颜色。** 如果颜色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，每个类别一种颜色。当颜色为连续数字范围时，它会根据范围字段的值发生变化。如果图形元素（例如，条或框）代表不止一个记录/案例并且范围字段用于颜色，则颜色会根据范围字段的平均数发生变化。
- **形状。** 形状由一个分类字段定义，该分类字段将直观表示分割为不同形状的元素，每个类别一种元素。
- **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，条或框）代表不止一个记录/案例并且范围字段用于透明度，则颜色会根据范围字段的平均数发生变化。处于最大值时，图形元素为完全透明。处于最大值时，图形元素为完全不透明。
- **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。** 如果大小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，每个类别一种尺寸。当大小为连续数字范围时，它会根据范围字段的值发生变化。如果图形元素（例如，条或框）代表不止一个记录/案例并且范围字段用于大小，则大小会根据范围字段的平均数发生变化。

面板和动画

镶面。 通过镶面（也称为分面）可创建图形表。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。镶面在检查直观表示是否符合镶面字段的条件时将非常有用。例如，您可以按性别嵌入直方图，确定男性和女性的频率分布是否相等。即，您可以检查薪水是否符合性别差异。选择一个用于镶面的分类字段。

动画。动画表示镶板，镶板是用动画字段值创建多个图形，但这些图形并不同时显示。更确切地说，使用浏览模式中的控件可以将输出生成动画，也可以在一系列的单个图片间跳转。而且，与镶面不同，动画不需要分类字段。可以指定连续字段，其值将自动分割为范围。可以在浏览模式中使用动画控件改变范围的大小。不是所有直观表示都具有动画。

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。

注意：分布节点图具有其他设置。

输出名称。指定运行节点时使用的图形名称。**自动** 根据生成输出的节点选择名称。（可选）可以选择**自定义**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生成输出。仅在“分布”节点中可用。
- **输出表。**选择此选项将以表格格式生成输出。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。使用省略号按钮 (...) 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*) （可用于其他 IBM SPSS Statistics 应用程序）。

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*)

编页码输出。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。）

每页的行数。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。

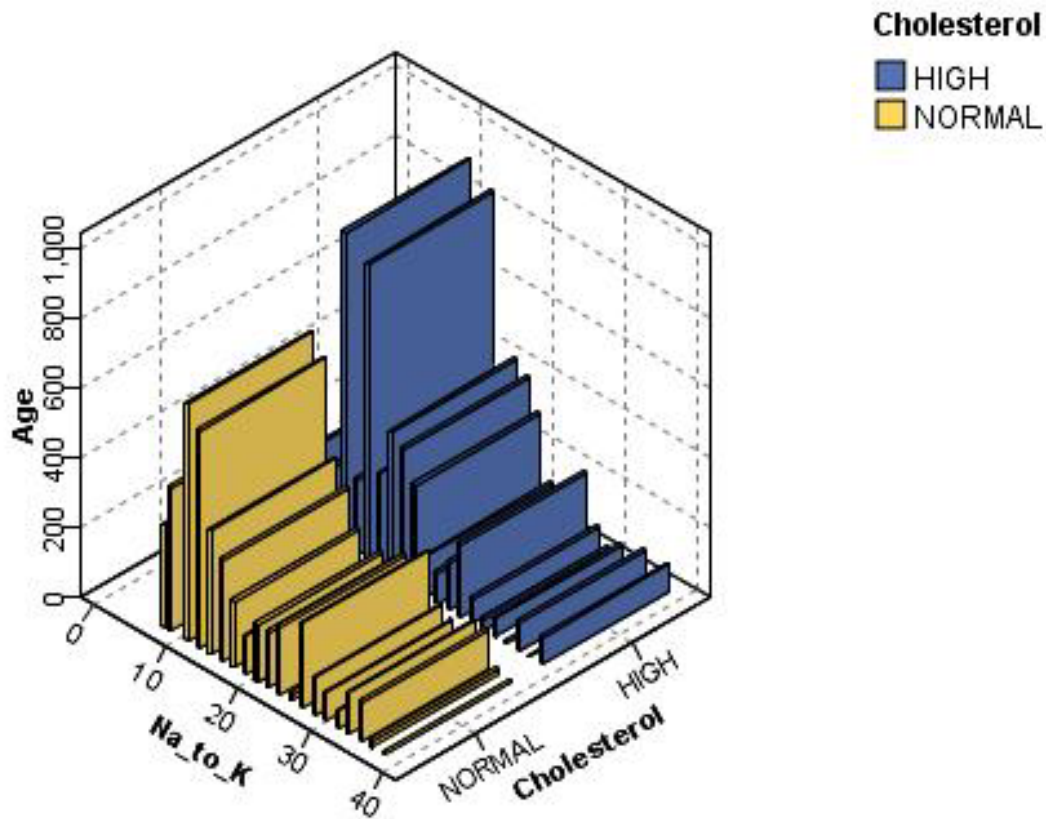


图9. 带有 x、y 和 z 轴的集合图

在 IBM SPSS Modeler 中创建 3D 图形有两种方法：在第三坐标轴上绘制信息的散点图（真实 3D 图形）和显示 3D 效果的图形。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 3D 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. 一旦创建图形后，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，您可以从单个节点上的许多不同图形输出（条形图、饼图、直方图、散点图和热图等）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。

注意：为了编辑节点或选择图形类型，必须将“图形板”节点连接到具有数据的流上。

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、样式表和映射。可以在本地计算机上导入、导出、重命名和删除可视化模板、样式表和映射。请参阅主题第 163 页的『管理模板、样式表和地图文件』，了解更多信息。

位置。更改可视化模板、样式表和映射的存储位置。当前位置列出在按钮右侧。请参阅主题第 162 页的『设置模板、样式表和地图的位置』，了解更多信息。

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。选择数据后，即可使用合适的直观表示类型子集显示您的数据。有关示例，请参阅第 151 页的『图形板示例』。

1. 从列表中选择一或多个字段（变量）。使用 **Ctrl+单击** 可选择多个字段。

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详细信息，请参阅第 142 页的『字段（变量）类型』。

2. 选择一个直观表示类型。有关可用类型的描述，请参阅第 145 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。哪些统计量子集可用，取决于该统计量是基于计数还是根据连续字段计算。可用统计还取决于模板自身。下一步就是可用统计的完整列表。
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 144 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 (Mean)**. 集中趋势的测量值。算术平均值，等于总和除以观测值数。
- **中位数 (Median)**. 第 50 个百分位，大于该值和小于该值的个案数各占一半。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。
- **最小值 (Minimum)**. 数字变量的最小值。
- **最大值 (Maximum)**. 数字变量的最大值。
- **范围**. 最大值与最小值之间的差。
- **中程数值**. 范围中间值，即与最小值的差等于与最大值的差的值。
- **Sum**. 所有具有非缺失值的观测值的和或总计。
- **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 (Variance)**。平均值离散度的测量值，等于均值的平方差除以观测值数减一的差。方差按单元计量，即变量自身单元数的平方。
- **标准差 (Standard Deviation)**。均数离差的度量值。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)**。因样本变化而变化的检验统计量值的个数的度量值。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)**。观测值聚类围绕中心点的程度的一种测度。对于正态分布，峰度统计量的值为零。正峰度值表示相对于正态分布，观察值在分布中心的聚集更多，同时尾部更薄，直到分布极值。在这一点，尖峰态分布的尾部比正态分布的尾部要厚。负峰度值表示相对于正态分布，观察值聚集得少并且尾部较厚，直到分布极值。在这一点，低峰态分布的尾部比正态分布的尾部要薄。
- **偏度 (Skewness)**。分布的不对称度量值。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，区域统计会产生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% 机会包含给定单个观测值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些列值

基于计数的汇总统计

- **计数**。行/观测值数量。
- **累积计数**。行/观测值的累积数目。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/观测值数量对比行/观测值的总数的百分比。
- **计数累积百分比**。每个子组中行/观测值数目对比行/观测值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，指示字段类型和数据类型。图标还标识多重响应集。

表 42. 测量级别图标。










测量级别	数值(N)	字符串	日期	时间
连续		n/a		
排序集合				
设置				

表 43. 多重响应集图标.

多重响应集类型	图标
多重响应集, 多分类	
多重响应集, 多二分法	

测量级别

创建直观表示时, 字段的测量级别非常重要。以下是对测量级别的描述。您可以在字段列表中右键单击字段并选择一个选项, 以暂时更改测量级别。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。带有限制数量的区分值或分类的数据 (例如, 性别或宗教信仰)。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)。分类字段也成为定性数据。集合、有序集合和标志都是分类字段。

- **集合。**此字段/变量的值采用某些固有的排秩方式 (例如, 员工工作的公司部门) 表示类别。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。**此字段/变量的值采用某些固有的排秩方式 (例如, 从极差到极高的服务满意度) 表示类别。有序集合的示例包括表示满意度或置信度的态度分数和优先选择评分。也称为有序变量。
- **标志。**此字段/变量具有两个不同值, 例如 Yes 与 No 或 1 与 2。也称为二分或二元变量。

连续。数据是按间隔或比例尺度进行测量的, 其中数据值指示各值的顺序和值与值间的距离。例如, \$72,195 的工资高于 \$52,398 的工资, 两个值之间的距离是 \$19,797。这也称为定量、尺度或数值范围数据。

分类字段定义直观表示中的类别, 通常画出单独的图形元素或将图形元素分组。通常在分类字段的类别内汇总连续字段。例如, 性别类别的收入默认直观表示会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, 散点图显示每个个案的当前工资和起始工资。可以使用分类字段按性别对个案进行分组。

数据类型

测量级别不是确定字段类型的唯一属性。字段也可保存为某个特定数据类型。可能的数据类型包括字符串 (非数字数据, 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。您必须更改数据在原始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”。多重响应集在通常意义上不是真正的“字段”。多重响应集使用多个字段记录对问题的响应, 响应者可以针对这些问题给出多个答案。可将多重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, 也可以对多重响应集执行。

多重响应集可为多二分法集或多类别集。

多二分集。多二分集通常由多个二分字段组成: 只具有是/否、存在/不存在、选中/不选中属性的两个可能值的变量。虽然字段可能不是严格二分, 但集合中的所有字段都采用同一方式编码。

例如，一项调查提供问题的五个可能回答，“以下哪些信息您需要通过新闻来获取？”响应者可选中每个选项旁边的复选框来表示多个选项。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，这些字段都采用同一方式编码，常常具有多个可能的响应类别。例如，一个调查项目，内容为“列出最能表现您的民族传统的国家”。可能会有数百个可能的回答，但为了进行编码，列表将限于 40 个最常见的国家，其他国家归于“其他”类别。在数据文件中，三个选项变为三个字段，每个字段具有 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

如果您知道要创建的直观表示类型，或者您想向直观表示添加可选外观、面板和/或动画，请使用“详细”选项卡。有关示例，请参阅第 151 页的『图形板示例』。

1. 如果已在“基本”选项卡中选择直观表示类型，那么将显示该类型。否则，请从下拉列表选择一个类型。有关可视化类型的信息，请参阅第 145 页的『可用的内置图形板可视化类型』。
2. 直观表示缩略图正右侧为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（例如在条形图中），可以使用这些汇总选项之一作为透明外观。有关摘要统计的描述，请参阅第 141 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 138 页的『审美原则、交叠、面板和动画』。请注意，脚本中不支持透明外观。
5. 如果您要创建地图直观表示，**地图文件组**将显示要使用的一个或多个地图文件。如果有默认的地图文件，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。请参阅主题『选择用于地图直观表示的地图文件』，了解更多信息。
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详细信息，请参阅第 138 页的『审美原则、交叠、面板和动画』。

选择用于地图直观表示的地图文件

如果选择了地图直观表示模板，那么需要地图文件来定义绘制地图的地理信息。如果有默认的地图文件，则会将其用于地图直观表示。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。如果您具有其他要使用的 ESRI shapefile，那么首先需要将 shapefile 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。在转换地图之后，请单击“模板选择器”对话框上的**管理...**以将地图导入至管理系统，以便它在“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。
- 如果模板不需要链接到数据关键字的地图关键字，那么它需要参考地图文件和字段，后者指定在参考地图上绘制元素的坐标（例如，经度和纬度）。
- 交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 164 页的『地图的关键概念』。

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您已导入的地图文件。有关管理地图文件的更多信息，请参阅第 163 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默认值，则选择此复选框。在指定了默认地图文件之后，您无需在每次创建地图直观表示时指定地图文件。

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。

在直观表示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直观表示中呈现，即使不存在匹配的数据关键字值时也是如此。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直观表示。地图关键字和数据关键字应该从同一域获取，例如，国家或地区和区域。单击**比较**以测试数据关键字和地图关键字值是否匹配。显示的图标将指示比较的状态。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关键字值。如果取消选中了**在直观表示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 44. 比较图标.

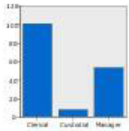
图标	描述
	未执行比较。这是您在单击 比较 前的默认状态。由于您不知道数据关键字和地图关键字值是否匹配，因此您需要继续小心操作。
	已执行比较，数据关键字和地图关键字值完全匹配。对于每个数据关键字值，都具有由地图关键字标识的匹配特征。
	已执行比较，某些数据关键字和地图关键字值不匹配。对于某些数据关键字值，不存在由地图关键字标识的匹配特征。您需要继续小心操作。如果继续，地图直观表示将不会包含所有数据值。
	已执行比较，数据关键字和地图关键字值完全不匹配。如果继续，将不会呈现任何地图，因此您应当选择其他数据关键字或地图关键字。

可用的内置图形板可视化类型

您可以创建多个不同的可视化类型。可以同时“基本”和“详细”选项卡上找到下列内置类型。模板（特别是地图模板）的某些描述使用**特殊文本**标识“详细”信息卡上指定的字段（变量）。

表 45. 可用的图形类型.

图表图标



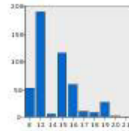
描述

条(B)

计算连续数字字段的汇总统计, 并将分类字段的各个类别的结果显示为条形。

所需项: 一个分类字段和一个连续字段。

图表图标

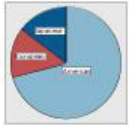


描述

计数条

将分类字段的各个类别中的行/观测值比例显示为条形。也可以使用分布图节点生成此图形。该节点提供了一些其他选项。请参阅主题第 175 页的『分布节点』以获取更多信息。

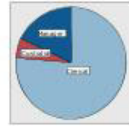
所需项: 单一分类字段。



饼(E)

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。

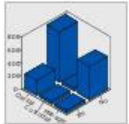
所需项: 一个分类字段和一个连续字段。



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。

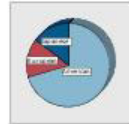
所需项: 单一分类字段。



3-D 条形

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。

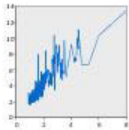
所需项: 一对分类字段和一个连续字段。



3-D 饼图

该图与饼图相似, 但具有 3-D 效果。

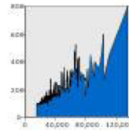
所需项: 一个分类字段和一个连续字段。



线(L)

根据各个其他字段值计算字段的汇总统计, 并绘制一条折线将这些值连接起来。也可以使用散点图节点生成折线散点图。该节点提供了一些其他选项。请参阅主题第 169 页的『散点图节点』以获取更多信息。

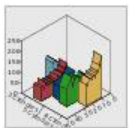
所需项: 一对任意类型的字段。



区域 (Area)

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。折线图和分区图之间的区别很小, 即分区图中包含了折线及折线下方着色的空间。但是, 如果您使用彩色审美原则, 将出现折线的简单分割和分区的堆叠。

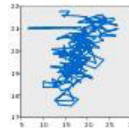
所需项: 一对任意类型的字段。



3-D 区域

显示根据其他字段值绘制成的一个字段值, 并按分类字段分割。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。



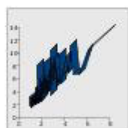
路径

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。

所需项: 一对任意类型的字段。

表 45. 可用的图形类型 (续).

图表图标

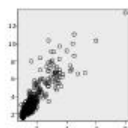


描述

带状(R)

根据各个其他字段值计算字段的汇总统计，并绘制一条带状将这些值连接起来。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。

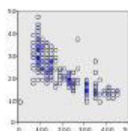
所需项: 一对任意类型的字段。



散点图

显示根据一个字段的值绘制的另一个字段的值。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。请参阅主题第 169 页的『散点图节点』以获取更多信息。

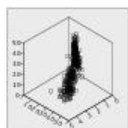
所需项: 一对任意类型的字段。



分级散点图

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，相似的值经过分级成为组，颜色或大小审美原则用于标明各个分级中的观测值数量。

所需项: 一对连续字段。

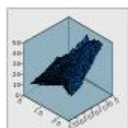


三维散点图

显示根据彼此绘制成的三个字段的值。此图形能够突出显示字段（如果有）之间的关系。也可以使用散点图节点生成 3-D 散点图。该节点提供了一些其他选项。请参阅主题第 169 页的『散点图节点』以获取更多信息。

所需项: 三个任意类型的字段。

图表图标

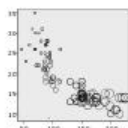


描述

曲面图

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。

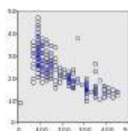
所需项: 三个任意类型的字段。



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各个点的大小。

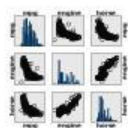
所需项: 三个任意类型的字段。



十六进制分级散点图

请参阅“分组散点图”的说明。区别在于基础分级的形状不同，即其形状近似于六角形，而不是近似于圆形。生成的十六进制分级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。

所需项: 一对连续字段。



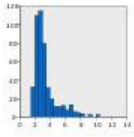
散点图矩阵图 (SPLOM)

显示根据各个字段的值绘制的另一个字段的值。SPLOM 类似于散点图的表。SPLOM 也包含了各个字段的直方图。

所需项: 两个或两个以上连续字段。

表 45. 可用的图形类型 (续).

图表图标



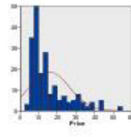
描述

直方图 (Histogram)

显示字段的频率分布情况。直方图有助于您确定分布类型和查看分布是否对称。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。请参阅主题第 179 页的『直方图选项卡』以获取更多信息。

所需项: 单一任意类型的字段。

图表图标

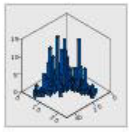


描述

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。

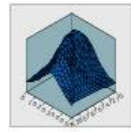
所需项: 单一连续字段。



3-D 直方图

显示连续字段对的频率分布情况。

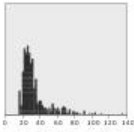
所需项: 一对连续字段。



3-D 密度

显示连续字段对的频率分布情况。这与 3-D 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。

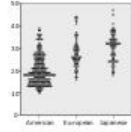
所需项: 一对连续字段。



点图

显示各个观测值/行，并将其堆叠在 x 轴的不同数据点上。该图与直方图的相似之处在于它显示了数据的分布情况，但它显示的是各观测值/行，而非特定分级（值范围）的汇总计数。

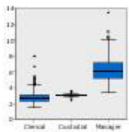
所需项: 单一任意类型的字段。



2-D 点阵图

显示各个观测值/行，并将其堆叠在分类字段中各个类别的 y 轴的不同数据点上。

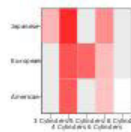
所需项: 一个分类字段和一个连续字段。



箱图

为分类字段的各个类别计算连续字段的五个统计量（最小值、第一四分位数、中位数、第三四分位数和最大值）。结果显示为长方形图/模式元素。长方形图有助于您查看类别中连续数据分布的变化情况。

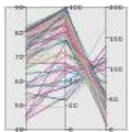
所需项: 一个分类字段和一个连续字段。



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。

所需项: 一对分类字段和一个连续字段。



平行图

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。

所需项: 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）各个类别的计数，并绘制使用色彩饱和度和表示地图特征中计数的地图，这些计数对应于类别。

所需项: 一个分类字段。关键字与数据关键字类别匹配的地图文件。

表 45. 可用的图形类型 (续).

图表图标	描述	图表图标	描述
	<p>平均值/中位数/总和的等值线图 计算分类字段 (数据关键字) 各个类别的连续字段 (颜色) 的平均值、中位数或总和, 并绘制使用色彩饱和度来表示地图特征中已计算统计量的图, 这些统计量对应于类别。</p> <p>所需项: 一个分类字段和一个连续字段。关键字与数据关键字类别匹配的地图文件。</p>		<p>值的等值线图 绘制使用颜色表示地图特征的分类字段 (颜色) 值的图, 这些值对应于另一分类字段 (数据关键字) 所定义的值。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。</p> <p>所需项: 一对分类字段。关键字与数据关键字类别匹配的地图文件。</p>
	<p>计数的等值线图上的坐标 这与计数的等值线图相类似, 区别在于前者具有两个额外的连续字段 (经度和纬度), 用于标识等值线图上绘制点的坐标。</p> <p>所需项: 一个分类字段和一对连续字段。关键字与数据关键字类别匹配的地图文件。</p>		<p>平均值/中位数/总和的等值线图上的坐标 这与平均值/中位数/总和的等值线图相类似, 区别在于前者具有两个额外的连续字段 (经度和纬度), 用于标识等值线图上绘制点的坐标。</p> <p>所需项: 一个分类字段和三个连续字段。关键字与数据关键字类别匹配的地图文件。</p>
	<p>值的等值线图上的坐标 这与值的等值线图相类似, 区别在于前者具有两个额外的连续字段 (经度和纬度), 用于标识等值线图上绘制点的坐标。</p> <p>所需项: 一对分类字段和一对连续字段。关键字与数据关键字类别匹配的地图文件。</p>		<p>地图上的计数图条 计算各个地图特征 (数据关键字) 的分类字段 (类别) 的各个类别中的行/观测值比例, 并绘制一个图, 条形图位于各个地图特征的中央。</p> <p>所需项: 一对分类字段。关键字与数据关键字类别匹配的地图文件。</p>
	<p>地图上的图条 计算连续字段 (值) 的汇总统计, 并将各个地图特征 (数据关键字) 的分类字段 (类别) 的各个分类的结果显示为条形图, 这些条形图位于各个地图特征的中央位置。</p> <p>所需项: 一对分类字段和一个连续字段。关键字与数据关键字类别匹配的地图文件。</p>		<p>地图计数饼图 显示各个地图特征 (数据关键字) 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。</p> <p>所需项: 一对分类字段。关键字与数据关键字类别匹配的地图文件。</p>

表 45. 可用的图形类型 (续).

图表图标	描述	图表图标	描述
	<p>地图上的饼图</p> <p>计算各个地图特征 (数据关键字) 的分类字段 (类别) 的各个类别中连续字段 (值) 的总和, 并在各个地图特征的中央位置绘制一个由总和和分区组成的饼图。</p> <p>所需项: 一对分类字段和一个连续字段。关键字与数据关键字类别匹配的地图文件。</p>		<p>地图上的折线图</p> <p>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。</p> <p>所需项: 一个分类字段以及一对任意类型的字段。关键字与数据关键字类别匹配的地图文件。</p>
	<p>参考地图上的坐标</p> <p>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。</p> <p>所需项: 一对范围字段。一个地图文件。</p>		<p>参考地图上的箭头</p> <p>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。</p> <p>所需项: 四个连续字段。一个地图文件。</p>
	<p>点重叠地图</p> <p>绘制参考地图并在其上方放置一个点地图, 点特征由分类字段 (颜色) 进行着色。</p> <p>所需项: 一对分类字段。关键字与数据关键字类别匹配的地图文件。一个参考地图文件。</p>		<p>多边形重叠地图</p> <p>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。</p> <p>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。一个参考地图文件。</p>
	<p>折线重叠地图</p> <p>绘制参考地图并在其上方放置另一个折线地图, 折线特征由分类字段 (颜色) 进行着色。</p> <p>所需项: 一对分类字段。关键字与数据关键字类别匹配的折线地图文件。一个参考地图文件。</p>		

创建地图直观表示

对于许多直观表示, 您必须作出的选择只有两个: 感兴趣的字段 (变量) 和用于查看这些字段的模板。不需要进行其他选择或操作。地图直观表示至少需要一个额外步骤: 选择定义了图形直观表示的地理信息的地图文件。

创建简单地图的基本步骤如下:

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 145 页的『可用的内置图形板可视化类型』。

2. 选择地图目标。
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. 使用“选择地图”对话框选择地图文件和地图关键字。地图关键字的值必须与**数据关键字**指定的字段值匹配。您可以使用**比较**按钮来对这些值进行比较。如果要选择交叠地图模板，那么还需要选择参考地图。参考地图不是数据的关键。它用作主要地图的背景。有关“选择地图”对话框的更多信息，请参阅第 144 页的『选择用于地图直观表示的地图文件』。
7. 单击**确定**以关闭“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

本节通过多个不同的示例来介绍可用的选项。这些示例还提供了解释生成可视化的信息。

这些示例使用名为 *graphboard.str* 的流，该流参考名为 *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 安装程序的 *Demos* 文件夹中找到。此目录可通过 Windows IBM SPSS Modeler 程序组进行访问。*graphboard.str* 文件位于 *streams* 文件夹中。

建议您按照显示的顺序读取示例。因为后面的示例以前面的示例作为构建基础。

示例：含汇总统计的条形图

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**。（使用 Ctrl+单击可选择多个字段/变量。）
4. 选择 **条形**。
5. 从“摘要”下拉列表中选择**均值**。
6. 单击**运行**。
7. 在生成的显示内容中，单击“显示字段和值标签”工具栏按钮（工具栏中心内两个按钮组的第二个按钮）。

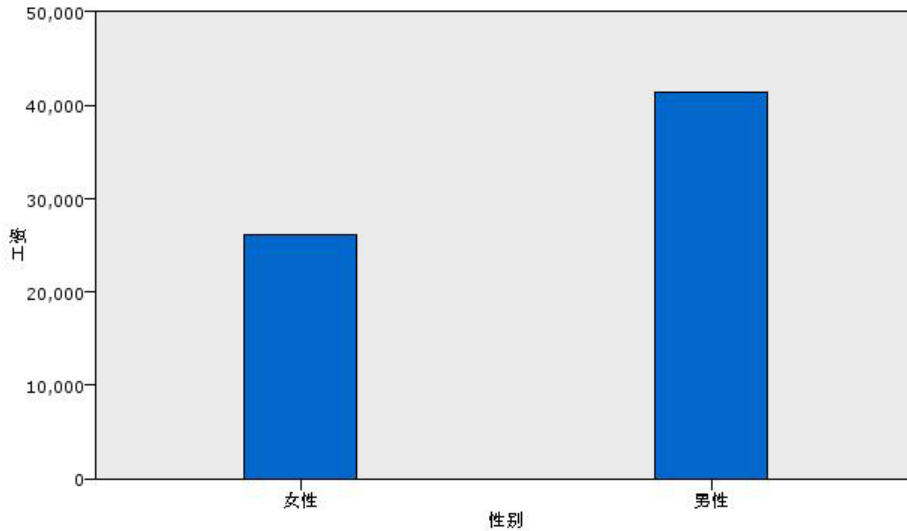


图 10. 含汇总统计的条形图

我们可以观察到以下内容:

- 根据条形的高度, 男性平均工资明显高于女性平均工资。

示例: 含汇总统计的复式条形图

现在我们将创建复式条形图, 看看男性与女性的平均工资之间的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。

注意: 此示例使用员工数据。

1. 添加图形板节点并打开它进行编辑。
2. 在“基本”选项卡上, 选择雇佣类别和当前工资。(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**。

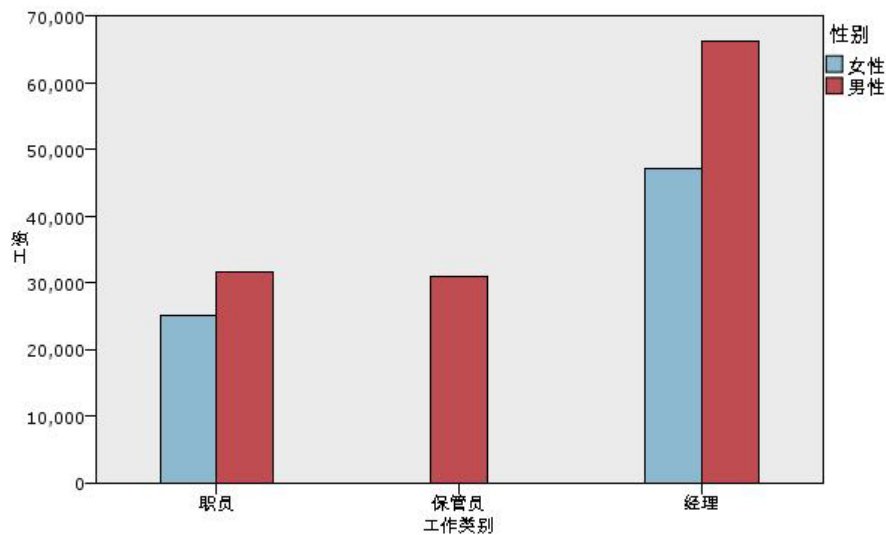


图 11. 簇状条形图

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: 直方图组

我们将按性别创建直方图组，以便比较男性工资和女性工资的频率分布。频率分布将反映特定工资范围内的观测值/行的数目。直方图组有助于我们进一步分析性别之间的工资差异。

注意: 此示例使用员工数据。

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 **当前工资**。
3. 选择 **直方图**。
4. 单击“详细”选项卡。
5. 在“面板和动画”组中，从“面板交叉”下拉列表选择 **性别**。
6. 单击**运行**。

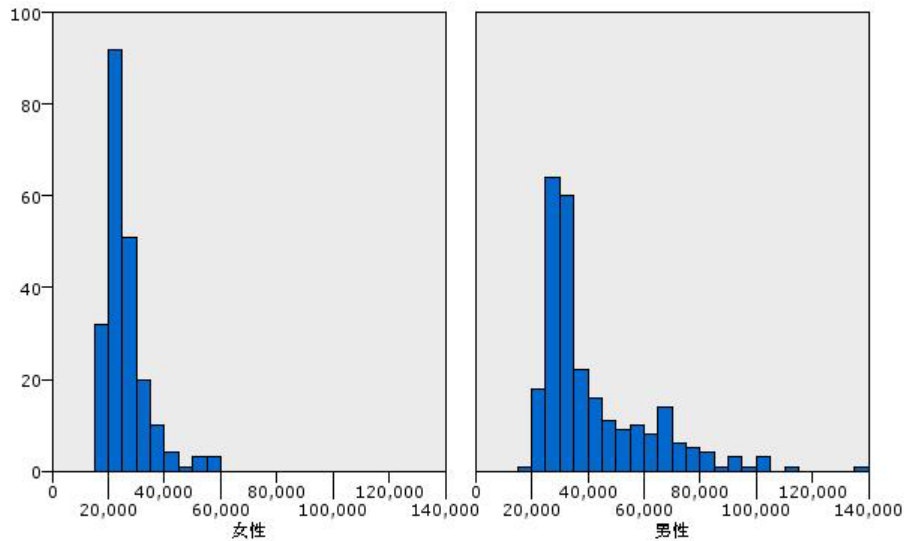


图 12. 直方图组

我们可以观察到以下内容:

- 没有频率分布属于正态分布。也就是说，直方图不呈现为正态分布曲线，如果数据呈正态分布，则该图也应呈正态分布。
- 较高的条形在各个图的左侧。因此，无论是男性还是女性，低工资的人都比高工资的人多。
- 男性和女性之间的工资频率分布不相等。请注意直方图的形状。高工资的男性比高工资的女性多。

示例: 点图组

与直方图相同之处在于，点阵图也显示连续数值范围的分布情况。不同之处在于，直方图显示数据分级范围的计数，而点阵图显示数据中的各个行/观测值。因此，点阵图比直方图的粒度更细。事实上，在分析频率分布时，可能开始使用点阵图更好。

注意: 此示例使用员工数据。

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 **当前工资**。
3. 选择 **点阵图**。
4. 单击“详细”选项卡。
5. 在“面板和动画”组中，从“面板交叉”下拉列表选择 **性别**。
6. 单击**运行**。
7. 最大化生成的输出窗口以更清楚地查看点图。

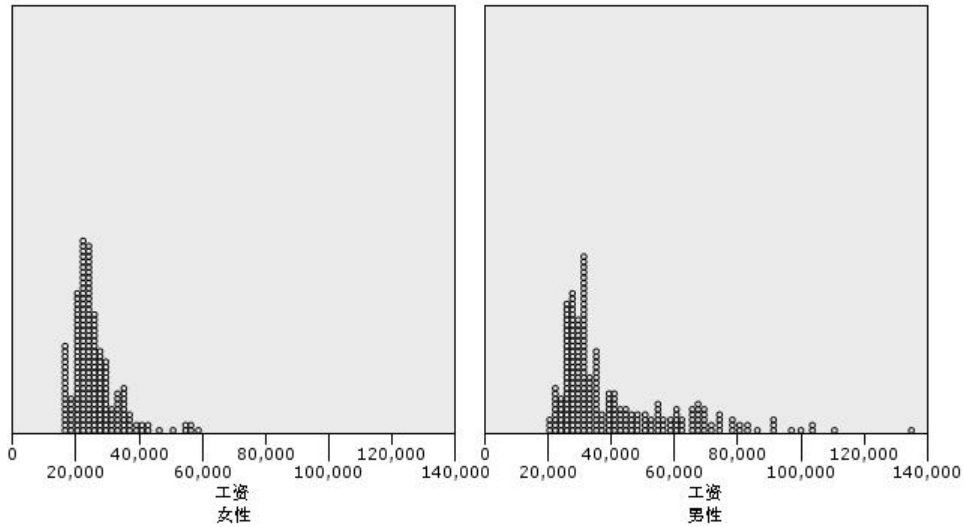


图 13. 点图组

通过与直方图进行比较（请参阅第 153 页的『示例：直方图组』），我们可以观察到以下内容：

- 直方图中女性的峰值为 20,000，比点阵图中的峰值要低很多。有许多观测值/行与该值有关，但这些值大多数都接近 25,000。该级别的粒度在直方图中并不明显。
- 虽然女性直方图表明女性平均工资在 40,000 以后逐渐下降，但点阵图显示该值为 80,000，两者大不相同。在该范围的任意一个工资值上，都有三个或三个以上的女性的工资为此特殊值。

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。长方形图中包含了多个统计测量量，我们在创建可视化工具后会讨论统计测量量。

注意：此示例使用员工数据。

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。（使用 Ctrl+单击可选择多个字段/变量。）
3. 选择 长方形图 。
4. 单击运行。

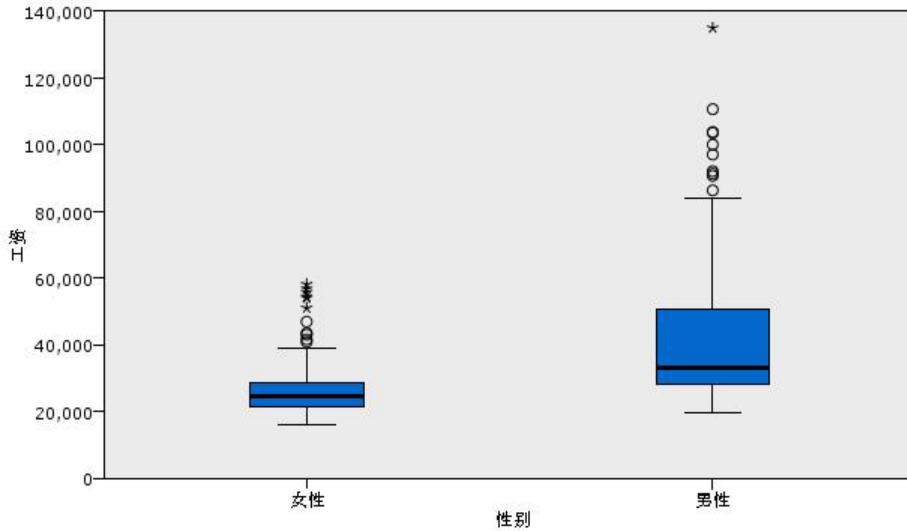


图 14. 箱图

现在我们来讨论长方形图的各个部分:

- 盒中间的暗线是 **工资** 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。与平均值相同，中位数是集中趋势的测量量。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。在此示例中，中位数小于平均值（与第 151 页的『示例：含汇总统计的条形图』相比较）。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。也就是说，有一小部分员工的工资非常高。
- 盒的底部表示第 25 个百分位数。25% 的观测值/行的值小于第 25 个百分位数。盒的顶部表示第 75 个百分位数。25% 的观测值/行的值大于第 75 个百分位数。这意味着 50% 的观测值/行位于该盒中。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 **内界限** 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，那么约有 95% 的数据落在内界限中。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。
- 这些点是 **离群值**。这些值是指落在内界限外的值。离群值是极值。星号或星星表示 **离群值极值**。这些值表示观测值/行的值是盒高度的三倍。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。平均值之所以比较大，原因就在于这些离群值。

示例：饼图

现在我们将使用不同的数据集研究一些其他可视化类型。数据集为 *customer_subset*，这是包含客户相关信息的虚拟数据文件。

我们将首先创建饼图以了解不同地理区域的客户比例。

1. 添加指向 *customer_subset.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 **地理指示**。
4. 选择 **计数饼图**。
5. 单击 **运行**。

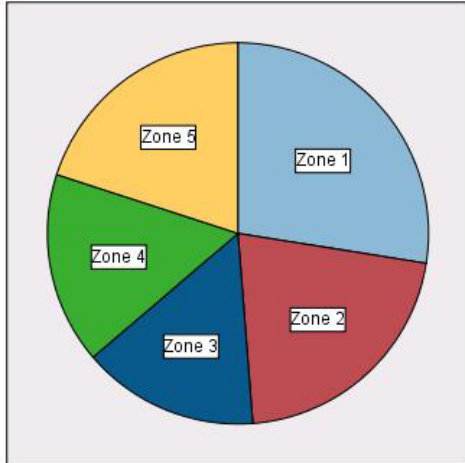


图 15. 饼图

我们可以观察到以下内容:

- 区域 1 中的客户数多于其他任何区域。
- 客户在其他区域之间均匀分布。

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。

注意: 此示例使用 *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)。(使用 Ctrl+单击可选择多个字段/变量。)
3. 选择 热图 。
4. 单击运行。
5. 在生成的输出窗口中, 单击“显示字段和值标签”工具栏按钮 (工具栏中心内两个按钮中右边的按钮)。

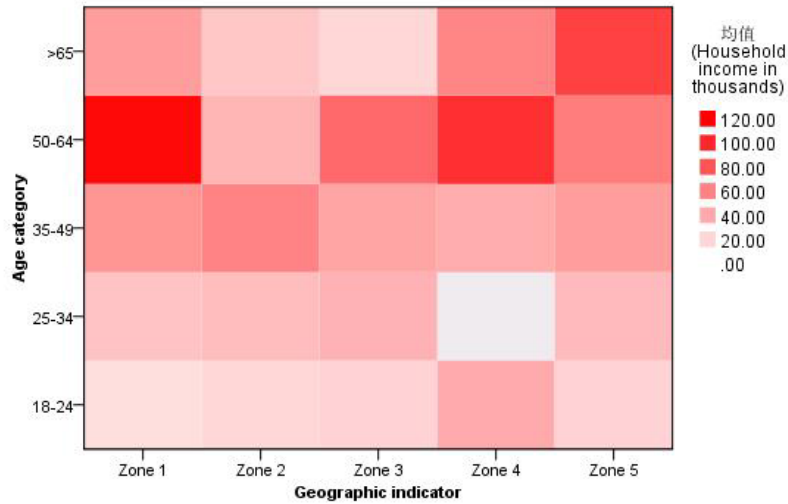


图 16. 分类热图

我们可以观察到以下内容:

- 热图与使用颜色（而不是数字）表示单元格值的表相似。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/变量的均值。
- 除区域 2 和区域 5 外，年龄在 50 到 64 之间的客户组的家庭收入高于其他组。
- 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)

我们将创建若干个不同变量的散点图矩阵图，以便确定数据集中的变量之间是否有关联。

注意: 此示例使用 *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，选择年龄（以年为单位）、家庭收入（以千美元为单位）和信用卡债务（以千美元为单位）。（使用 Ctrl+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。

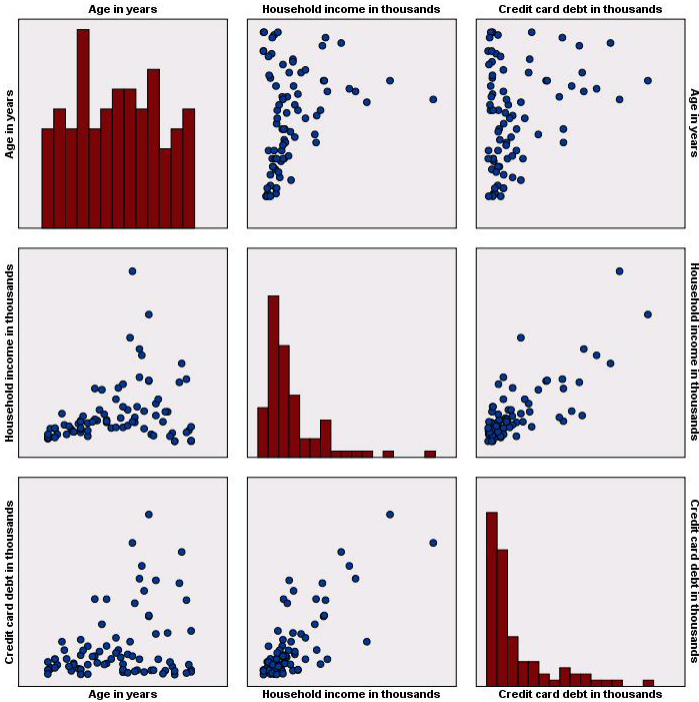


图 17. 散点图矩阵图 (SPLOM)

我们可以观察到以下内容:

- 对角上显示的直方图展示了 SPLOM 中各个变量的分布。*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。没有一个变量是正态分布。也就是说, 没有一个直方图呈现为正态分布曲线。另请注意, *income* 和 *creddebt* 的直方图为正偏态直方图。
- *age* 与其他变量之间似乎不存在任何关系。
- *income* 与 *creddebt* 之间存在一种线性关系。即, *creddebt* 随 *income* 增加而增加。您可能希望分别创建这些变量的散点图和其他相关变量, 以进一步研究它们之间的关联情况。

示例: 总和的等值线图 (颜色图)

现在, 我们将创建一个可视化图形。然后, 在随后的实例中, 我们将创建此可视化图形的变体。数据集为 *worldsales*, 这是一个包含按洲和产品分类的销售收入的虚拟数据文件。

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。(使用 Ctrl+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关键字匹配。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。另外, 我们还能看到没有大洋洲的数据。
9. 在“选择地图”对话框中, 单击确定。
10. 单击运行。



图 18. 总和的等值线图

从该可视化图形中，我们可以轻松地看到北美的收入最高，而南美和非洲的收入最低。由于我们使用洲作为数据标签外观，因此，对每个洲进行了标注。

示例：关于地图的条形图

此示例显示如何在各洲按产品对收入进行分类。

注意：此示例使用 *worldsales*。

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、产品和收入。（使用 Ctrl+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。

使用多个特定类型的字段时，请检查每个字段是否都分配到正确的通道，这一点非常重要。

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“数据关键字”下拉列表中选择洲。
8. 从“摘要”下拉列表中选择总和。
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关键字匹配。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。另外，我们还能看到没有大洋洲的数据。
12. 在“选择地图”对话框中，单击确定。

- 单击运行。
- 最大化生成的输出窗口以更清楚地查看显示。

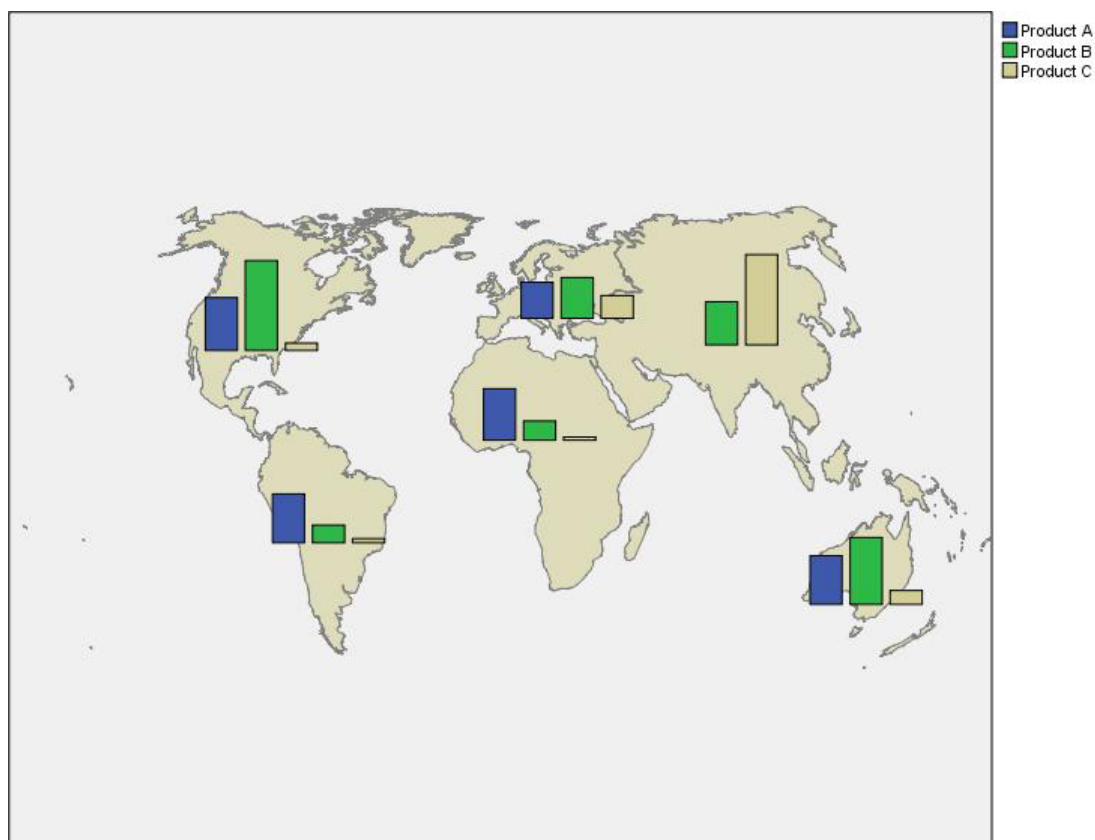


图 19. 关于地图的条形图

我们可以观察到以下内容：

- 在南美和非洲，跨产品总收入的分布非常相似。
- 产品 C 在除亚洲以外的任何地方都产生了最低收入。
- 产品 A 在亚洲没有收入或收入最少。

“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。 输入用于图形标题的文本。

子标题。 输入用于图形子标题的文本。

文字说明。 输入用于图形说明的文本。

抽样。 为较大数据集指定一种方法。可以指定数据集大小上限，或使用缺省记录数。如果选择 **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。

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、样式表和映射。可以在本地计算机上导入、导出、重命名和删除可视化模板、样式表和映射。请参阅主题第 163 页的『管理模板、样式表和地图文件』，了解更多信息。

位置。更改可视化模板、样式表和映射的存储位置。当前位置列出在按钮右侧。请参阅主题『设置模板、样式表和地图的位置』，了解更多信息。

以下示例显示外观选项在图形上的位置。（注意：并非所有图形都使用所有这些选项。）

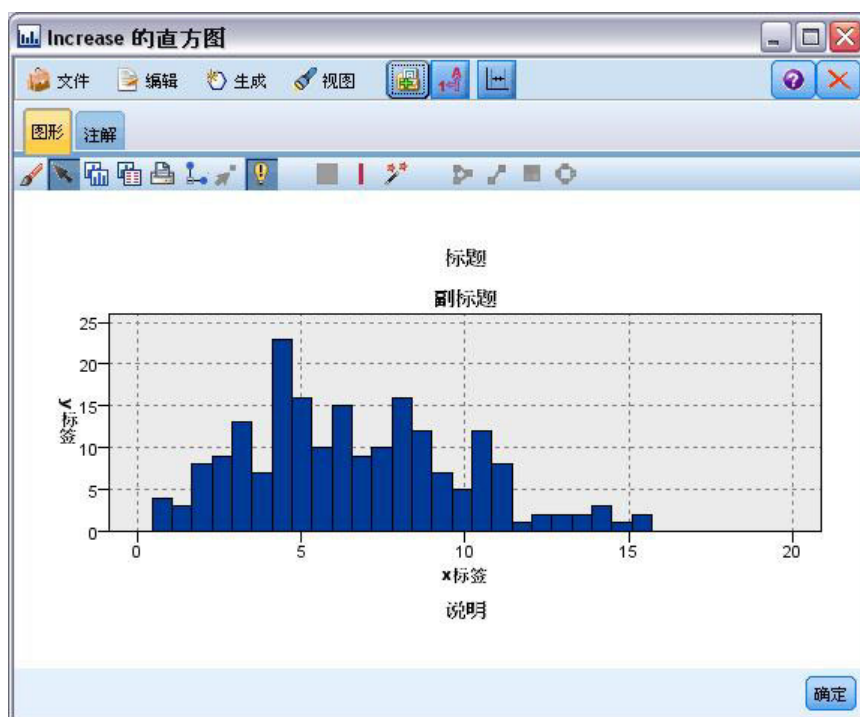


图 20. 各个图形外观选项的位置

设置模板、样式表和地图的位置

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 163 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默认位置选项：

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 中，此文件夹是 *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『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. 在具有“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**。

注意：如果您尚未连接到 IBM SPSS Collaboration and Deployment Services Repository，那么系统将提示您提供连接信息。

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定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。

单击您在其中选择模板、样式表或地图的一个对话框中的**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。请参阅主题第 225 页的『应用样式表』以获取更多信息。“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

使用图形板模板选择器可以根据可视化模板和 SMZ 文件的组合创建地图直观表示。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。图形板模板选择器随选定数目的 SMZ 文件一起预先安装。如果您具有要用于地图直观表示的现有 ESRI shapefile, 那么您首先需要使用地图转换实用程序将 shapefile 转换为 SMZ 文件。地图转换实用程序支持包含单一点的、折线或多边形 (形状类型分别为 1、3 和 5) ESRI shapefile。

除转换 ESRI shapefile 外, 使用地图转换实用程序还可以修改地图的详细信息级别、更改特征标签、合并特征、移动特征以及许多其他可选更改。另外, 还可以使用地图转换实用程序修改现有 SMZ 文件 (包括预先安装的 SMZ 文件)。

编辑预先安装的 SMZ 文件

1. 从管理系统导出 SMZ 文件。请参阅主题第 163 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 163 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 () 和 美国人口普查局 (<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。在本信息材料中对任何非 IBM 网站的引用仅为了方便用户, 并不以任何方式表明对这些网站的认可。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。

地图的关键概念

了解一些与 shapefile 相关的关键概念将帮助您高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下列三种类型的 shapefile:

- **点。** 此 shapefile 用于标识点的位置, 例如城市。
- **折线。** 此 shapefile 用于标识路径及其位置, 例如河流。
- **多边形。** 此 shapefile 用于标识有界性的区域及其位置, 例如国家或地区。

通常, 您将使用多边形 shapefile。等值线图根据多边形 shapefile 而创建。等值线图使用颜色来表示各个多边形 (区域) 内的值。通常, 点和折线 shapefile 在多边形 shapefile 上重叠。例如, 点 shapefile 美国城市在多边形 shapefile 美国各州上重叠。

shapefile 由特征组成。特征是各个地理实体。例如，特征可以是国家或地区、洲和城市等。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。至少存在一个作为特征的地图关键字的属性。地图关键字可以是标签，例如国家或地区或洲名。地图关键字是将链接到数据文件中的变量/字段以创建地图直观表示的内容。

注意，您将只能在 SMZ 文件中保留关键字属性或属性。地图转换实用程序不支持保存其他属性。这意味着，如果要在不同级别进行汇总，那么您将需要创建多个 SMZ 文件。例如，如果要对美国州和区域进行汇总，那么您需要不同的 SMZ 文件：具有标识州的关键字的文件以及具有标识区域的关键字的文件。

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

地图转换实用程序中存在四个主屏幕（步骤）。其中一个步骤还涉及对编辑地图文件进行更加详细控制的子步骤。

步骤 1 - 选择目标文件和源文件

首先，您需要选择一个源地图文件，并为转换后的地图文件选择目标。您将同时需要 shapefile 的 .shp 文件和 .dbf 文件。

选择要进行转换的 .shp (ESRI) 或 .smz 文件。浏览至计算机上的现有地图文件。这是将转换为并作为 SMZ 文件保存的文件。shapefile 的 .dbf 文件必须存储在基本文件名所在的位置，此基本文件名与 .shp 文件匹配。.dbf 文件是必需文件，因为它包含 .shp 文件的属性信息。

设置已转换地图文件的目标和文件名。输入将根据原始地图源创建的 SMZ 文件的路径和文件名。

- **导入至模板选择器。**除在文件系统上保存文件外，您还可以（可选）将地图添加到模板选择器的管理列表。如果选择此选项，那么地图将自动显示在模板选择器中，以供计算机上安装的 IBM SPSS 产品使用。如果现在未将地图导入至模板选择器，那么您需要在稍后以手动方式进行导入。有关将地图导入模板选择器管理系统的更多信息，请参阅第 163 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，您将选择要包括在 SMZ 文件中的地图关键字。接着，您可以更改某些将影响地图呈现的选项。地图转换实用程序中的后续步骤包括预览地图。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，某些属性可能包含已转换的标签。如果期望使用其他语言编码数据，那么您可能希望保留这些属性。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，如果主要关键字是 U.S. states 的全名，那么您只能选择那些表示 U.S. States 的替换关键字，例如，state 缩写。

自动平滑地图。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。自动平滑地图时，最大角度为 15 度，要保留的百分比为 99。有关这些设置的信息，请参阅『平滑地图』。请注意，您可以稍后在其他步骤中应用更多平滑功能。

除去同一特征中接触多边形之间的边界。某些特征可能包括子特征，这些子特征的边界位于相关主特征的内部。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，选择此选项将除去国家或地区边界，同时保留大洲边界。

步骤 3 - 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。这些修改是可选项的。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。

每个任务在地图转换实用程序的左侧都有下列公共控件。

在地图上显示标签。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。

对地图预览进行着色。缺省情况下，预览地图以纯色显示区域。所有特征都是同一颜色。您可以选择向每个地图特征都分配一种颜色。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。

平滑地图：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。此选项对于点地图和折线地图不适用。

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此线作为端点之间的直线进行处理并删除中间点。

要保留的百分比。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。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。即，特征将始终至少存在一个多边形，而与应用的平滑量无关。

选择最大角度和要保留的百分比之后，请单击**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。如果反复执行平滑，那么您将达到无法对地图应用任何平滑的程度。

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。缺省情况下不会在预览中显示标签，因此，您可能还希望选择**在地图上显示标签**以显示这些标签。

关键字。选择包含要查看和/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，请在列表中双击该标签。如果标签显示在地图上，那么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，请单击**比较**。

X/Y。这些文本框列出地图上所选特征的标签的当前中心点。单元在地图的坐标中显示。这些坐标可以是本地坐标、笛卡尔坐标（例如，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。如果显示了标签，那么您还可以单击标签并在地图上拖动来移动此标签。文本框将通过新位置进行更新。

比较。如果您具有一个数据文件，此文件包含的数据值旨在与特定关键字的特征标签匹配，请单击**比较**以显示“与外部数据源进行比较”对话框。您可以在此对话框中打开此数据文件，并直接将其值与地图关键字的特征标签中的值进行比较。

“与外部数据源进行比较”对话框： 使用“与外部数据源进行比较”对话框可以打开一个制表符分隔的值文件（带有 .txt 扩展名）、以逗号分隔的值文件（带有 .csv 扩展名）或针对 IBM SPSS Statistics 格式化的数据文件（带有 .sav 扩展名）。打开文件后，您可以在数据文件中选择一个字段来与特定地图关键字中的特征标签进行比较。然后，可以更正地图文件中的任何差异。

数据文件中的字段。选择要将其值与特征标签进行比较的字段。如果 .txt 或 .csv 文件中的第一行包含各个字段的描述性标签，请选中**使用第一行作为列标签**。否则，将通过字段在数据文件中的位置来标识各个字段（例如，“第 1 列”，“第 2 列”，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。准备好比较值后，请单击此选项。

比较结果。缺省情况下，“比较结果”表仅列出数据文件中不匹配的字段值。通常，应用程序通过查找感兴趣的区域或缺失的区域来尝试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**。如果要查看所有字段值（甚至包括那些已与特征标签匹配的值），请取消选中**仅显示不匹配的观测值**。如果要覆盖一个或多个匹配项，您可能需要执行此操作。

只能使用各个特征一次来将其与字段值匹配。如果要将多个特征与单一字段值匹配，那么可以合并这些特征，并将新的合并后的特征与字段值匹配。有关合并特征的更多信息，请参阅『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。

关键字。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。请参阅主题第 166 页的『编辑特征标签』，了解更多信息。

“命名合并后的特征”对话框： 使用“命名合并后的特征”对话框可以将标签分配给新合并的特征。

“标签”表显示地图文件中每个关键字的信息，并使您能够分配各个关键字的标签。

新标签。输入合并后的特征的新标签以将其分配给特定地图关键字。

关键字。将新标签分配到的地图关键字。

旧标签。这是将合并到新特征中的特征的标签。

除去接触多边形之间的边界。选中此选项可除去已合并在一起的特征中的边界。例如，如果您将州合并到地理区域，那么此选项将除去各个州周围的边界。

移动特征：可以除去地图中的特征。如果要将特征集中在一起（例如大陆及周围的岛屿），那么此选项非常有用。

关键字。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除去的特征。

特征。单击要移动的特征。注意，将在地图预览中选中这些特征。另外，也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。单元在地图的坐标中显示。这些坐标可以是本地坐标、笛卡尔坐标（例如，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。您还可以在地图上单击并拖动特征来进行移动。文本框将通过新位置进行更新。

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。

关键字。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。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。但是，某些投影更加适合，具体取决于您查看的是全局地图还是本地地图。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。此选项仅适用于具有地理坐标（经度和纬度）的地图。

与地图转换实用程序中的其他选项不同，可以在创建地图直观表示后更改投影。

投影。选择地图投影。如果要创建全局地图或半球面地图，请使用**局部**、**墨卡托**或**温克尔三重**投影。对于较小的区域，请使用**局部**、**兰勃特等角圆锥**或**横轴墨卡托**投影。所有投影都对数据使用 **WGS83 ellipsoid**。

- 如果地图是使用本地坐标系（例如，国家平面坐标系）创建的，那么将始终使用**局部**投影。这些坐标系由笛卡尔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尔坐标系中等距排列。局部投影不是正形投影。
- **墨卡托**投影是全局地图的正形投影。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。
- **温克尔三重**投影不是全局地图的正形投影。虽然它不是正形投影，但它提供了形状与大小之间的很好的平衡。除赤道和本初子午线外，所有线都是曲线。如果全局地图包含北极或南极，那么这是非常好的投影选项。
- 顾名思义，**兰勃特等角圆锥**投影是一个正形投影，用于大陆或更小大陆块的地图，这些大陆或大陆块东部和西部的长度大于北部和南部。

- **横轴墨卡托投影**是另一正行投影，用于大陆或更小大陆块的地图。对于北部与南部较东部与西部较长的大陆块，将使用此投影。

步骤 4 - 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。如果地图文件中存在多个关键字，请选择希望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。此注释将显示在“图形板模板选择器”的管理系统中。

根据特征标签创建数据集。如果要根据显示的特征标签创建数据文件，请选中此选项。单击**浏览...**之后，您将能够指定位置和文件名。如果添加了 *.txt* 扩展名，那么将此文件作为制表符分隔的值文件进行保存。如果添加了 *.csv* 扩展名，那么将此文件作为逗号分隔的值文件进行保存。如果添加了 *.sav* 扩展名，那么将以 **IBM SPSS Statistics** 格式保存此文件。未指定扩展名时，**SAV** 是缺省值。

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，这些用户可以使用管理系统来导入地图。您可以只发送在步骤 1 中指定了其位置的文件。如果要发送位于管理系统中的文件，那么需要首先将其导出：

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，您可以将物理地图文件发送给其他用户。用户将需要反转此过程并将地图导入管理系统。

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交叠，则将在图形上生成多条不同色调的、贯穿左右的线。

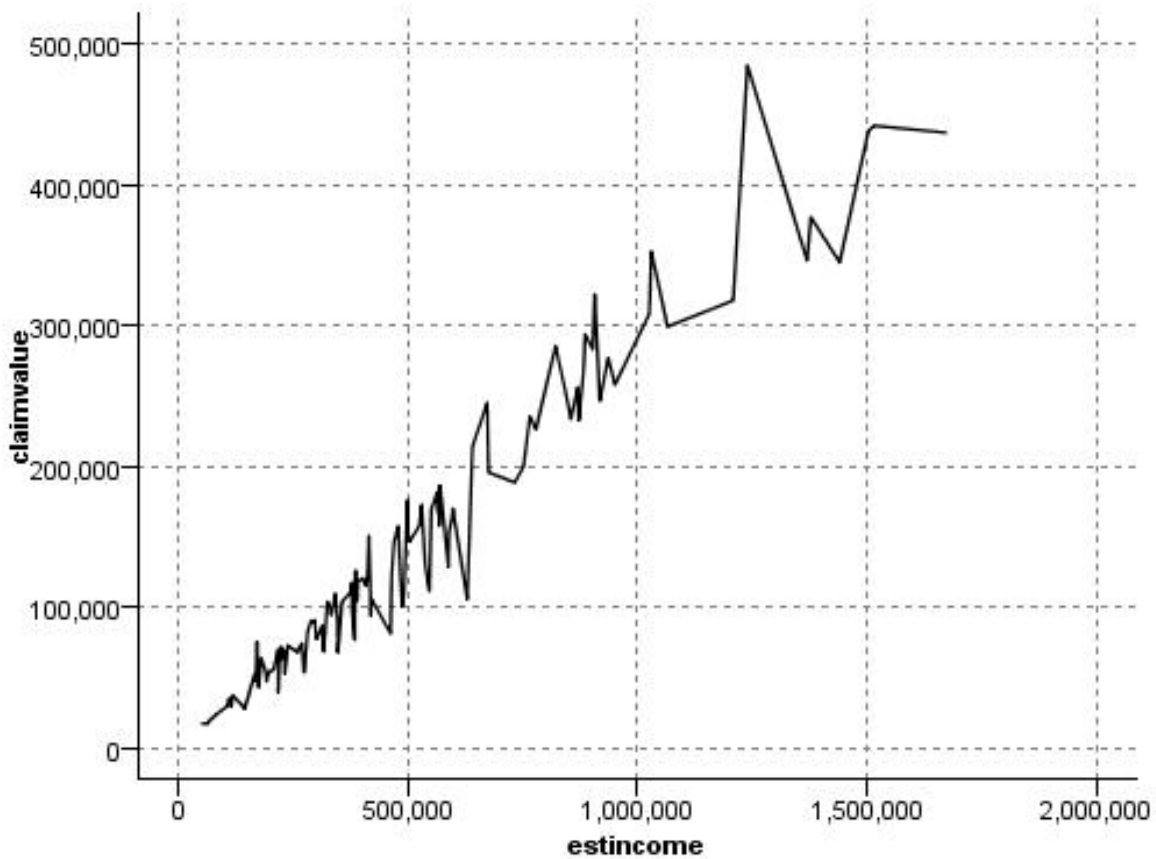


图 21. X 模式设置为“排序”的线散点图

X 模式 = 交叠

如果将 X 模式设置为 **交叠**，则将在同一图形上创建多线散点图。不会对交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。

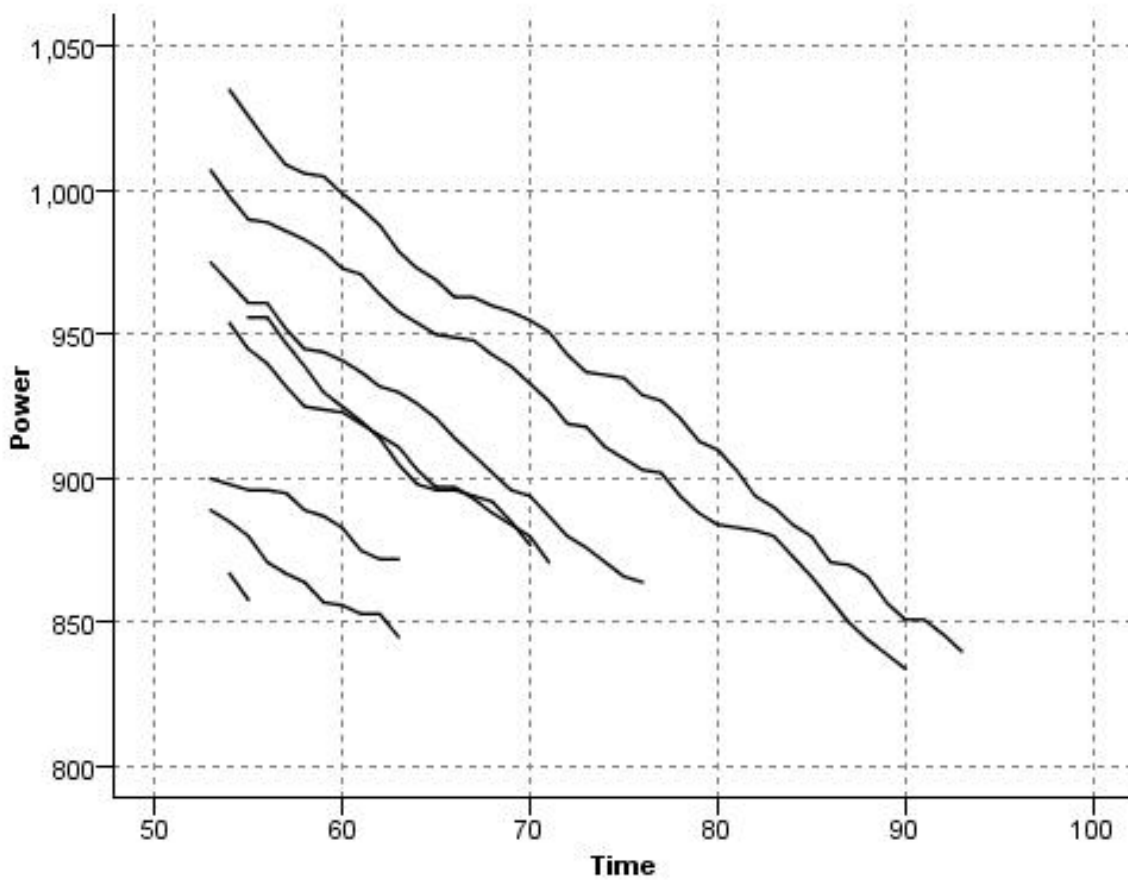


图 22. X 模式设置为“交叠”的线散点图

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。

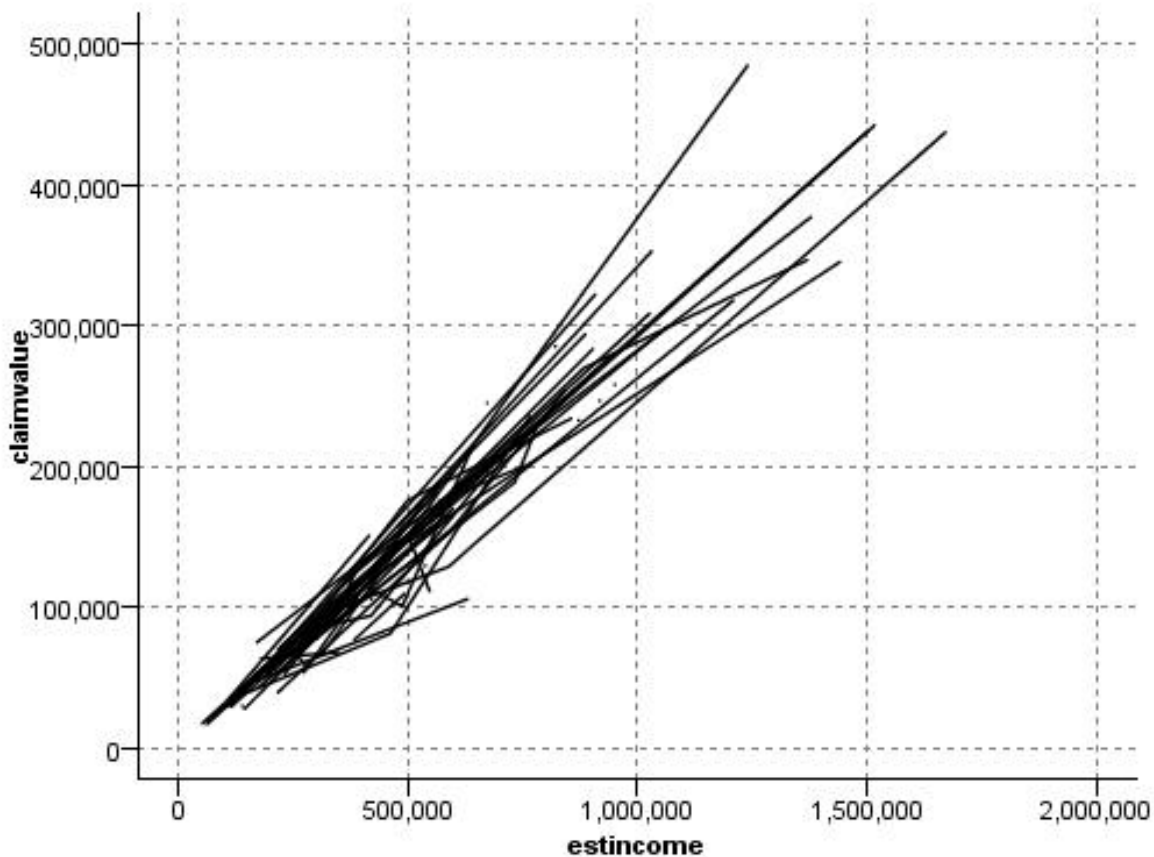


图 23. 线散点图起先显示为“排序”，随后以“如所读取”的 X 模式再次执行

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。请参阅主题第 145 页的『可用的内置图形板可视化类型』，了解更多信息。

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。

交叠。 有几种方式可用于说明数据值类别。例如，可以使用 *maincrop* 作为颜色交叠来指定补贴申请人种植的主要作物的 *estincome* 和 *claimvalue* 值。请参阅主题第 138 页的『审美原则、交叠、面板和动画』，了解更多信息。

交叠类型。 指定是否显示交叠函数或光滑线。光滑线和交叠函数总是作为 y 的函数来计算。

- 无。不显示任何交叠。
- 光滑线。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。

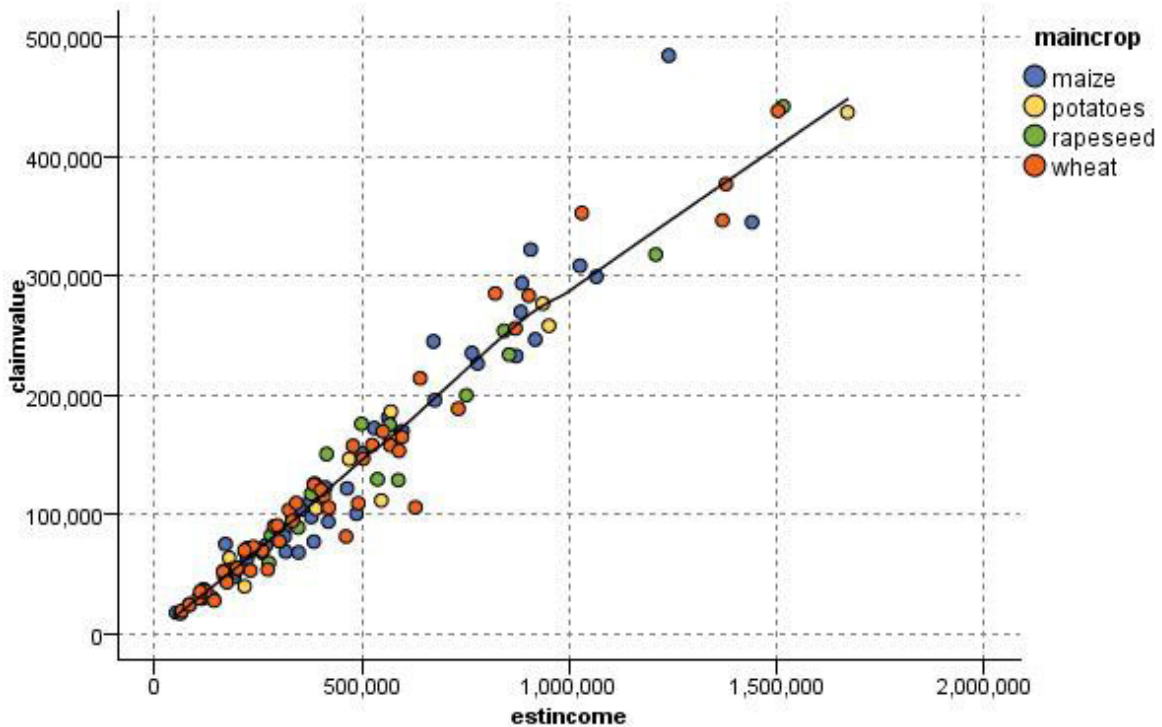


图 24. 具有 LOESS 光滑交叠的图

- **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 $y = x$ 作为交叠。在文本框中指定函数 $y =$ 。默认函数为 $y = x$ ，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: 交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 (+) 作为默认的点形状。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**、**交叠**或**如所读取**。如选择 **交叠** 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。默认情况下，将选择自动范围以快速构建图形。

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。默认情况下，将选择自动范围以快速构建图形。

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。默认情况下，将选择自动范围以快速构建图形。

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。

以前版本的 *IBM SPSS Modeler* 用户请注意：在 *IBM SPSS Modeler* 的本发行版中，散点图的抖动值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**交叠**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（除非您用大小作为交叠）。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。输入用于图形标题的文本。

子标题。输入用于图形子标题的文本。

文字说明。输入用于图形说明的文本。

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实质都是 X 相对于 Y 的散点图。例如，如果要调查农业补助申请中的潜在欺诈行为（如 *IBM SPSS Modeler* 安装的 *Demos* 文件夹中的 *fraud.str* 所示），那么您可能希望绘制申请中声明的收入相对于由神经网络算法估算的收入的散点图。可以使用交叠，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。

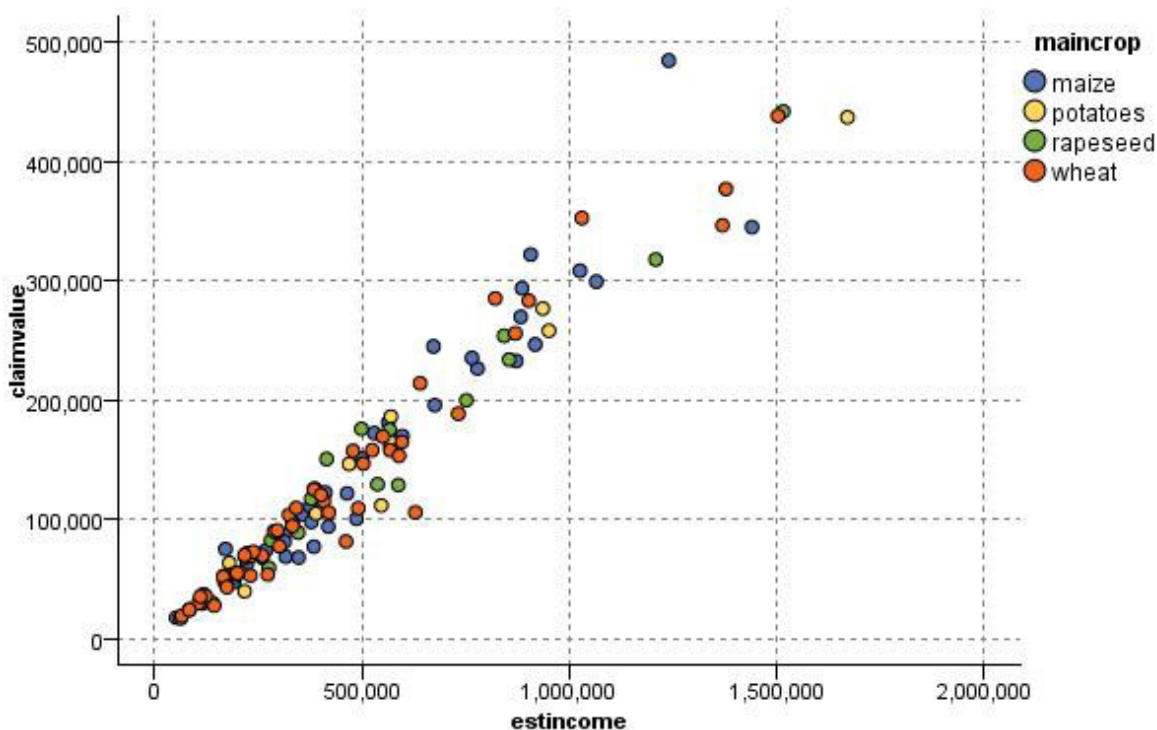


图 25. 将主要作物类型作为交叠来绘制估计收入与索赔值之间关系的散点图

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。请参阅主题第 203 页的『探索图形』，了解更多信息。

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。请参阅主题第 145 页的『可用的内置图形板可视化类型』，了解更多信息。

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*）数量的交叠字段，说明对活动的热情与家庭大小的关系。请参阅主题第 138 页的『审美原则、交叠、面板和动画』，了解更多信息。

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。交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。输入用于图形标题的文本。

子标题。输入用于图形子标题的文本。

文字说明。输入用于图形说明的文本。

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。

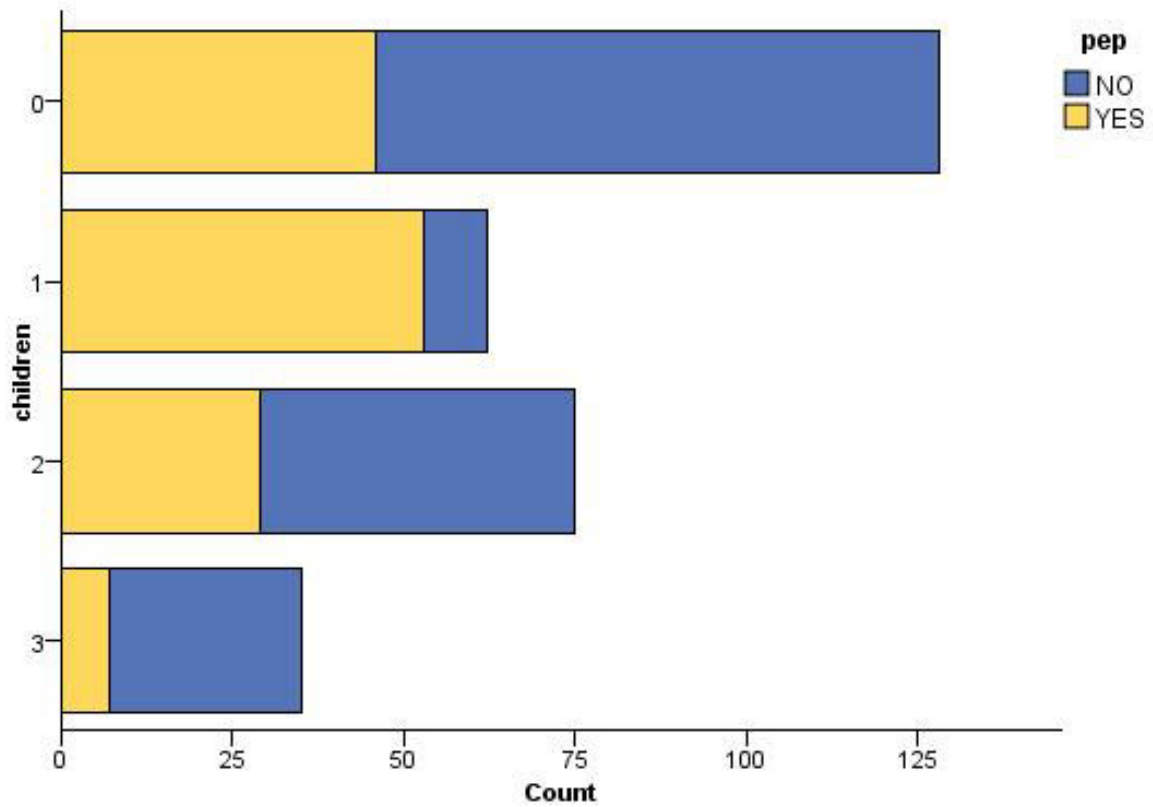


图 26. 分布图显示对市场活动有响应的人（有孩子或没孩子）数所占比例

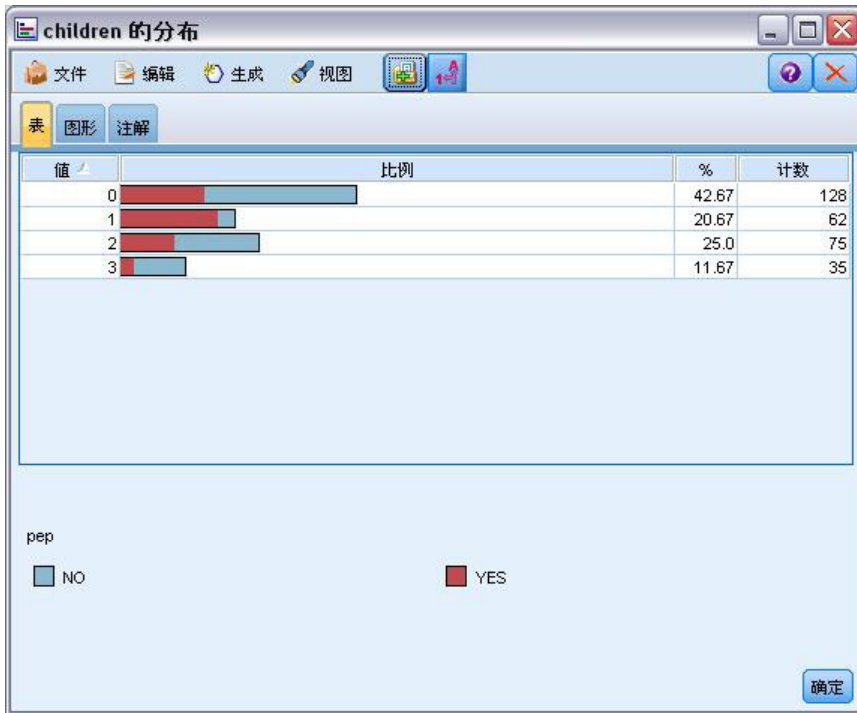


图 27. 分布表显示对市场活动有响应的人（有孩子或没孩子）数所占比例

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。请参阅主题第 226 页的『打印、保存、复制和导出图形』，了解更多信息。

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. 在“编辑”菜单中，选择**复制表**或**复制表（包括字段名称）**。
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交叠的值。

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。请参阅主题第 211 页的『从图形中生成节点』，了解更多信息。

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。请参阅主题第 145 页的『可用的内置图形板可视化类型』，了解更多信息。

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。

交叠。选择一个符号字段以显示指定字段中值的类别。选择交叠字段将使直方图变成堆积图，图中以各种颜色显示交叠字段的不同类别。如果使用“直方图”节点，那么将存在三种类型的交叠：颜色、面板和动画。请参阅主题第 138 页的『审美原则、交叠、面板和动画』，了解更多信息。

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。默认情况下，将选择自动范围以快速构建图形。

分级。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图条调整为同一高度，以便交叠的值在每个图条中显示为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显示了数据平均值和方差的正态曲线。

每种颜色的独立带状区域。选择此选项将在图形中为每个交叠的值分别显示一个带状区域。

直方图外观选项卡

可以在创建图形前指定外观选项。

标题。输入用于图形标题的文本。

子标题。输入用于图形子标题的文本。

文字说明。输入用于图形说明的文本。

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。请参阅主题第 203 页的『探索图形』，了解更多信息。

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。请参阅主题第 211 页的『从图形中生成节点』，了解更多信息。

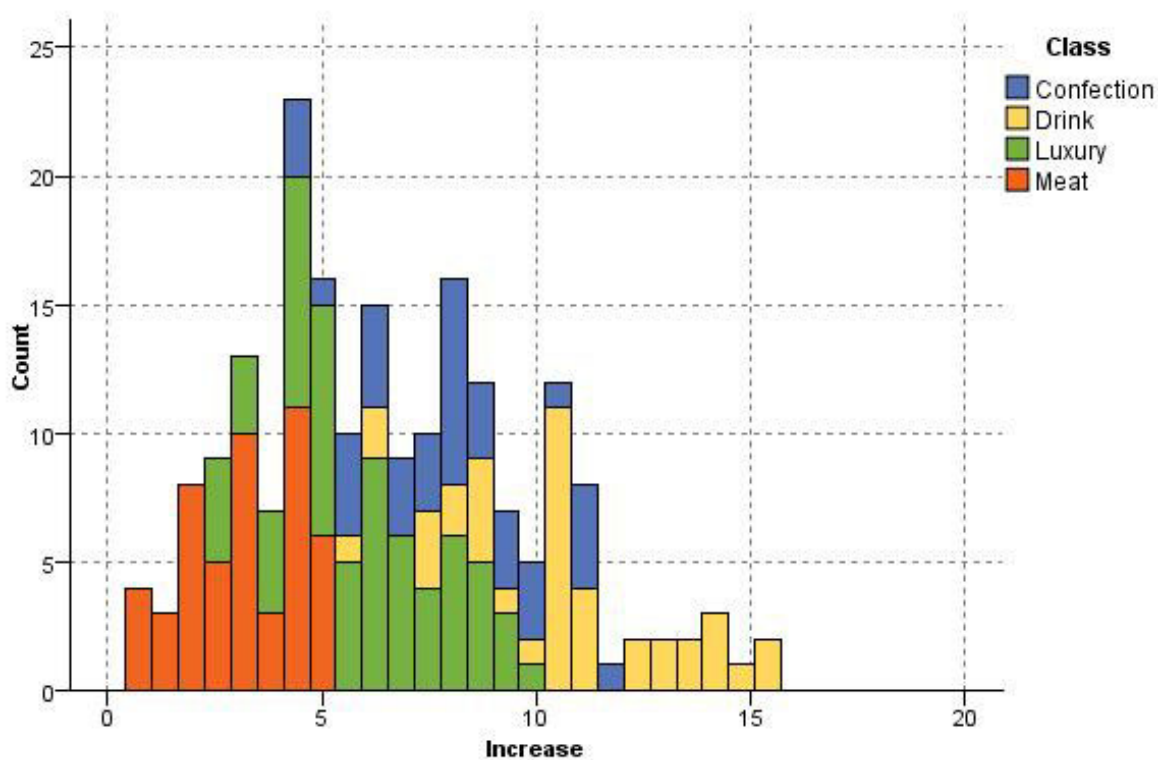


图 28. 直方图分类显示由促销活动带来的购买增长情况的分布

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交叠的堆积条形图。请参阅主题第 138 页的『审美原则、交叠、面板和动画』，了解更多信息。

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。

超出。 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**。

交叠。 选择一个符号字段以显示选定字段中值的类别。选择一个交叠字段，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交叠：**颜色**、**面板**和**动画**。请参阅主题第 138 页的『审美原则、交叠、面板和动画』，了解更多信息。

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。默认情况下，将选择自动范围以快速构建图形。

分级。 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。 输入用于图形标题的文本。

子标题。 输入用于图形子标题的文本。

文字说明。 输入用于图形说明的文本。

超出标签。 接受自动生成的标签，或选择**自定义**指定标签。

收集标签。 接受自动生成的标签，或选择**自定义**指定标签。

按照标签。 接受自动生成的标签，或选择**自定义**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。

以下示例显示外观选项在 3D 版本图形上的位置。

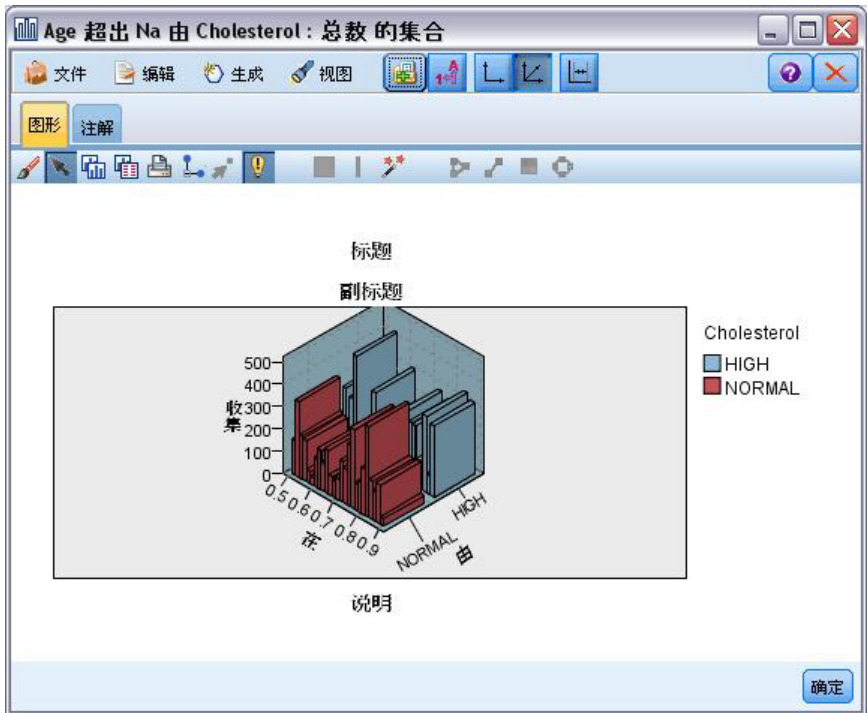


图 29. 三维集合图外观选项的位置

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。请参阅主题第 203 页的『探索图形』，了解更多信息。

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 *Derive Flag* 节点以添加一个字段，将符合条件的记录标记出来；或生成 *Select* 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。请参阅主题第 211 页的『从图形中生成节点』，了解更多信息。

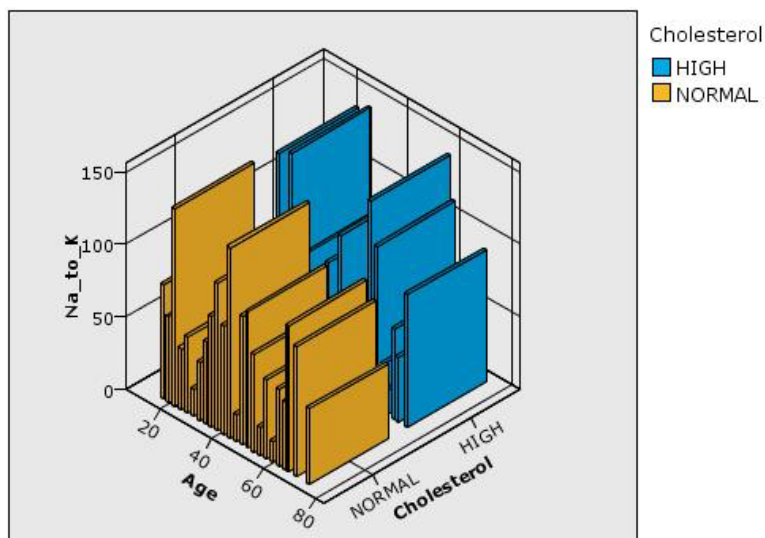


图 30. 三维集合图显示针对高、正常胆固醇水平，Na_to_K 相对于 Age 的合计

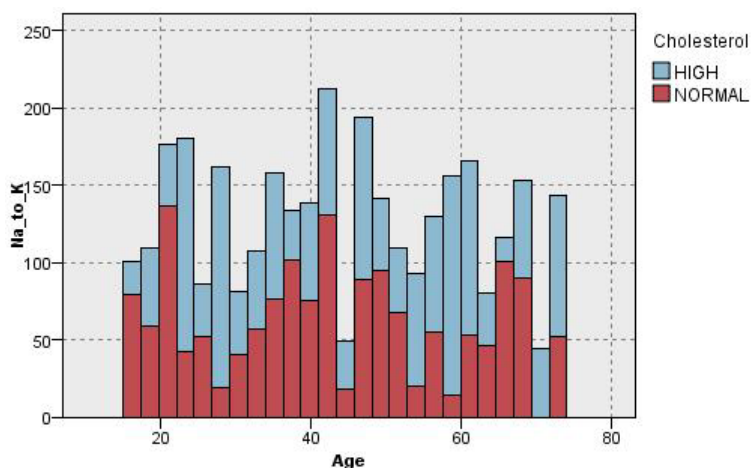


图 31. 显示没有 z 轴的集合图，并将胆固醇作为交叠以颜色标出

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。

交叠。有几种方式可用于说明数据值类别。例如，您可以使用动画交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。请参阅主题第 138 页的『审美原则、交叠、面板和动画』，了解更多信息。

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。）

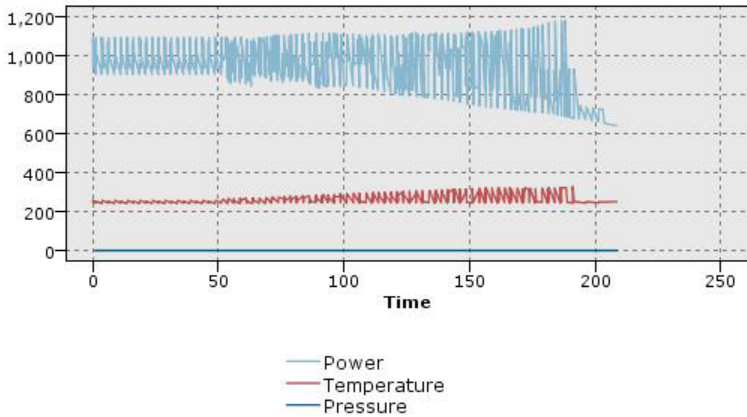


图 32. 显示发电厂在一段时间中的变化的标准多重散点图（请注意，如果未选择“标准化”，则压力的散点图不可见）

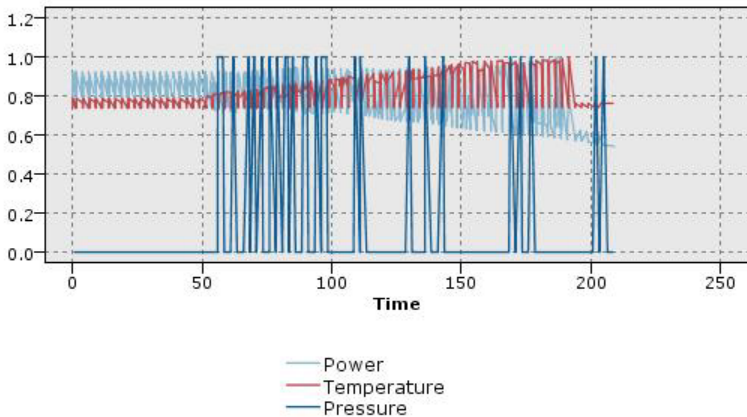


图 33. 显示压力散点图的标准化多重散点图

交叠函数。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 $y = x$ 作为交叠。在文本框中指定函数 $y =$ 。默认函数为 $y = x$ ，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：交叠函数不可用于面板图或动画图。

当记录数大于。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**交叠**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。输入用于图形标题的文本。

子标题。输入用于图形子标题的文本。

文字说明。输入用于图形说明的文本。

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实质都是 X 相对于 Y 的散点图。例如，如果要调查农业补助申请中的潜在欺诈行为（如 IBM SPSS Modeler 安装的 *Demos* 文件夹中的 *fraud.str* 所示），那么您可能希望绘制申请中声明的收入相对于由神经网络算法估算的收入的散点图。可以使用**交叠**，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。

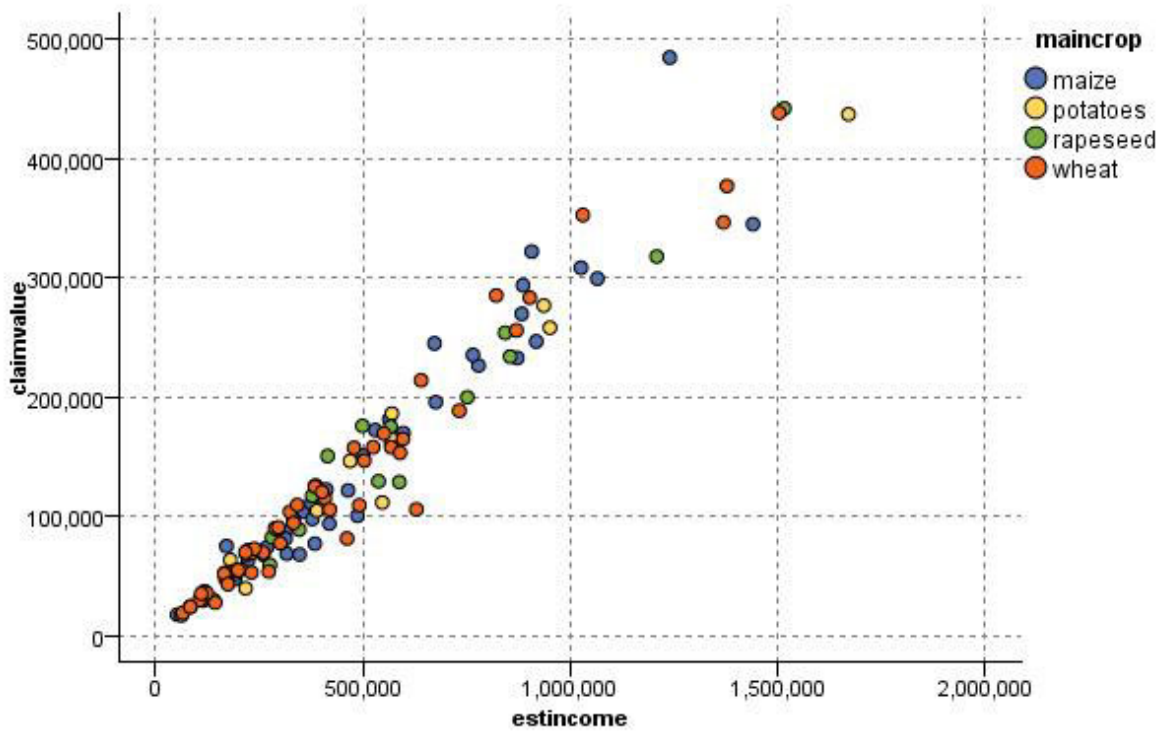


图 34. 将主要作物类型作为交叠来绘制估计收入与索赔值之间关系的散点图

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。请参阅主题第 203 页的『探索图形』，了解更多信息。

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。

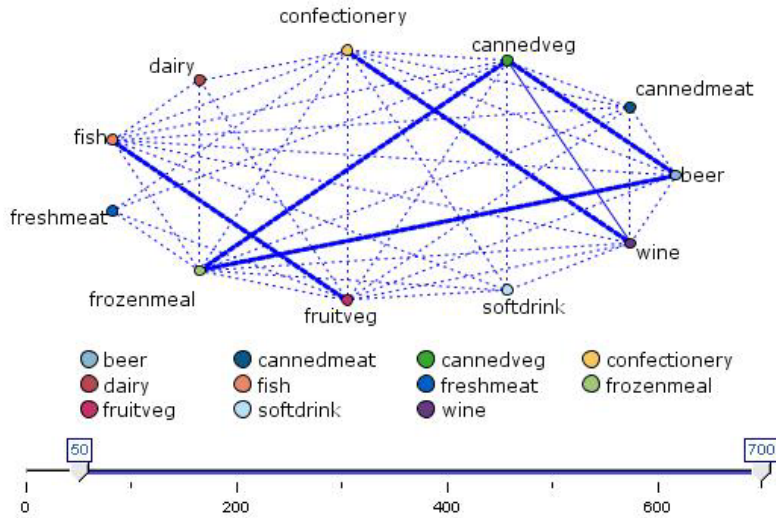


图 35. Web 图形显示购买杂货物品之间的关系

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。

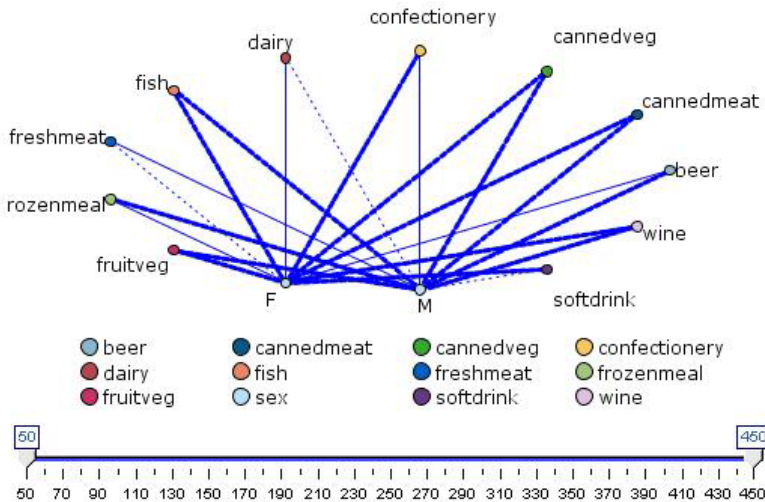


图 36. 定向 Web 图形显示购买杂货物品与性别之间的关系

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。

结束字段（仅用于定向 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”。

仅显示真值标志。 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。

线值为。 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时经常使用此选项。

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。

若记录过少则废弃。 选择此选项将忽略受少量记录支持的连接。在 **最小记录/行** 中输入一个数字，为此选项设置阈值。

若记录过多则废弃。 选择此选项将忽略受到较强支持的连接。在 **最大记录/行** 中输入数字。

以下弱链接。 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。所有低于该值的链接都被认为是弱链接。

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。

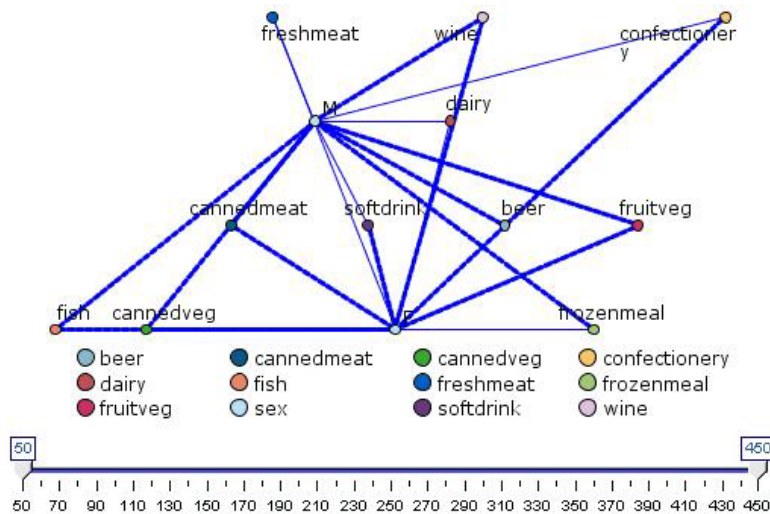


图 37. 显示从 *frozenmeal* 及 *cannedveg* 到其他杂货的强链接的网络图形

网络外观选项卡

可以在创建图形前指定外观选项。

标题。输入用于图形标题的文本。

子标题。输入用于图形子标题的文本。

文字说明。输入用于图形说明的文本。

显示图例。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网络图形

使用网络图形

Web 节点用于显示两个或多个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。

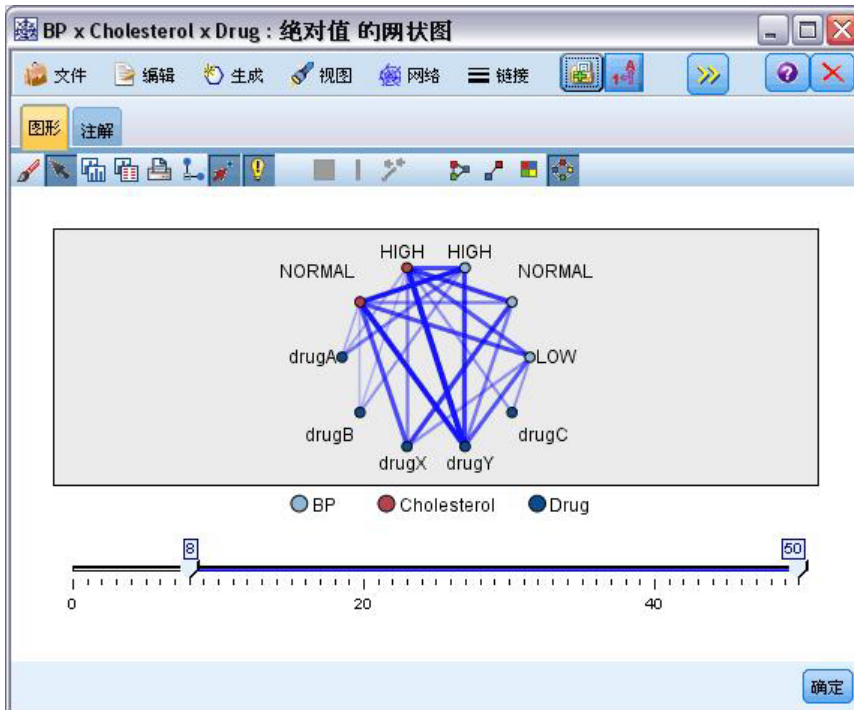


图 39. 网络图形说明一组紧密程度高的关系，如正常血压与 DrugX 及高胆固醇与 DrugY。

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。

以不同布局查看网络

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**。

选择或标记出单个关系中的记录

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。标志字段的名称以两个值的名称中间加上下划线表示，例如， *LOW_drugC* 或 *drugC_LOW* 。

选择或标记出一组关系中的记录

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。

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。

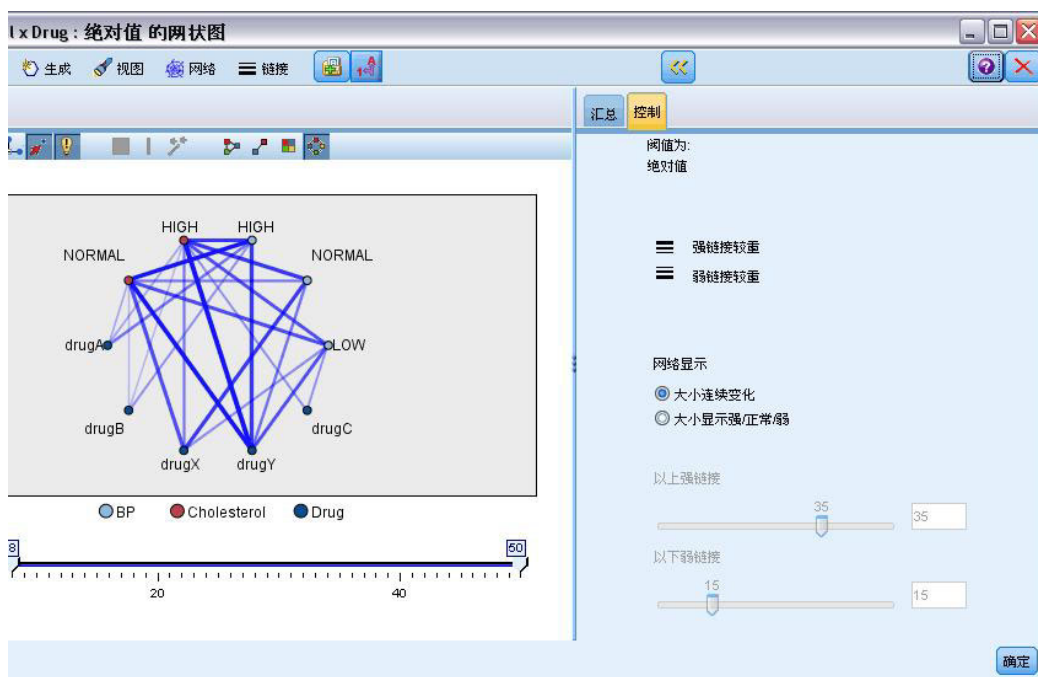


图 40. 扩展窗口中的选项主要针对显示和阈值

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。所有高于该值的连接都被认为是强连接。使用滑块调整值或在字段中输入一个数字。

以下弱链接。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。

创建网络汇总

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

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。通常，您在使用时间散点图节点前，应使用时间区间节点，创建 *TimeLabel* 字段。默认情况下，在图形中，该字段为 *x* 轴标签。请参阅主题第 125 页的『时间区间节点』，了解更多信息。

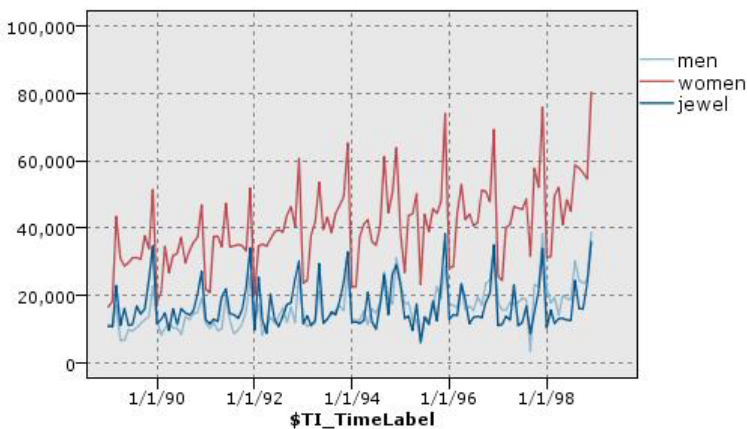


图 41. 绘制一段时间内男士及女士服装、珠宝的销售额

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。请参阅主题第 105 页的『导出节点』，了解更多信息。

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。

- **选定的序列。** 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。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“时间区间”节点中创建的 TimeLabel 字段作为上游。如果没有“时间区间”节点，那么将选择有序整数列。请参阅主题第 125 页的『时间区间节点』，了解更多信息。

在单独面板中显示序列。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。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。输入用于图形标题的文本。

子标题。输入用于图形子标题的文本。

文字说明。输入用于图形说明的文本。

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。请参阅主题第 203 页的『探索图形』，了解更多信息。

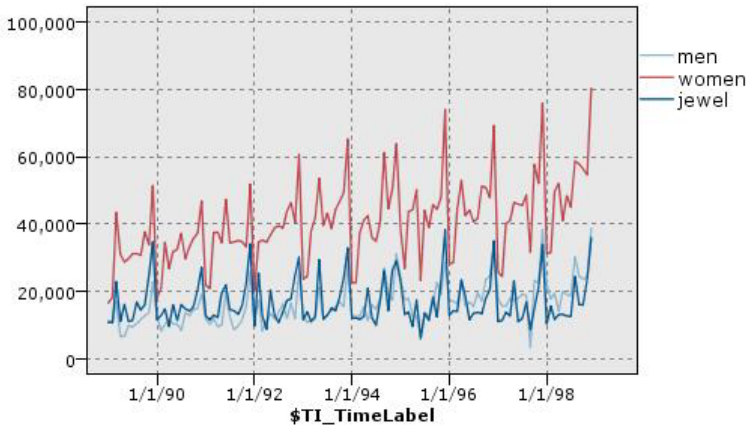


图 42. 绘制一段时间内男士及女士服装、珠宝的销售额

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。请参阅主题第 211 页的『从图形中生成节点』，了解更多信息。

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，或使用以下描述的默认匹配标准：

- 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。

收益图表

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%。

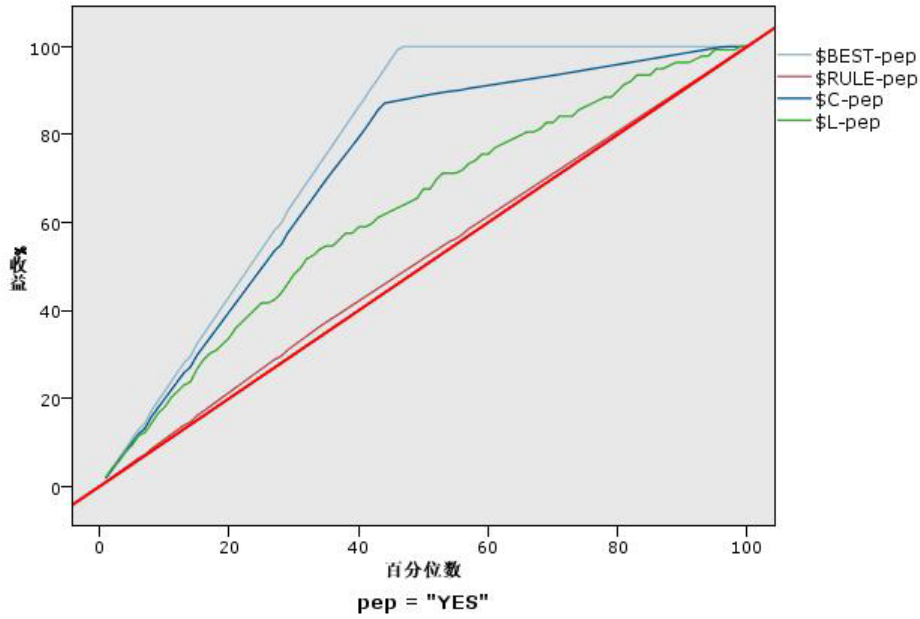


图 43. 显示带有基线、最佳线及业务规则的收益图（累积）

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）。

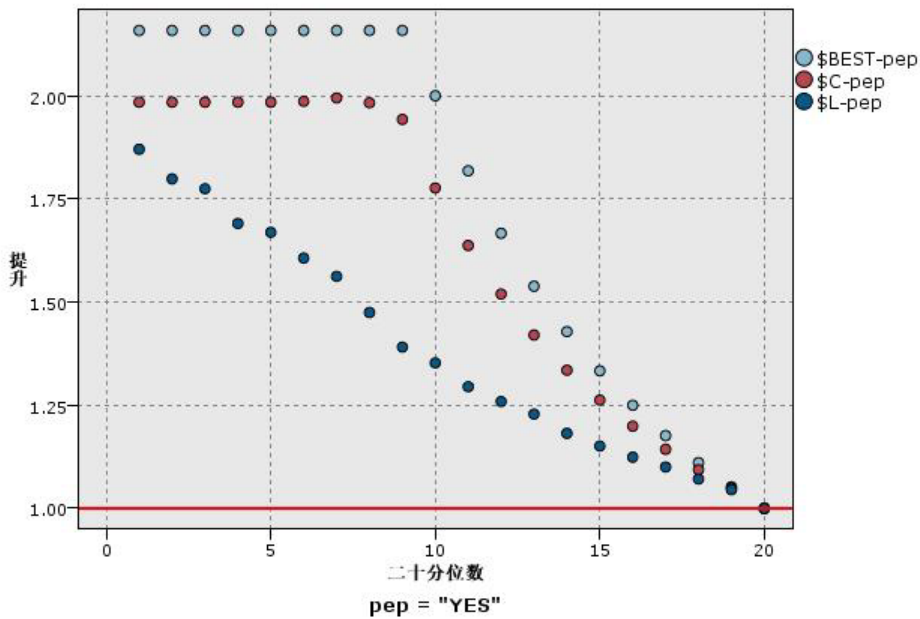


图 44. 使用点和最佳线的提升图（累积）

响应图

响应即分位数中，匹配记录的比例。其计算方式为（分位数中的匹配/分位数中的记录）× 100%。

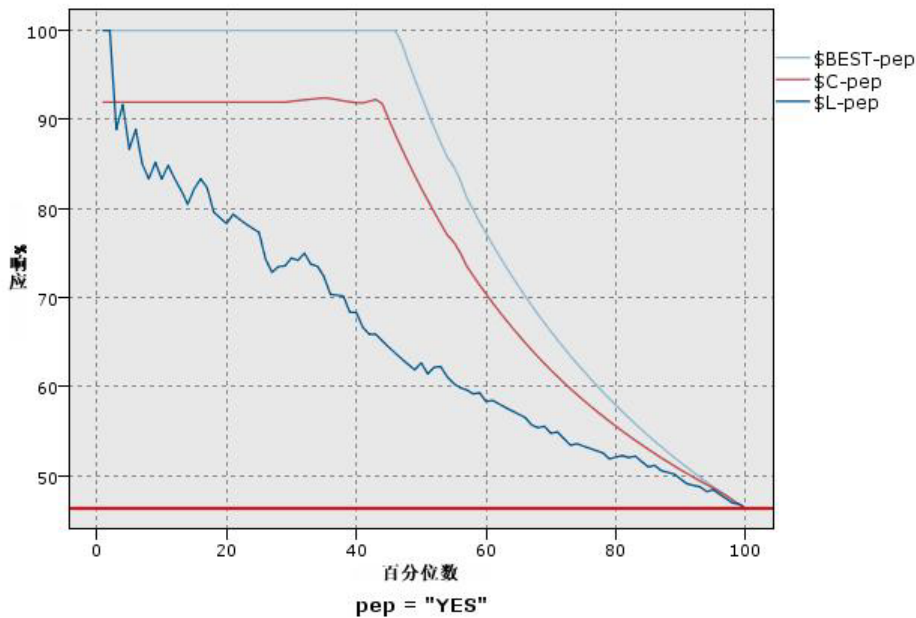


图 45. 具有最佳线的响应图（累积）

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）。

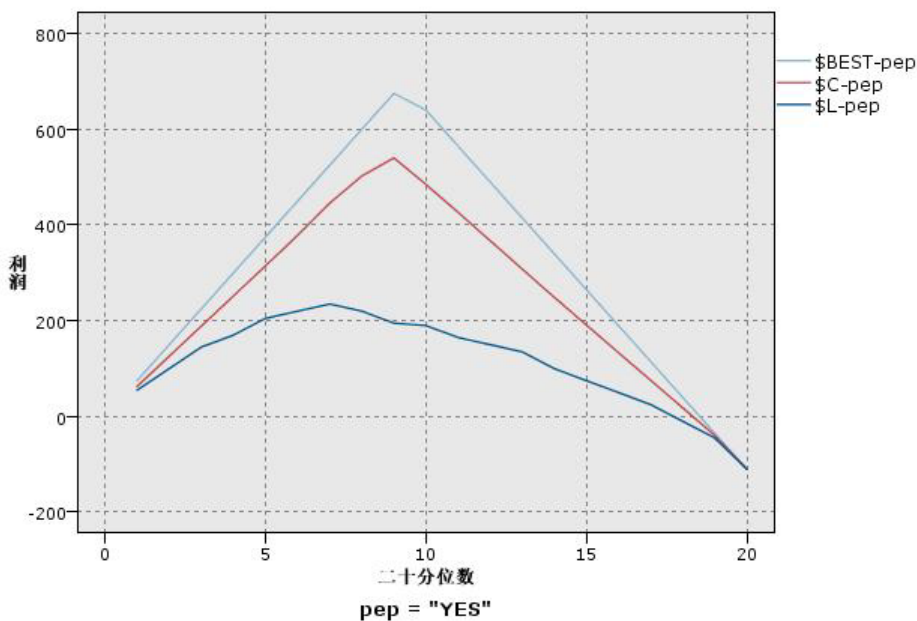


图 46. 具有最佳线的利润图（累积）

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%。

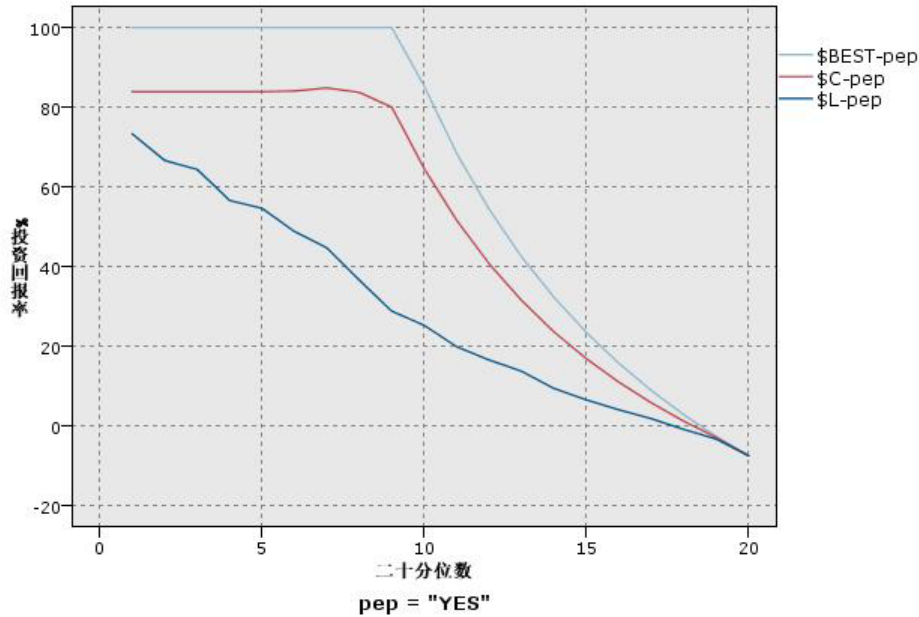


图 47. 具有最佳线的投资回报图 (累积)

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。

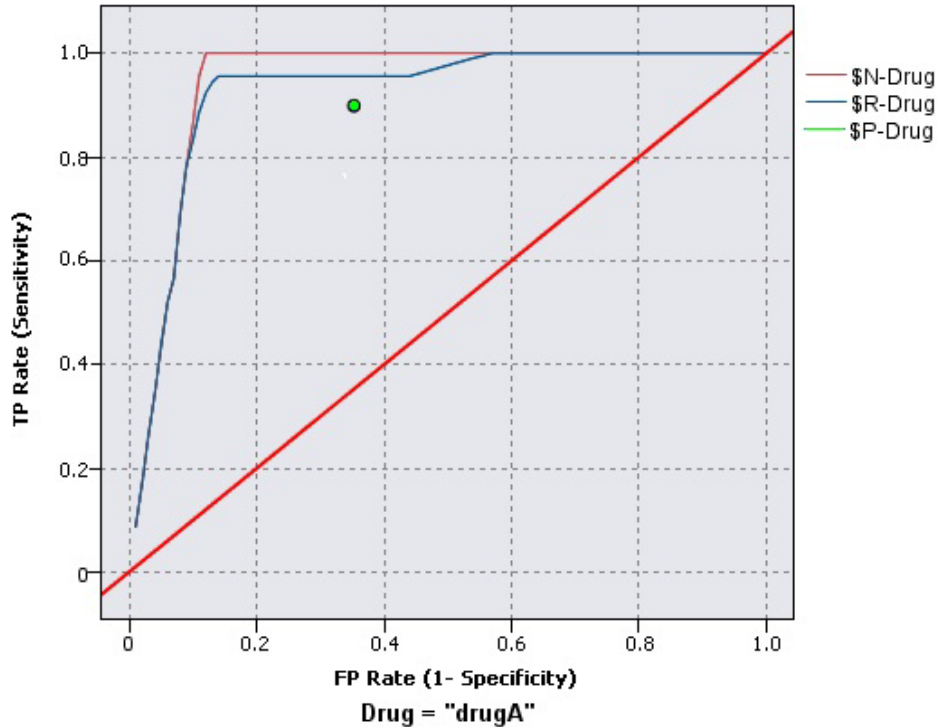


图 48. 具有最佳线的 ROC 图表

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。

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。）

将利润标准用于所有图表类型。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。）

查找预测字段/预测变量字段所使用的方法。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得分字段；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得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。请参阅主题第 121 页的『分区节点』，了解更多信息。

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"` 即是一个有效条件：凡是目标字段的值为 *Yes* 的记录，都将在评估中算作匹配。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。

用户定义的评分。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`，且相应的得分为 `1 - @PRE-DICTED`。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。

包括业务规则。选择此选项后，您可指定满足相关标准的规则条件。例如，您可能想要显示用于所有满足 `mortgage = "Y" and income >= 33000` 这一条件的观测值的规则。业务规则将在图表中绘制出来，在关键字字段中，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。

包括字段名称。选择此选项可使字段名称在输出文件的第一行显示出来。

每条记录后换新行。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。输入用于图形标题的文本。

子标题。输入用于图形子标题的文本。

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）。

效益图。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。

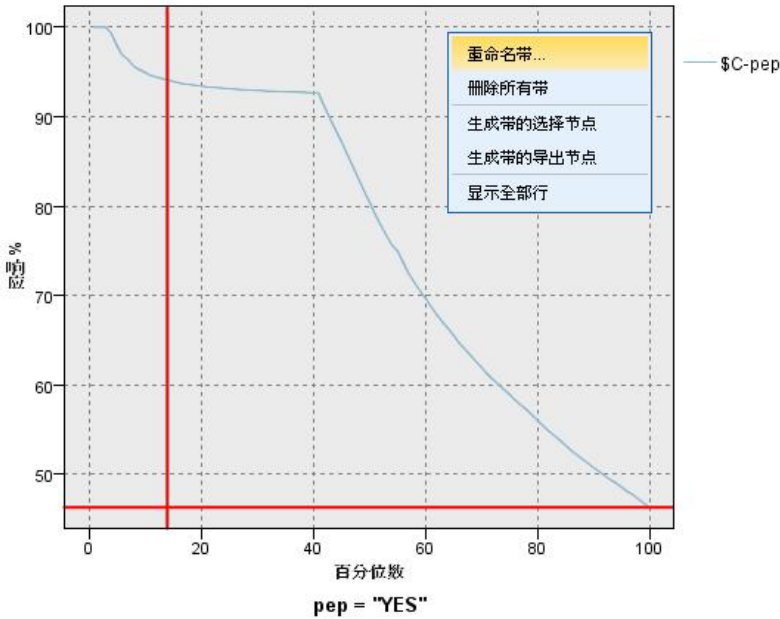


图 49. 处理评估图表

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。请参阅主题『探索图形』，了解更多信息。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。请参阅主题第 211 页的『从图形中生成节点』，了解更多信息。

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。

探索图形

使用编辑模式可以编辑图形的布局 and 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。要选择此模式，请从菜单中选择 **视图 > 探索模式**（或单击工具栏图标）。

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。请参阅主题第 204 页的『使用带状区域』，了解更多信息。
- 定义和编辑区域，这些区域用于标识矩形区域中的一组值。请参阅主题第 208 页的『使用区域』，了解更多信息。
- 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。请参阅主题第 210 页的『使用标记后的元素』，了解更多信息。
- 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。请参阅主题第 211 页的『从图形中生成节点』，了解更多信息。



图 50. 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。其中一些可以包含带状区域的图形包括：直方图、条形图和分布图、散点图（线、散点、时间等）、收集图和评估图。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。

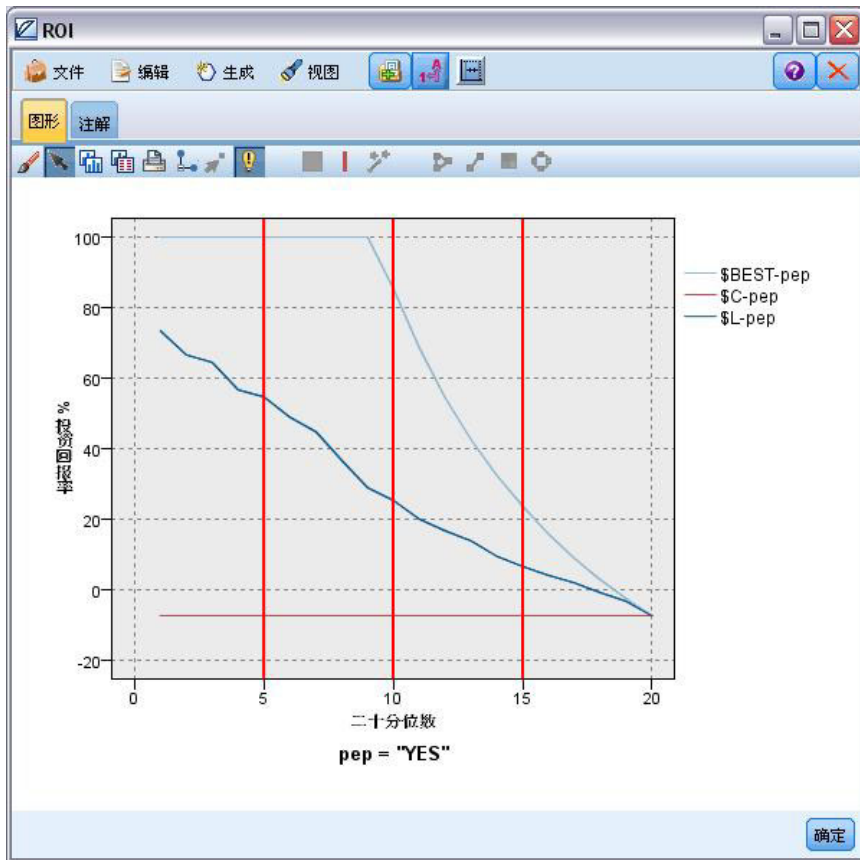


图 51. 包含三个带状区域的图形

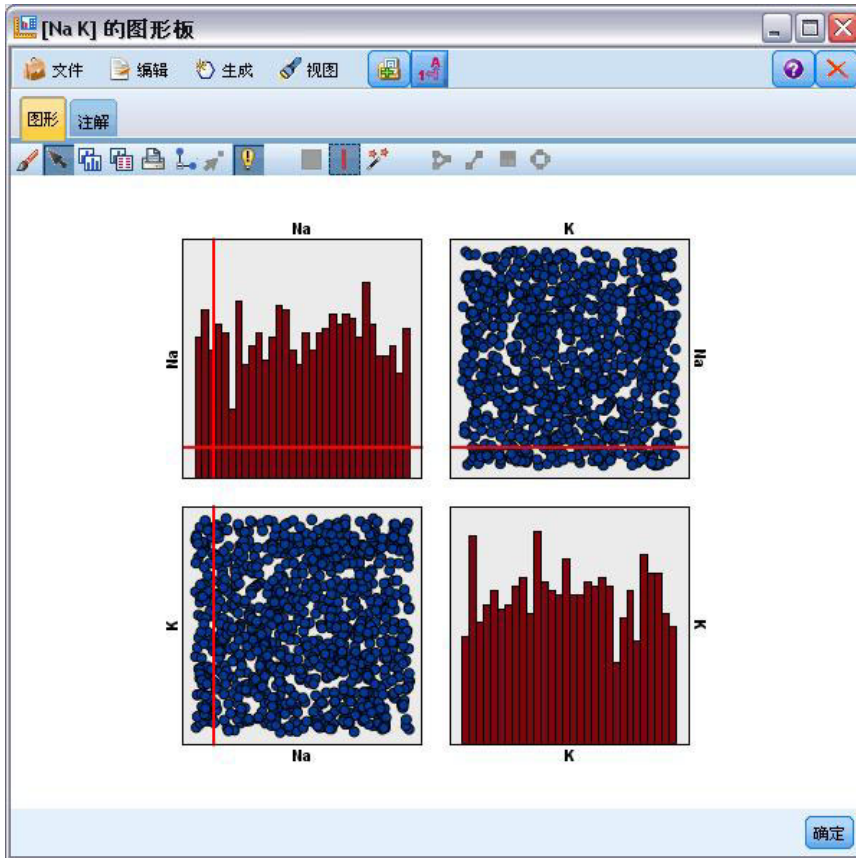


图 52. 包含带状区域的 SPLOM

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。默认情况下，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，通过拖动带状区域可在 *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 53. “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 54. 分割图标可扩展工具栏，提供用于分割为带状区域的选项

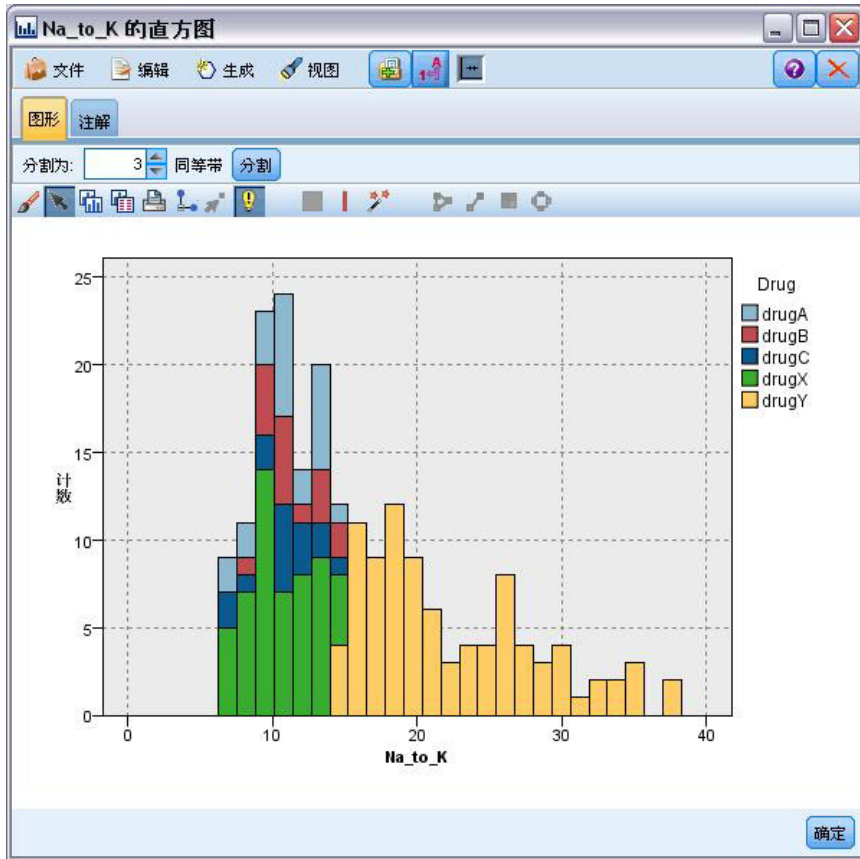


图 55. 创建启用带状区域的相等带状区域工具栏

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 56. “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** 键开始新行。
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。可使用区域的图形包括：散点图（线、散点、气泡、时间等）、SPLOM 图和集合图。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。

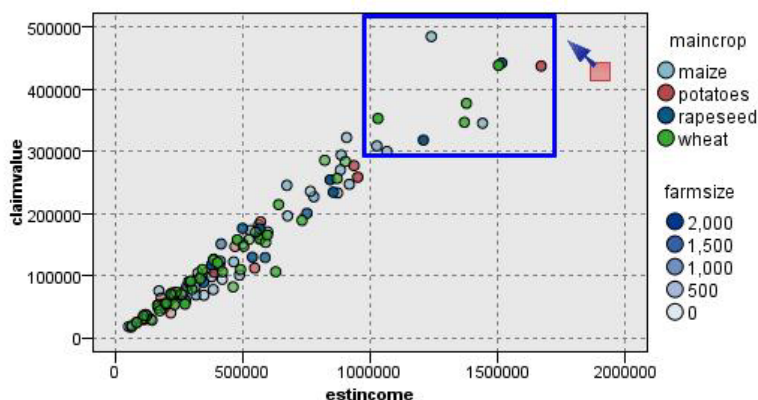


图 57. 定义具有高索赔值的区域

定义区域

无论在哪个位置上定义区域，都要对值进行分组。默认情况下，每个新区域都称为**区域<N>**，其中 **N** 表示创建的区域量。

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。请参阅主题第 211 页的『从图形中生成节点』，了解更多信息。

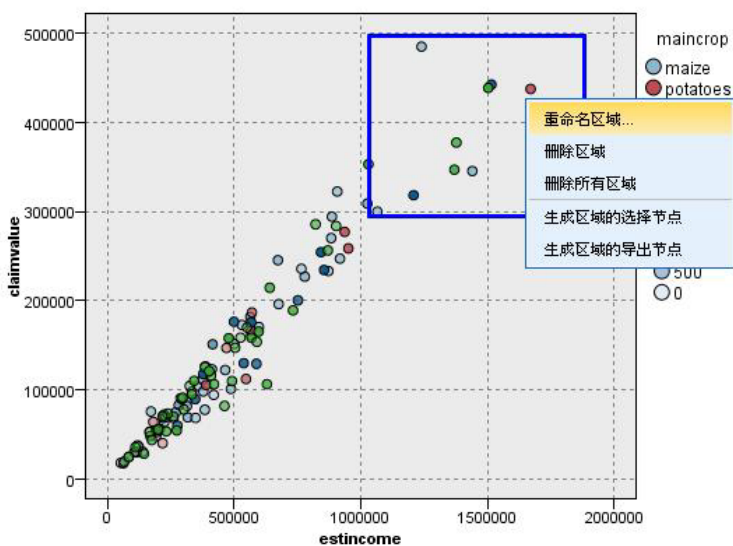


图 58. 探索具有高索赔值的区域

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 59. “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 60. 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** 键开始新行。
6. 通过调整 **A** 和 **B** 的 **Min** 值和 **Max** 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。

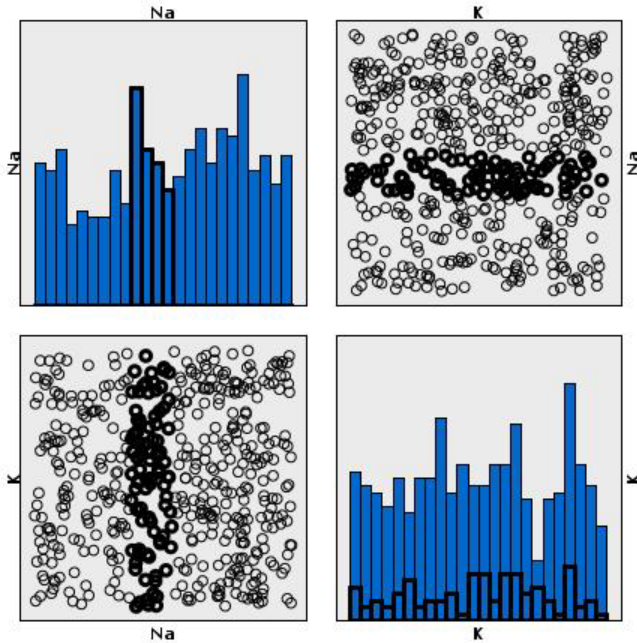


图 61. 对 SPLOM 中的元素进行标记

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。请参阅主题『从图形中生成节点』，了解更多信息。

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 62. 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。可以生成用于在区域中包含或排除记录的选择节点。
- 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。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，可以生成一个导出节点，该节点可以为轴上所标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默认区域同名。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**分组**生成重新分类节点, 以根据特定值在组 (在“表”选项卡中使用 **Ctrl** 键并同时单击可选择组) 中的包含关系对显示字段的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。

注意: 如果值已进行预定义, 那么可以将它们作为平面文件的形式读入 **IBM SPSS Modeler**, 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。

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。

编辑直观表示

探索模式允许您以分析的方式探索直观表示代表的数据和值, 而编辑模式允许您更改直观表示的布局 and 外观。例如, 可以更改字体和颜色以符合您组织的样式指南。要选择此模式, 请从菜单中选择 **视图 > 编辑模式** (或单击工具栏图标)。

在编辑模式中, 有几种工具栏会影响直观表示布局的不同方面。如果发现任何未使用的工具栏, 可以将其隐藏, 以增大显示图形的对话框的空间量。要选择工具栏或取消选择, 可单击“视图”菜单上相关的工具栏名称。

注意: 要将进一步详细信息添加到您的直观表示, 可以应用标题、脚注和轴标。请参阅主题第 224 页的『添加标题和脚注』, 了解更多信息。

在 **编辑模式** 中有几个编辑直观表示的选项。可以进行下列操作:

- 编辑文本并对其进行格式化。
- 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（例如条形和点）的大小。
- 使用边距和填充调整项目周围的空白。
- 指定数字格式。
- 更改轴和刻度设置。
- 在分类轴上排序、排除和折叠类别。
- 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。
- 更改图例的位置。
- 应用直观表示样式表。

下列主题说明了如何执行上述各种任务。同时建议您阅读关于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

编辑直观表示的一般规则

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:

视图 > 编辑方式

选择

可用于编辑操作的选项取决于选定的内容。根据不同的选定内容，会启用不同的工具栏和属性选用板选项。仅已启用的项目会应用于当前选定的内容。例如，如果已选定某个轴，则属性选项板中可用的选项卡有“尺度”、“主要核对项”和“次要核对项”。

以下是有关在直观表示中选择项目的一些提示:

- 单击一个项目以选定它。
- 单击可选择图形元素（例如散点图中的点和条形图中的条形）。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** 键可取消选定所有元素。

选项板

当在直观表示中选择了个项目时，各种选用板会进行更新以反映选择。选用板包含用于编辑选择的控件。选用板可以是工具栏或带有多个控件和选项卡的面板。选用板可以隐藏，以确保显示必要选用板用于编辑。查看“视图”菜单了解当前显示的选项板。

您可以通过单击并拖动工具栏选项板中或其他选项板左侧的空白空间更改选项板位置。可视反馈让您知道您可以将选用板放在哪里。对于非工具栏选用板，您还可以单击关闭按钮隐藏选用板，或单击取消放置按钮以在单独窗口中显示选用板。单击帮助按钮以显示特定选项板的帮助。

自动设置

某些设置提供 **-auto-** 选项。这表示可应用自动值。使用哪些自动设置取决于特定直观表示和数据值。可以输入一个值覆盖自动设置。如果需要恢复自动设置，可删除当前值并按 **Enter** 键。此设置将再次显示 **-auto-**。

删除/隐藏项目

您可以除去/隐藏直观表示中的各种项目。例如，可以隐藏图例或轴标签。要删除某个项目，可将其选中并按“删除”。如果不允许删除此项目，那么将不会删除任何内容。如果意外地删除了某个项目，则按 **Ctrl+Z** 可撤销删除操作。

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。

编辑和格式化文本

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，不能编辑勾号标签，因为此标签的内容源自原始数据。然而，您可以设置直观表示中任何文本的格式。

在适当的位置编辑文本的步骤

1. 双击该文本块。此操作会选定所有文本。此时禁用所有工具栏，因为您在编辑文本时无法更改直观表示的任何其他部分。
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。将游标放在所需位置并输入附加文本。

格式化文本的步骤

1. 选定包含文本的框。不要双击此文本。
2. 使用字体工具栏格式化文本。如果未启用此工具栏，请确保仅选定包含文本的框。如果已选定文本本身，则将禁用此工具栏。

可以对字体进行以下更改：

- 颜色(C)
- 族（例如，Arial 或 Verdana）
- 字号（单位为 pt，除非指定了其他单位，例如 pc）
- 宽度(W)
- 相对于文本框进行调整

可对框中的所有文本应用格式化操作。不能更改任何特定文本块中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。条形的颜色就是填充颜色。这些条形四周还会有黑色的实线边框。

在有填充颜色的直观表示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，请看轴标签中的文本。此文本看起来像是个“浮动”文本，但它实际上位于一个具有透明填充颜色的框中。通过选定轴标签可以看到此框。

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拆分，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、边框颜色或填充样式，可使用颜色工具栏。

注意：此工具栏不反映当前选定内容的状态。

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。在颜色中，注意有一种颜色的外观是白底上加红斜线。这是一种透明颜色。例如，可以使用此颜色隐藏直方图中条形的边框。

- 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关联，那么此按钮将更改与数据中的最高值关联的颜色的填充颜色。可以使用属性选用板上的“颜色”选项卡更改与最低值和缺失数据关联的颜色。随着基本数据值的增加，元素的颜色会逐渐由“低”色彩更改为“高”色彩。
 - 第二个按钮可控制边框颜色。
 - 第三个按钮可控制填充模式。填充模式使用边框颜色。因此，只有当边框颜色可见时填充样式才可见。
 - 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。百分比越低表示不透明度越低，透明度越高。100% 表示完全不透明。
3. 要更改边框或线的划线类型，可使用线工具栏。

注意：此工具栏不反映当前选定内容的状态。

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高宽比

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）。

修改点元素的步骤

1. 选定点元素。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。
 - 使用第一个按钮可以更改点的形状。单击下拉箭头并选择一个预定义的形状。
 - 使用第二个按钮可以将点旋转到特定的圆周位置。单击下拉箭头然后将指针拖动到所需位置。
 - 使用第三个按钮可以更改宽高比。单击下拉箭头然后单击并拖动弹出的矩形。矩形的形状表示宽高比。

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**。

更改图形元素的大小的步骤

1. 选定要更改其大小的图形元素。
2. 在使用位于符号工具栏上的选项时，可使用滑块或输入特定大小。此单位为像素，除非指定了其他单位（参阅以下单位缩写的完整列表）。还可以指定百分比（如 30%），表示图形元素将使用指定百分比的可用空白。可用空间取决于图形元素类型和特定直观表示。

表 46. 有效单元缩写

缩写	单位
cm	厘米
英寸	英寸
mm	毫米
pc	派卡
pt	点
px	像素

指定边距和填充

如果直观表示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** 指位于框及其周围的其他项目之间的空白量。**填充** 指位于框的边框和框的 **内容** 之间的空白量。

指定边距和填充的步骤

1. 选定要为其指定边距和填充的框。此框可以是文本框、图例周围的框，乃至显示图形元素（例如条形和点）的数据框。
2. 使用属性选用板上的“边距”选项卡指定这些设置。所有大小的单位都是像素，除非指定了其他单位（例如厘米或英寸）。

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勾号标签的数字格式。例如，您可以指定刻度勾号标签中显示的数字以千显示。

如何指定数字格式

1. 选择连续轴刻度勾号标签或数据值标签（如果包含数字）。
2. 单击“属性”选项板上的**格式**选项卡。
3. 选择所需数字格式选项:

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 (%)。

最小整数位数。 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，那么该值的整数部分将替换为星号。

最小小数位数。 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。

最大小数位数。 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，那么小数将四舍五入为适当的位数。

科学表示法。 是否以科学表示法显示数字。科学表示法对很大或很小的数字很有用。**-auto-** 使应用程序确定何时使用科学表示法比较适当。

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，假设您的刻度轴显示薪金，标签为 30,000、50,000 和 70,000。您可以输入一个 1000 的刻度因子，这样显示的就是 30、50 和 70。然后您就应该编辑刻度轴标题，使其包括“单位为千”这样的文本。

括号用于 -ve。 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。您计算机的当前区域设置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. 选定轴的任何部分（例如，轴标签或勾号标签）。
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。

“尺度”选项卡

注意：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。 指定刻度是线性的还是变换的。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。尺度变换还可用于使非对称的直方图更加对称以与正态分布类似。注意，此操作仅能够变换在其上显示数据的尺度；而不能变换实际的数据。

- **线性。** 指定线性非变换的刻度。
- **对数。** 指定底数为 10 的对数变换刻度。为适用于零和负数，此变换使用对数函数的修改版。此“安全对数”函数定义为 $\text{sign}(x) * \log(1 + \text{abs}(x))$ 。因此，`safeLog(-99)` 等于：

$$\text{sign}(-99) * \log(1 + \text{abs}(-99)) = -1 * \log(1 + 99) = -1 * 2 = -2$$

- **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 $\text{sign}(x) * \text{pow}(\text{abs}(x), 0.5)$ 。因此，`safePower(-100)` 等于：

$$\text{sign}(-100) * \text{pow}(\text{abs}(-100), 0.5) = -1 * \text{pow}(100, 0.5) = -1 * 10 = -10$$

最小值/最大值/略低值/略高值。 指定刻度的范围。选择 **略低值** 和 **略高值** 可使应用程序能够基于数据选择适当的尺度。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，那么尺度的略低值和略高值可能是 0 和 100，而不是实际的最小数据值和最大数据值。注意，不要将范围设置得过小而隐藏了重要的项目。还需注意，如果选择了 **包含零** 选项，则不能精确地设置最小值和最大值。

低边距/高边距。在轴的低端和/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，如果为垂直轴设置的 **高边距** 为 5，则沿数据框的顶部水平显示一个 5px 宽的边框。

反转。指定刻度是否反转。

包括 0。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。如果选中此选项，则将禁用 **最小值** 和 **最大值**，因为无法为尺度范围设置自定义的最小值和最大值。

主要核对项/次要核对项选项卡

刻度或刻度标记是在轴上显示的数据行。它们表示特定间隔或类别上的值。**主刻度标记**是带标签的刻度标记。这些记号也比其他勾号标记长。**次刻度标记**是出现在主刻度标记之间的刻度标记。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。**网关线**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。指定刻度标记的长度。此单位为像素，除非指定了其他单位（例如厘米或英寸）。

基准。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。

增量。仅适用于主刻度标记。指定主刻度标记之间的差值。即，每隔 n 个值显示一个主要核对项，其中 n 是 delta 值。

区域数。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。次刻度标记数为区域数减一。例如，假设在 0 和 100 的位置上是主要核对项。如果输入 2 作为次要核对项的区域数，则在 50 的位置上会有一个次要核对项，这个次要核对项将 0-100 的范围分成两个区域。

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。类别选项板显示轴上的类别。

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. 在类别选项板的下拉列表中选择排序选项：

定制。根据类别在选项板中的显示顺序排列类别。使用箭头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。使用显示在选用板中的名称，按字母顺序排列类别。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。

统计量。根据为每一个类别计算的统计量排列类别。包括计数、百分比和平均值的统计量示例。仅在统计量用于图形中时此选项可用。

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直观表示中。

1. 选择分类轴。类别选项板显示轴上的类别。

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”选项板中，单击添加类别按钮：



图 63. 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。类别选项板显示轴上的类别。

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

可以合并太小而无需单独显示的类别。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量。例如，您无法一起添加均值，因为均值不可加。因此，使用均值合并/拼并类别不可用。

1. 选择分类轴。类别选项板显示轴上的类别。

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. 在类别选项板中，选择 **折叠** 并指定百分比。将所有总数百分比低于指定值的类别合并为一个类别。百分比基于图表中显示的统计量。折叠功能仅能用于基于计数的统计量和总计统计量 (sum)。

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。

更改面板方向的步骤

1. 选择直观表示的任何部分。
2. 单击“属性”选项板上的**面板**选项卡。
3. 从 **布局** 中选择选项：

表。将面板布局成类似表的样子，在此布局中，会将行或列分配给每个单独的值。

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。将面板布局成类似列表的样子，在此布局中，每个单元格都代表值的组合。列和行不再分配给单独的值。此选项使面板能够在必要时换行。

转换坐标系统

许多直观表示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，并转置轴。如果已应用于当前直观表示，那么还可以取消任何这些转换。例如，在极坐标系中绘制饼图。您可以取消极转换，并在矩形坐标系统中将饼图显示为单个堆积条形图。

如何转换坐标系统

1. 选择要转换的坐标系。您可以通过选择单个图形周围的边框选择坐标系统。
2. 单击“属性”选项板上的**坐标**选项卡。
3. 选择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**。这类似于对调二维直观表示中的水平和垂直轴。

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。饼图是一维直观表示，其中极转换以特定角度绘制单个条形图。雷达图表是一个二维直观表示，其中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。三维直观表示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。

同一比例。应用同一比例指定每个刻度上的同一距离代表数据值中的同一差异。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之前，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 %。如果要更改图形的宽高比，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之后，缩进距离会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。

更改统计量和图形元素

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”选项板上的**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 47. 图形元素类型

图形元素类型	描述
点	这是标识特定数据点的标记。点元素在散点图及其他相关直观表示中使用。
区间	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。间隔元素在条形图和直方图中使用。
线(L)	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(A)	用线连接数据元素，并填充线与原点之间的区域。
多边形	多边形构成闭合数据区域。多边形元素在分级散点图或映射中使用。
模式	该元素由带有须和标记以表示离群值的箱体构成。模式元素用于箱图。

如何更改统计量

1. 选择要更改统计量的图形元素。
2. 单击“属性”选项板上的**元素**选项卡。
3. 从“摘要”下拉列表中选择新统计量。注意选择一个汇总数据的统计量。如果希望直观表示显示未汇总的数据，请从“摘要”列表中选择（**非统计量**）。

根据连续字段计算的汇总统计

- **平均值 (Mean)**. 集中趋势的测量值。算术平均值，等于总和除以观测值数。
- **中位数 (Median)**. 第 50 个百分位，大于该值和小于该值的个案数各占一半。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。
- **最小值 (Minimum)**. 数字变量的最小值。
- **最大值 (Maximum)**. 数字变量的最大值。
- **范围**. 最大值与最小值之间的差。
- **中程数值**. 范围中间值，即与最小值的差等于与最大值的差的值。
- **Sum**. 所有具有非缺失值的观测值的和或总计。
- **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 (Variance)**. 平均值离散度的测量值，等于均值的平方差除以观测值数减一的差。方差按单元计量，即变量自身单元数的平方。
- **标准差 (Standard Deviation)**. 均数离差的度量值。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)**. 因样本变化而变化的检验统计量值的个数的度量值。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)**. 观测值聚类围绕中心点的程度的一种测度。对于正态分布, 峰度统计量的值为零。正峰度值表示相对于正态分布, 观察值在分布中心的聚集更多, 同时尾部更薄, 直到分布极值。在这一点, 尖峰态分布的尾部比正态分布的尾部要厚。负峰度值表示相对于正态分布, 观察值聚集得少并且尾部较厚, 直到分布极值。在这一点, 低峰态分布的尾部比正态分布的尾部要薄。
- **偏度 (Skewness)**. 分布的不对称度量值。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, 区域统计会产生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% 机会包含给定单个观测值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些列值

基于计数的汇总统计

- **计数**。行/观测值数量。
- **累积计数**。行/观测值的累积数目。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/观测值数量对比行/观测值的总数的百分比。
- **计数累积百分比**。每个子组中行/观测值数目对比行/观测值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。

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”选项板上的**元素**选项卡。
3. 从“修饰符”下拉列表中选择冲突修饰符。**-自动-** 让应用程序确定哪个冲突修饰符适合该图形元素类型和统计量。

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。

更改图例位置的步骤

1. 选择图例。
2. 单击“属性”选项板上的图注选项卡。
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 64. 复制直观表示按钮

复制直观表示。此操作会将直观表示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 65. 复制直观表示数据按钮

复制直观表示数据。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。

键盘快捷键

表 48. 键盘快捷键

快捷键	函数
Ctrl+Space	在“探索”和“编辑”模式之间切换
删除	删除一个直观表示项目
Ctrl+Z	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。

添加图形的标题

1. 从菜单中选择 **编辑 > 添加图形标题**。含有 **<标题>** 的文本框将显示在图形上方。
2. 确保处于编辑模式。从菜单选择 **视图 > 编辑方式**。
3. 双击 **<标题>** 文本。

4. 输入需要的标题并按“返回”按钮。

添加图形的脚注

1. 从菜单中选择 **编辑 > 添加图形脚注**。含有 **<脚注>** 的文本框将显示在图形下方。
2. 确保处于编辑模式。从菜单选择 **视图 > 编辑方式**。
3. 双击 **<脚注>** 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 提供了一个默认样式表；不过可以根据需要修改该样式表。例如，可以在图形中使用公司惯用的图示颜色。请参阅主题第 213 页的『编辑直观表示』，了解更多信息。

在图形节点中，可使用编辑模式来更改图形外观的样式。然后使用 **编辑 > 样式** 菜单将更改保存为样式表，以便应用于后来通过当前图形节点生成的所有图形，或者用作使用 IBM SPSS Modeler 制作的所有图形的新默认样式表。

“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。请参阅主题『应用样式表』，了解更多信息。
- **将样式存储在节点中**。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。
- **将样式另存为缺省值**。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

您可以应用规定直观表示样式属性的直观表示样式表。例如，样式表可定义其他选项间的字体、划线和颜色。样式表在某种程度上提供您需要手动进行的编辑工作的快捷方式。但请注意，样式表仅限于样式更改。其他更改，如图形位置或尺度范围不保存在样式表中。

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”对话框选择样式表。
3. 单击**应用**以应用样式表到直观表示而不关闭对话框。单击**确定**以应用样式表并关闭对话框。

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。有些样式表是预先安装的，其他样式表可能是在 IBM SPSS Visualization Designer（另外一个产品）中创建的。

对话框底部显示具有样本数据的示例直观表示。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，样式表可覆盖直观表示中的所有样式。您可以更改这种行为。

- **覆盖所有样式。** 应用样式表时，覆盖直观表示中的所有样式，包括在当前编辑会话期间在直观表示中修改的样式。
- **保持修改样式。** 应用样式表时，只覆盖在当前编辑会话中未在直观表示中修改的样式。保留在当前编辑会话期间修改的样式。

管理。 管理计算机上的可视化模板、样式表和映射。可以在本地计算机上导入、导出、重命名和删除可视化模板、样式表和映射。请参阅主题第 163 页的『管理模板、样式表和地图文件』，了解更多信息。

位置。 更改可视化模板、样式表和映射的存储位置。当前位置列出在按钮右侧。请参阅主题第 162 页的『设置模板、样式表和地图的位置』，了解更多信息。

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，可以从“编辑”菜单选择复制其中的图形或数据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。有关配置页眉和页脚的详细信息，请参阅。

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

要将图形保存在存储库中，请从菜单中选择**文件 > 保存输出**。

复制图形

要复制图形，以用于其他应用程序（比如 MS Word 或 MS PowerPoint），请从菜单中选择 **编辑 > 复制图形**。

复制数据

要复制数据，以用于其他应用程序（比如 MS Excel 或 MS Word），请从菜单中选择 **编辑 > 复制数据**。缺省情况下，将数据格式化为 HTML 文件。粘贴时，使用另一个应用程序上的 **选择性粘贴** 来查看其他格式化选项。

导出图形

使用**导出图形**选项，可以通过下列其中一种格式导出图形：Bitmap (.bmp)、JPEG (.jpg)、PNG (.png)、HTML (.html) 或用于其他 IBM SPSS Statistics 应用程序的 ViZml 文档 (.xml)。

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**选项可以通过下列其中一种格式导出表：制表符分隔 (.tab)、逗号分隔 (.csv) 或 HTML (.html)。

要导出表格，请从菜单中选择 **文件 > 导出表格**，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”节点以表格形式显示数据，这些数据还可以写入到文件中。每当您需要检查数据值或将其导出为可轻松读取的形式时，此节点非常有用。



“矩阵”节点将创建一个字段关系表。此节点最常用于显示两个符号字段间的关系，但也可用于显示标志字段或数字字段间的关系。



“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。“分析”节点也可以对比各个预测模型。



“数据审核”节点将首先全面检查数据，这些数据包括每个字段的汇总统计量、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，该矩阵可以排序并且可以用于生成完整大小的图表和数据准备节点。



通过变换节点可首先选择和以可视方式预览变换结果，然后再将其应用于选择的字段。



“统计量”节点可提供有关数值字段的基本汇总信息。它可计算单个字段以及字段间的相关性的汇总统计量。



“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。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。通过使用模板中的 CLEM 表达式，可以包括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，可以用该节点为一个名为年龄的字段计算统计量并通过插入函数 @GLOBAL_MEAN(age) 在 CLEM 表达式中使用年龄的总均值。



“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**。请参阅主题第 230 页的『导出输出』，了解更多信息。

- 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 Collaboration and Deployment Services 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。请参阅主题第 231 页的『选择单元格和列』，了解更多信息。

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 Collaboration and Deployment Services 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 Internet 访问和 IBM SPSS Collaboration and Deployment Services 帐户进行查看，而无需安装 IBM SPSS Modeler。

注意：访问 IBM SPSS Collaboration and Deployment Services 存储库需要单独许可证。有关更多信息，请参阅<http://www.ibm.com/software/analytics/spss/products/deployment/cds/>。

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。

其他类型的输出只能在用户计算机上安装了相关应用程序（例如，对于流对象，需要 IBM SPSS Modeler）时才能进行查看。

表 49. 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
	Statistics
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 Collaboration and Deployment Services 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。

注意: 访问 IBM SPSS Collaboration and Deployment Services 需要单独许可证。有关更多信息, 请参阅 <http://www.ibm.com/software/analytics/spss/products/deployment/cds/>。

要在网络上查看发布的输出:

1. 将您的浏览器指向 `http://<repos_host>:<repos_port>/peb`

其中 `repos_host` 和 `repos_port` 为 IBM SPSS Collaboration and Deployment Services 主机的主机名和端口号。

2. 输入您的 IBM SPSS Collaboration and Deployment Services 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。该信息是 HTML 形式的输出, 因而您可以保存它并在别处 (如公司内部网或 Internet 站点) 重复使用。

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。

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。

对列进行重新排序。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。

表节点的“设置”选项卡

突出显示符合以下条件的记录。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。请参阅主题第 100 页的『字段格式设置选项卡』，了解更多信息。

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**自定义**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。

数据以系统缺省编码格式输出，编码格式可在 Windows 控制面板中指定，如果以分布式模式运行，那么将在服务器计算机上指定。

- **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“平均值”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。

注意：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“平均值”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视图间切换。请参阅主题第 248 页的『平均值节点输出浏览器』，了解更多信息。

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。请参阅主题第 250 页的『报告节点的模板选项卡』，了解更多信息。建议不要在使用**定制格式**时时有此选项。

每页的行数。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。

注意：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。请参阅主题第 274 页的『平面文件导出节点』，了解更多信息。

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。请参阅主题第 231 页的『选择单元格和列』，了解更多信息。这些操作与预览节点中数据的操作相同。

导出表格数据。要从表格浏览器导出数据，请选择：

文件 > 导出

请参阅主题第 230 页的『导出输出』，了解更多信息。

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。

搜索表格。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。
- **所有标志字段（真值）。**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。

包括缺失值。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，或选择交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。请参阅主题第 234 页的『矩阵节点的输出浏览器』，了解更多信息。
- **函数。**如果选择汇总函数，那么单元格值是所选交叠字段值的函数（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**，交叠字段为**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）。默认汇总函数是 **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。默认值为**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。

交叠。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。

- **突出显示前几个值。**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的数目。
- **突出显示后几个值。**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：

$$p(\text{row value}) * p(\text{column value}) * \text{total number of records}$$

- **残差。**观测值和期望值之间的差值。
- **行百分比。**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。
- **合计百分比。**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。
- **包括行和列的合计。**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。

矩阵节点的输出浏览器

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。请参阅主题第 228 页的『查看输出』，了解更多信息。

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。

- **Select 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

使用 Analysis 节点，可以对模型生成准确预测的能力进行评估。Analysis 节点执行一个或多个模型块的预测值和实际值（您的目标字段）之间的各种比较。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 Analysis 节点时，对于在执行的流中的每个模型块，都会自动将分析结果汇总添加到“汇总”选项卡上的“分析”部分中。详细分析结果显示在管理器窗口的“输出”选项卡上，或者，也可将其直接写至文件中。

注意：因为 Analysis 节点将预测值与实际值进行比较，所以它们只适用于监督式模型（需要目标字段的模型）。对于非监督式模型（如聚类算法），没有实际结果可以用作比较基础。

分析节点的分析选项卡

使用“分析”选项卡，可指定分析的详细信息。

符合矩阵（适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。将会显示一个表格，其中包含实际值定义的行和预测值定义的列，以及每个单元格中符合该模式的记录数。这适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，但这些字段是由不同的模型生成的，那么将为这些字段相同和不相同的情况进行计数并显示合计值。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）的平均信息内容进行测量的测量标准。考虑到不同类别在分类问题中的难度不一，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。如果对于某类别，模型并不比随机猜测效果好，则该类别的性能评估指数将为 0。

评估度量（仅二元分类器 AUC 和 Gini）。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 系数的按照 ROC 曲线与对角线之间面积的两倍进行计算，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，此选项报告有关置信度值及其与预测的关系的统计量。此选项有两个设置：

- **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。**报告按指定系数提高的准确性的置信度级别。例如，如果总准确性为 90%，而此选项设置为 2.0，则报告的值将是准确性为 95% 时所需的置信度。

查找预测字段/预测变量字段所使用的方法。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，从而在即使重命名预测字段的情况下，也可以进行匹配。使用 **Type** 节点，从“值”对话框也可以访问任何预测字段的模型字段信息。请参阅主题第 96 页的『使用值对话框』，了解更多信息。
- **字段名称格式。**基于命名规则匹配字段。例如，C5.0 模型块为名为 *response* 目标生成的预测值必须位于名为 *\$C-response* 的字段中。

按分区分割。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。请参阅主题第 121 页的『分区节点』，了解更多信息。

注意：当按分区分割时，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。您可以指定自己的分析计算以在评估模型时使用。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得分合并为总得分。使用函数 **@TARGET** 和 **@PREDICTED** 可分别指定目标（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。**指定条件为真时的计算。
- **Else。**指定条件为假时的计算。
- **Use。**选择用于根据单个分值计算总分值的统计量。

按字段分解分析。显示适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 **Analysis** 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。请参阅主题第 228 页的『查看输出』，了解更多信息。

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。对于具有由生成的模型创建的相应预测字段的每个输出字段，分析输出都会包含一个相应的部分。

比较。在输出字段部分中，这是与该输出字段相关联的每个预测字段的子部分。对于分类输出字段，此部分的顶层包含一个表格，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，此节显示以下信息：

- **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。
- **均数误差。**显示所有记录的误差的平均值（均数）。这指出模型中是否有系统**偏差**（过高估计的趋势强于过低估计的趋势，或反之）。

- **平均绝对误差。**显示所有记录的误差绝对值的平均值。指出误差的平均大小（不考虑方向）。
- **标准差。**显示误差的标准差。
- **线性相关。**显示预测值和实际值之间的线性相关。此统计量在 -1.0 和 1.0 之间变化。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，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，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。注意：此处的空白条目表示由于实际或预测值为常量，在此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，列表示预测值。表格中的单元格表示预测值和实际值的每个组合的记录数。

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。每个输出类别均与它的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，则此处会显示这些值。对于模型置信度值，会报告下列统计量：

- **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）。
- **正确分类的平均置信度。**显示正确分类的记录的平均置信度。
- **未正确分类的平均置信度。**显示未正确分类的记录的平均置信度。
- **始终正确的置信度下界。**显示预测始终正确的置信度的下界，并显示符合此条件的案例的百分比。
- **始终不正确的置信度上界。**显示预测始终错误的置信度的上界，并显示符合此条件的案例的百分比。
- **准确度为 X%。**显示准确度为 X% 时的置信度级别。X 是分析选项中指定的阈值的近似值。对于某些模型和数据集，不存在能给出精确阈值（在选项中指定的）的置信度值（通常是对具有相同的、接近阈值的置信度值的类似案例进行聚类所致）。所报告的阈值是最接近于指定准确度条件的值，使用单个置信度值阈值，可以获得指定的准确性条件。
- **比整体准确性好 X 倍。**显示比整个数据集的准确性好 X 倍时，其相应的置信度值。X 等于在分析选项中指定的改善准确性值。

之间的一致性。如果流中包括两个或更多个对相同输出字段进行预测的已生成模型，则您还将查看与模型生成的预测之间的一致性相关的统计量。这包括预测相同的记录数和百分比（对于分类输出字段）或误差汇总统计量数和百分比（对于连续输出字段）。对于分类字段，它包括对模型相同（生成相同的预测值）的记录子集的预测与实际值进行的比较分析。

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。

筛选数据或对数据抽样。因为初始审核在处理“大数据”时特别有效，所以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**以根据默认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“交叠”字段。如果指定了多个目标字段，则不指定默认交叠。

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。

交叠字段。交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，则使用它作为默认交叠字段，如上所述。或者，您可以选择**使用自定义字段**以指定交叠。

显示。用于指定输出中是否显示图形以及选择默认显示的统计量。

- **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。请参阅主题第 239 页的『数据审核输出浏览器』，了解更多信息。
- **基本/高级统计量。**指定默认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。请参阅主题第 239 页的『显示统计量』，了解更多信息。

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。请参阅主题第 242 页的『过滤含缺失数据的字段』，了解更多信息。

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

- **含有效值的记录的计数。**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。
- **含无效值的记录的分类计数。**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:

与平均值的标准差。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。请参阅主题第 228 页的『查看输出』,了解更多信息。
- 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。请参阅主题『显示统计量』,了解更多信息。
- 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。请参阅主题第 19 页的『设置字段存储类型和格式』,了解更多信息。

查看和生成图形

如果未选择交叠,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交叠,将根据交叠的值为图形着色。

对于连续字段交叠,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 轴映射到交叠字段,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**。最终节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。
- 如果交叠集具有 100 个以上的值,则会发出警告,并且不会包括交叠。

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。请参阅主题第 238 页的『数据审核节点的设置选项卡』,了解更多信息。

最小值 (Minimum). 数字变量的最小值。

最大值 (Maximum). 数字变量的最大值。

Sum. 所有具有非缺失值的观测值的和或总计。

范围 (Range). 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得出的值。

平均值 (Mean). 集中趋势的测量值。算术平均值，等于总和除以观测值数。

平均值标准误差 (Standard Error of Mean). 来自同一分布的各样本之间均值差异的测量值。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 (standard deviation). 均数离差的度量值，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 (Variance). 平均值离散度的测量值，等于均值的平方差除以观测值数减一的差。方差按单元计量，即变量自身单元数的平方。

偏度 (Skewness). 分布的不对称度量值。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). 偏态与其标准误差的比率可用作正态检验（即，如果该比率小于 -2 或大于 +2，则可以拒绝正态）。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). 观测值聚类围绕中心点的程度的一种测度。对于正态分布，峰度统计量的值为零。正峰度值表示相对于正态分布，观察值在分布中心的聚集更多，同时尾部更薄，直到分布极值。在这一点，尖峰态分布的尾部比正态分布的尾部要厚。负峰度值表示相对于正态分布，观察值聚集得少并且尾部较厚，直到分布极值。在这一点，低峰态分布的尾部比正态分布的尾部要薄。

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检验（即如果比率小于 -2 或大于 +2，则可以拒绝正态）。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）。

唯一 (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 (Valid). 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。

中位数 (Median). 第 50 个百分位，大于该值和小于该值的个案数各占一半。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。请参阅主题第 238 页的『数据审核节点的设置选项卡』，了解更多信息。

交叠的统计量

如果连续（数值范围）交叠字段正在使用，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：

- **空值或系统缺失值。**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。系统缺失值显示为 **\$null\$**。请注意，空字符串在 IBM SPSS Modeler 中不被视为空值，但它们可能会被某些数据库视为空值。
- **空字符串和空格。**空字符串值和空白（带有不可见字符的字符串）不同于空值。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，那么此设置也应用于空字符串。
- **空白值或用户定义的缺失值。**这些是在源节点或“类型”节点中明确定义为缺失值的值，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 函数将这些值以及其他类型的缺失值处理为空白值。

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：

固定。替换为固定值（可以是字段平均值、范围中间值，或者您指定的常数）。

随机。根据正态分布或均匀分布替换为随机值。

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。根据 C&RT 算法替换为模型预测的值。对于使用此方法填补的每个字段，都会有一个单独的 C&RT 模型，还有一个“填充”节点，此节点使用该模型预测的值替换空白值和空值。然后使用“过滤”节点除去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，默认情况，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。请参阅主题第 238 页的『数据审核的质量选项卡』，了解更多信息。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内的最高值或最低值。
- **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。
- **质量百分比超过以下值的字段。**“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。默认阈值为 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. 要确保将所有字段（包括空字段或无类型字段）包括在审核中，可在上游的源节点或“类型”节点中单击**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。
- **质量百分比超过以下值的字段。**选择节点将检查完整记录的百分比大于指定阈值的字段。默认阈值为 50%。

在以下位置发现无效值时将记录视为无效。指定将记录确定为无效的条件。

- **任何上述字段。**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。
- **所有上述字段。**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。请参阅主题第 245 页的『生成图形』，了解更多信息。

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，此偏移与 IBM SPSS Modeler 中的 @OFFSET 序列函数执行的偏移无关。）

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。

- 取倒数 ($1/x$)。指出逆变换应显示在输出中。
- 取对数 ($\log n$)。指出 \log_n 变换应显示在输出中。
- 取对数 ($\log 10$)。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，可以指定输出的格式和位置。您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。请参阅主题第 231 页的『输出节点的“输出”选项卡』，了解更多信息。

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。请参阅主题第 228 页的『查看输出』，了解更多信息。

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：

平均值（标准差）

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 50. “派生”节点生成示例。

字段	转换
AGE	当前分布
INCOME	对数
OPEN_BAL	逆模型
BALANCE	逆模型

超节点中包含下列节点：

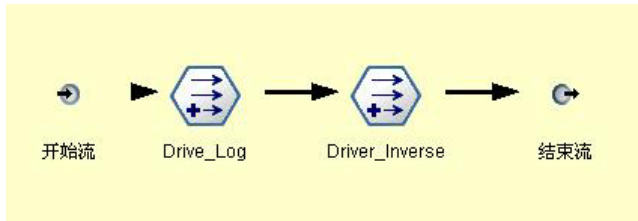


图 66. 工作区上的超节点

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：

$$(\log(\text{AGE}) - \text{Mean}) / \text{SD}$$

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

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。请参阅主题第 230 页的『导出输出』，了解更多信息。

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。

统计信息。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。默认情况下，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。默认情况下，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。默认情况下，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。请参阅主题第 228 页的『查看输出』，了解更多信息。

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。
- **合计**。对所有记录的该字段值求和。
- **最小值**。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：求出每个值和总平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**。一种对字段平均值估计值的不确定性（如果该平均值应用到新数据）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来的值。
- **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。请参阅主题第 246 页的『相关设置』，了解更多信息。

“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。

包括/排除显示在以下位置的字段。定义用于选择相关的条件。

- **相关总数**。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。
- **最大相关值的百分比**。选择指定百分比 ($n\%$) 的相关并包括/排除显示在这些相关中出现的所有字段。
- **大于该值的相关**。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：

- **在字段中的组之间**。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 检验本质上一致。请参阅主题第 248 页的『比较独立组的平均值』，了解更多信息。

- **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。请参阅主题『在成对字段之间比较平均值』，了解更多信息。

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可区分值，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。

相关性设置。使您可以指定用于标注相关性强度的选项。请参阅主题第 246 页的『相关设置』，了解更多信息。

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、边界或不重要的阈值 p 。您也可以编辑每个排序的标签。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。默认标签为**重要、边界和不重要**。

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。请参阅主题第 228 页的『查看输出』，了解更多信息。

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。
- **按组分类的平均值。**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。请参阅主题第 248 页的『平均值节点选项』，了解更多信息。

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。
- **df。**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。

- **字段一/二。**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。
- **平均值一/二。**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。请参阅主题第 246 页的『相关设置』，了解更多信息。
- **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。请参阅主题第 248 页的『平均值节点选项』，了解更多信息。

高级输出

高级输出增加了以下各列：

95% 置信度区间。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。

df。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：

固定。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

这是我的报告，打印在 [@TODAY] 上

将一个行复制到报告中，包含文本和当前日期。

全局（迭代 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

```
#ALLFor 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（迭代 WHERE）。对于满足指定条件的每条记录，会将包含在特殊标记 #WHERE <condition> 和 # 之间的行向报告复制一次。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。请参阅主题第 231 页的『输出节点的“输出”选项卡』，了解更多信息。

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。请参阅主题第 228 页的『查看输出』，了解更多信息。

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来计算名为年龄的字段的统计量，然后通过插入函数 @GLOBAL_MEAN(age) 在 CLEM 表达式中使用年龄的总平均值。

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。
- **MIN。**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**按钮将默认运算应用到列表中的所有字段。

执行前清除所有全局量。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。

执行后显示已创建全局量的预览。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。要将分布拟合到数据，拟合组件需要 2000 个有效值。对于数据较为稀疏的情况，您应该先除去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，只有在已除去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。请参阅主题第 35 页的『“模拟生成”节点』，了解更多信息。

“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。

要除去“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**除去链接**。

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 43 页的『分布』中描述了可用于拟合的分布。对理论分布的参数进行了调整，以便根据拟合度的度量实现与数据的最佳拟合；Anderson-Darling 标准或 Kolmogorov-Smirnov 标准。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算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：

- 使用上游节点除去包含缺失值的记录。
- 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 51. 根据字段的存储类型和测量级别进行的分布拟合

存储类型	测量级别					
	连续	分类	标记	名义	有序	无类型
字符串	不可能	对分类分布、骰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						
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。

- **要抽样的观测值数。**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。
- **拟合度标准（仅限连续字段）。**对于连续字段，将分布拟合到字段时选择拟合度 Anderson-Darling 检验或 Kolmogorov-Smirnoff 检验来对分布进行排序。缺省情况下将选择 Anderson-Darling 检验，尤其是在您希望确保尾部区域中实现可能的最佳拟合时，建议使用此检验。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。
- **分级（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。请参阅主题第 42 页的『迭代』，了解更多信息。

“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。

选择项。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。

已按迭代对输入数据进行排序。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。

- **均数。**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 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”节点的输出

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。请参阅主题第 228 页的『查看输出』，了解更多信息。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、导出或打印带有参考线的图表。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 52. 导航面板缩略图

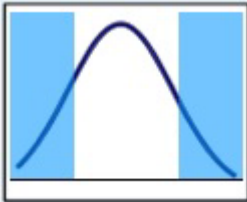
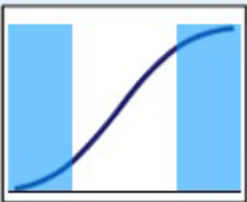
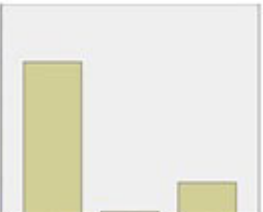
缩略图	描述	注释
	概率密度函数	只有在目标字段的测量级别为连续，并且在“模拟评估”节点对话框的“密度函数”视图中选择了 概率密度函数 (PDF) 的情况下，才会显示此缩略图。 如果目标字段的测量级别为分类，那么将不会显示此缩略图。
	累积分布函数	只有在目标字段的测量级别为连续，并且在“模拟评估”节点对话框的“密度函数”视图中选择了 累积分布函数 (CDF) 的情况下，才会显示此缩略图。 如果目标字段的测量级别为分类，那么将不会显示此缩略图。
	预测类别值	只有在目标字段的测量级别为分类，在“模拟评估”节点对话框的“密度函数”视图中选择了 概率密度函数 (PDF) ，以及在 要报告的类别值 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。

表 52. 导航面板缩略图 (续)

缩略图	描述	注释
	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。
	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。请参阅主题『图表选项』，了解更多信息。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。请参阅主题『图表选项』，了解更多信息。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。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。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系列。

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**来除去左参考线，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**来除去右参考线，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。您不能同时除去这两条参考线；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。

- **均数。**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，您可以除去定制参考线。

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**来除去参考线。

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。

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 Statistics 变换、Statistics 模型、Statistics 输出或 Statistics 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 交互。输入直接在“Statistics 导出”节点生成的数据文件上启动 IBM SPSS Statistics 时要使用的命令的完整路径和名称（例如，`C:\Program Files\IBM\SPSS\Statistics\<nn>\stats.exe`）。请参阅主题第 290 页的『Statistics 导出节点』，了解更多信息。

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。默认设置为 **本地**。

端口。为 IBM SPSS Statistics Server 指定服务器端口。

IBM SPSS Statistics 位置实用程序。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。单击该按钮以指定希望使用其许可的本地 IBM SPSS Statistics 安装的位置。
- 此外，如果是对远程 IBM SPSS Modeler Server 运行分布式方式，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。在 Windows 系统中，要完成此操作，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：

```
statisticsutility -location=<IBM SPSS Statistics_installation_path>/
```

或者，在 UNIX 系统中运行以下命令：

```
./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。

注释(O)

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。请参阅主题第 291 页的『重命名或过滤 IBM SPSS Statistics 的字段』，了解更多信息。
- 如果在 IBM SPSS Modeler 之后安装了 IBM SPSS Statistics，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”节点将数据写入与 ODBC 兼容的关系数据源。要将数据写入 ODBC 数据源，数据源必须存在且您必须对其具有写许可权。



平面文件导出节点将数据输出到已分隔的文本文件。这对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。IBM SPSS Statistics Base 和其他产品可以读取 *.sav* 或 *.zsav* 文件。这种格式也用于 IBM SPSS Modeler 中的某些缓存文件。



IBM SPSS Data Collection 导出节点以 IBM SPSS Data Collection 市场调查软件使用的格式输出数据。必须安装 IBM SPSS Data Collection 数据库才可使用此节点。



IBM Cognos BI 导出节点以 Cognos BI 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”节点可以 SAS 格式输出数据，以便读入 SAS 或与 SAS 兼容的软件包中。有三种可用的 SAS 文件格式：SAS for Windows/OS2、SAS for UNIX 或 SAS Version 7/8。



Excel 导出节点以 Microsoft Excel 格式 (*.xls*) 输出数据。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。

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。请参阅主题第 10 页的『数据库源节点』，了解更多信息。

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

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。请参阅主题第 11 页的『添加数据库连接』，了解更多信息。

表名称。输入接收数据的表名称。如果选择 **插入表中** 选项，则可以通过单击 **选择** 按钮在数据库中选择一个现有表。

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。选中此项可将数据作为新行插入现有的数据库表中。

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。

注意：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。请参阅主题第 264 页的『数据库导出模式选项』，了解更多信息。

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。请参阅主题第 266 页的『数据库导出索引选项』，了解更多信息。

单击**高级**可指定批量加载和数据库提交选项。请参阅主题第 267 页的『数据库导出高级选项』，了解更多信息。

给表名和列名加上引号。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。
- **从不。**选中此项将禁用引号。

针对此数据生成一个导入节点。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画布。

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。数据库列名后面将显示单词 **<NEW>** 以表示这是新字段。

合并行。使用关键字字段（如交易 ID）合并关键字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关键字字段的值；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关键字字段。具有相应映射数据库列的任何映射字段都可用作关键字，但作为新数据库列添加的字段（在名称后显示有**<新>**）不能用作关键字。
- **用于合并的关键字。**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。要从列表中除去关键字段，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。
- **只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。**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”下单击您要除去映射的字段名称。

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，指定主要关键字字段，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，该字段的默认格式为：

```
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。该对话框将根据表中的设定自动生成 <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。

- 在 DB2 9.1 或更高版本上运行的 IBM InfoSphere Warehouse。请参阅主题第 265 页的『用于 IBM DB2 InfoSphere Warehouse 的选项』，了解更多信息。
- SQL Server 2008 或更高的 Enterprise 与 Developer 版本。请参阅主题第 265 页的『用于 SQL Server 的选项』，了解更多信息。
- Oracle 10g 与 11gR1 或更高的 Enterprise 或 Personal 版本。请参阅主题第 265 页的『用于 Oracle 的选项』，了解更多信息。

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。请务必保留文本参数 <table-name > 和 (<table-columns >)，因为这两个参数代表 IBM SPSS Modeler 定义的实际表名和列。

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。

对于包含长度、精确度或比例参数的类型（ BINARY 、 VARBINARY 、 CHAR 、 VARCHAR 、 NUMERIC 和 NUMBER ， 应由您指定一个长度，而不可让数据库服务器自动指定长度。例如，如果为长度指定一个理想值，如 VARCHAR(25) ，则必会如您所愿覆盖 IBM SPSS Modeler 中的存储类型。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）。请参阅主题第 266 页的『数据库导出索引选项』以获取更多信息。）

用于 IBM DB2 InfoSphere Warehouse 的选项

表空间。用于导出的表空间。数据库管理员可以创建表空间或将表空间配置为分区。建议选择其中一个表空间（而不是缺省表空间）以用于数据库导出。

按字段分区数据。指定要用于分区的输入字段。

使用压缩。如果选中此选项，那么将使用压缩创建要导出的表格（例如，相当于 SQL 中的 CREATE TABLE MYTABLE(...) COMPRESS YES;）。

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。

压缩支持。选择压缩层级。

- **行。**启用低级别压缩（例如，相当于 SQL 中的 CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW);）。
- **页面。**启用页面级别压缩（例如，SQL 中的 CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);）。

用于 Oracle 的选项

Oracle 10g 设置

使用压缩。如选中，将使用压缩为导出创建表格。对于此版本的数据库，仅可使用基本压缩（例如，SQL 中的 CREATE TABLE MYTABLE(...) COMPRESS;）。

Oracle 11gR1 设置

使用压缩。如选中，使用压缩为导出创建表格。

压缩支持。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**直接加载操作**选项的效果相同。
- **直接加载操作。**仅对批量（直接路径）插入操作启用压缩（例如，SQL 中的 CREATE TABLE MYTABLE(...) COMPRESS FOR DIRECT_LOAD OPERATIONS;）。
- **所有操作。**对所有操作启用压缩（例如，SQL 中的 CREATE TABLE MYTABLE(...) COMPRESS FOR ALL OPERATIONS;）。

Oracle 11gR2 设置 - 基本选项

使用压缩。如选中，使用压缩为导出创建表格。

压缩支持。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 11gR2 设置 -“高级”选项

使用压缩。如选中，使用压缩为导出创建表格。

压缩支持。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。可指定要包含的字段集合并根据需要自定义 CREATE INDEX 命令。

该对话框由两部分组成：

- 上方的文本字段显示可用于生成一个或多个 CREATE INDEX 命令的模板，该字段的默认格式为：

```
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。对话框将依次自动生成 <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。每一行将单独生成一个 CREATE INDEX 命令。

自定义 CREATE INDEX 命令

可以为所有索引或仅为特定索引自定义 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，对话框将根据表中的设定自动生成 <index-name > 和 <table-name >参数的值，但无法直接编辑这些参数值。

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。由于大多数数据库会自动为 CREATE TABLE 命令中指定的任何主要关键字段创建索引，因此不必明确创建这些字段的索引。请参阅主题第 264 页的『数据库导出模式选项』，了解更多信息。

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。

使用批处理提交。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。

InfoSphere Warehouse 选项。只有在连接到 InfoSphere Warehouse 数据库（IBM DB2 9.7 或更高版本）时才显示。不记录更新允许您避免在创建表格和插入数据时记录事件。

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。

- **逐行绑定。**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。
- **逐列绑定。**选中此项可使用逐列绑定将数据加载到数据库。逐列绑定可通过为每列数据库（位于参数化 INSERT 语句中）绑定一组 *N* 值提高处理速度。每次执行 INSERT 语句时都会在数据库中插入 *N* 行。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。*scripts* 子目录下已随附技术文档提供多个软件脚本。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。请参阅主题『批量加载程序设计』，了解更多信息。

- **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。默认情况下，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。默认情况下，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。*scripts* 子目录下已随附技术文档提供多个软件脚本。
- **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。
- **其他加载程序选项。**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\""` 的选项既包含 `-comment` 标志，还包含显示为 `This is a "comment"` 的注解内容。

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\\"` 的选项既包含标志 `-specialdir`，还包含显示为 `C:\Test Scripts\` 的目录。

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。请参阅主题第 272 页的『开发批量加载程序』，了解更多信息。

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。

注意：IBM SPSS Modeler 安装并不包含 Python 运行时解释器，因此需要另外安装 Python。请参阅主题第 267 页的『数据库导出高级选项』，了解更多信息。

在 IBM SPSS Modeler 安装目录的 `\scripts` 文件夹下，提供了用于下表中列出的数据库的脚本。

表 53. 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	<code>db2_loader.py</code>	请参阅主题『向 IBM DB2 数据库批量加载数据』，了解更多信息。
IBM Netezza	<code>netezza_loader.py</code>	请参阅主题第 270 页的『向 IBM Netezza 数据库批量加载数据』，了解更多信息。
Oracle	<code>oracle_loader.py</code>	请参阅主题第 270 页的『向 Oracle 数据库批量加载数据』，了解更多信息。
SQL Server	<code>mssql_loader.py</code>	请参阅主题第 271 页的『向 SQL Server 数据库批量加载数据』，了解更多信息。
Teradata	<code>teradata_loader.py</code>	请参阅主题第 271 页的『向 Teradata 数据库批量加载数据』，了解更多信息。

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。确保在要执行 `db2_loader.py` 的服务器（通常为运行 IBM SPSS Modeler Server 的主机）上安装了命令行处理器（对于 UNIX，应为 `db2`；对于 Windows，应为 `db2cmd`）。

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 客户机软件所使用的名称，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。请参阅主题第 135 页的『字段重排节点』，了解更多信息。

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

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。并不总是能检测到批量加载进程中的故障，启用此选项将执行附加检查，以确保加载正确数量的记录行。

空字符串

空字符串将作为 NULL 值导出到数据库。

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 ID (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除去空格。

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。

使用 Python 构建批量加载程序

默认情况下，IBM SPSS Modeler 会根据数据库类型搜索默认的批量加载程序。请参阅第 269 页的表 53。

可使用脚本 `test_loader.py` 来协助开发批量加载程序。请参阅主题第 274 页的『测试批量加载程序』，了解更多信息。

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 54. 传递给批量加载程序的参数.

参数	描述
schemafilename	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	要装入的表的名称。
ownername	表所有者的名称 (也称为模式名称)。
logfile	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 55. <table> 元素的属性.

属性	描述
delimiter	字段分隔符 (TAB 将表示为 \t)。

表 55. <table> 元素的属性 (续).

属性	描述
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 56. <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* 不会将数据装入数据库。

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。

注意：无法以过时的高速缓存格式写入文件，因为 IBM SPSS Modeler 已不再使用该高速缓存文件格式。IBM SPSS Modeler 高速缓存文件现以 IBM SPSS Statistics *.sav* 格式保存，您可以通过 Statistics 导出节点写入文件。请参阅主题第 290 页的『Statistics 导出节点』，了解更多信息。

“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。

- **包括字段名称。** 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。

每条记录后换新行。 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。您可以选择系统默认值、流默认值或 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。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。通常为计算机区域设置中定义的小数分隔符。
- **句点 (.)。**将使用句点字符作为小数分隔符。
- **逗号 (,.)。**将使用逗号字符作为小数分隔符。

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“变量文件”源节点。请参阅主题第 16 页的『变量文件节点』，了解更多信息。

IBM SPSS Data Collection 导出节点

IBM SPSS Data Collection 导出节点基于 IBM SPSS Data Collection 数据模型，以 IBM SPSS Data Collection 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（描述如何收集和組織观测值数据）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。请参阅主题第 20 页的『Data Collection 节点』，了解更多信息。

注意：此节点需要 IBM SPSS Data Collection Data Model 4.0 或更高版本，此版本已随 IBM SPSS Data Collection 软件一起发行。有关更多信息，请参阅 IBM SPSS Data Collection 网页：<http://www.ibm.com/software/analytics/spss/products/data-collection/>。除安装 Data Model 以外，不需要任何其他配置。

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。默认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。

针对此数据生成一个导入节点。选择此选项可自动生成将读取已导出数据文件的 IBM SPSS Data Collection 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。请参阅主题第 103 页的『编辑多重响应集』，了解更多信息。

Analytic Server 导出

Analytic Server 导出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。

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。请参阅主题第 24 页的『Analytic Server 选择数据源』，了解更多信息。

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**可添加到现有数据源中，或者选择**覆盖**可替换数据源的内容。

针对此数据生成一个导入节点。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。

IBM Cognos BI 导出节点

IBM Cognos BI 导出节点允许您采用 UTF-8 格式将数据从 IBM SPSS Modeler 流导出到 Cognos BI。这样，Cognos BI 可利用来自 IBM SPSS Modeler 的转换或评分数据。例如，可使用 Cognos BI Report Studio 创建一个基于导出数据的报告，包括预测和置信度值。然后可在 Cognos BI 服务器上保存报告并分配给 Cognos BI 用户。

注意：只能导出关系数据，不能导出 OLAP 数据。

要将数据导出到 Cognos BI，需要指定以下各项：

- Cognos 连接 - 到 Cognos BI 服务器的连接
- ODBC 连接 - 到 Cognos BI 服务器使用的 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 BI 服务器。

对于任何其他导出节点，也可使用节点对话框的“发布”选项卡，为使用 IBM SPSS Modeler Solution Publisher 的部署发布流。

Cognos 连接

您可以在此处指定与要用于导出的 Cognos BI 服务器的连接。该程序涉及将元数据导出到 Cognos BI 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**按钮将显示一个对话框，您可在该对话框中定义 URL 以及要将数据导出到的 Cognos BI 服务器的其他详细信息。如果已经通过 IBM SPSS Modeler 登录 Cognos BI 服务器，也可编辑当前连接的明细。请参阅主题第 30 页的『Cognos 连接』，了解更多信息。

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。要在其上创建导出数据包的 Cognos BI 服务器的路径和文件夹名称。

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。**（默认）单击**运行**即开始执行导出操作。
- **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。

针对此数据生成一个导入节点。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。

注意：必须确保此处指定的数据源与在 **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。请参阅主题第 11 页的『添加数据库连接』，了解更 多信息。

表名称。输入接收数据的表名称。如果选择 **插入表中** 选项，则可以通过单击 **选择** 按钮在数据库中选择一个 现有表。

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。选中此项可将数据作为新行插入现有的数据库表中。

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。

注意：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。请参阅主题第 264 页的『数据库导出模式选项』，了解更多 信息。

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。请参阅主题第 266 页的『数据库导出索 引选项』，了解更多信息。

单击 **高级** 可指定批量加载和数据库提交选项。请参阅主题第 267 页的『数据库导出高级选项』，了解更多信 息。

给表名和列名加上引号。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。
- **从不。**选中此项将禁用引号。

针对此数据生成一个导入节点。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 BI 导出节点，您可以将数据从 IBM SPSS Modeler 流导出至 Cognos TM1。这样，Cognos BI 可利用来自 IBM 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 BI，需要指定以下各项：

- 到 Cognos TM1 服务器的连接
- 将向其中导出数据的多维数据集
- 从 SPSS 数据名到对等 TM1 维度和测量值的映射

对于任何其他导出节点，也可使用节点对话框的“发布”选项卡，为使用 IBM SPSS Modeler Solution Publisher 的部署发布流。

连接到 Cognos TM1 多维数据集可导出数据

将数据导出到 IBM Cognos TM1 数据库的第一步是在 IBM Cognos TM1 对话框的**连接**选项卡中选择相关的 TM1 PM 主数据中心以及相关服务器和多维数据集。

注：必须确保将向其中导出数据的视图的数据结构与从中导入数据的视图的数据结构匹配。只要用于导出的视图与用于导入的视图共享同一结构，这两个视图就可以位于不同的多维数据集中。

PM 系统。输入包含要连接到的 TM1 服务器的主数据中心。

TM1 服务器。建立 Cognos 主数据中心连接后，请选择包含要导入的数据的服务器，然后单击**登录**。如果您先前未连接到此服务器，那么系统将提示您输入**用户名和密码**；或者，您可以选择其他服务器。

选择要导出的 TM1 多维数据集。显示可以向其中导出数据的 TM1 服务器内多维数据集的名称。双击多维数据集可显示该多维数据集内的视图。

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。选中多维数据集之后，请使用**映射**选项卡，从而使 TM1 字段与相关的 SPSS 字段相匹配。

映射可供导出的 Cognos TM1 数据

选中 TM1 PM 主数据中心以及相关服务器和多维数据集之后，您可以在 IBM Cognos TM1 对话框的“**连接**”选项卡中确定将哪些 SPSS 字段映射到相关 TM1 字段中。

在可能的情况下，TM1 字段将自动映射到相关 SPSS 字段。

注：必须确保将向其中导出数据的视图的数据结构与从中导入数据的视图的数据结构匹配。只要用于导出的视图与用于导入的视图共享同一结构，这两个视图就可以位于不同的多维数据集中。

字段。列出由 SPSS 数据文件提供、并且数据可供导出的数据字段名称。需要时，请使用列表底部的按钮来更改显示的字段；例如，您可以显示所有字段、仅显示连续类型的字段、或仅显示分类类型的字段。

TM1 维度。显示**连接**选项卡中所选的多维数据集及其常规维度和测量维度。如果未自动填写此字段，请选择要映射到 SPSS 数据字段的 TM1 元素的名称。

SAS 导出节点

注意：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 Version 7/8。

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。

导出。指定导出文件格式。选项有 **SAS for Windows/OS2**、**SAS for UNIX** 和 **SAS Version 7/8**。

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。
- **将名称用作变量标签。**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**。

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 SAS 源节点。请参阅主题第 32 页的『SAS 源节点』，了解更多信息。

Excel 导出节点

Excel 导出节点以 Microsoft Excel 格式 (*.xls*) 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。默认文件名为 *excelxp.xls*。

文件类型。选择要导出的 Excel 文件类型。

创建新文件。新建 Excel 文件。

插入到现有文件。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。

包括字段名称。指定字段名是否可以包含在工作表的第一行中。

从单元格开始。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。

针对此数据生成一个导入节点。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。请参阅主题第 32 页的『Excel 源节点』，了解更多信息。

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 导出文件。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默认结构：

```
<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。请参阅主题第 281 页的『XML 映射记录选项』，了解更多信息。

映射字段。表示已映射的字段数。

针对此数据生成一个导入节点。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。请参阅主题第 33 页的『XML 源节点』，了解更多信息。

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。

第 8 章 IBM SPSS Statistics 个节点

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 节点显示在节点选项板的专门部分中。

注意: 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 选项板包含下列节点:



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 中的高速缓存文件 (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。IBM SPSS Statistics Base 和其他产品可以读取 *.sav* 或 *.zsav* 文件。这种格式也用于 IBM SPSS Modeler 中的某些缓存文件。

注意: 如果您的 SPSS Statistics(S) 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 Statistics(S) 节点, 那么您可能会得到许可授权错误。当某个分支的 SPSS Statistics(S) 会话尚未结束, 而另一个分支试图启动时, 就会出现此错误。如可行, 应重新设计流, 以确保带有 SPSS Statistics(S) 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 早期版本中的高速缓存文件。如果想要导入已保存的高速缓存文件, 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。

文件已通过密码进行加密。 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。
- **读取用作名称的标签。** 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 中将它们用作变量名称。

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。
- **读取用作数据的标签。** 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, 请选择此选项。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。

使用字段格式信息确定存储类型。 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 *Fn.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。请参阅主题第 103 页的『编辑多重响应集』，了解更多信息。

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。请参阅主题第 259 页的『IBM SPSS Statistics 帮助应用程序』，了解更多信息。请参阅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。请参阅主题第 291 页的『重命名或过滤 IBM SPSS Statistics 的字段』，了解更多信息。

语法参考。有关特定的 IBM SPSS Statistics 过程的详细信息，请参阅《IBM SPSS Statistics 命令语法参考》指南，该文档包含在您的 IBM SPSS Statistics 软件副本中。要从“语法”选项卡上查看该指南，请选择**语法编辑器**选项，并单击“启动 IBM SPSS Statistics 语法帮助”按钮。

注意：此节点并不支持所有 IBM SPSS Statistics 语法。请参阅主题『允许的语法』，了解更多信息。

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

检查。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-!ENDDFINE
DELETE VARIABLES
DO IF
DO REPEAT
ELSE
ELSE IF
END CASE
END FILE
END IF
END INPUT PROGRAM
END LOOP
END REPEAT
EXECUTE
FILE HANDLE
FILE LABEL
FILE TYPE-END FILE TYPE
FILTER
FORMATS
IF
INCLUDE
INPUT PROGRAM-END INPUT PROGRAM
INSERT
LEAVE
LOOP-END LOOP
MATRIX-END MATRIX
MISSING VALUES

命令名称
N OF CASES
NUMERIC
PERMISSIONS
PRESERVE
RANK
RECODE
RENAME VARIABLES
RESTORE
RMV
SAMPLE
SELECT IF
SET
SORT CASES
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。请参阅主题第 259 页的『IBM SPSS Statistics 帮助应用程序』，了解更多信息。请参阅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。请参阅主题第 228 页的『查看输出』，了解更多信息。

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。请参阅主题第 259 页的『IBM SPSS Statistics 帮助应用程序』，了解更多信息。请参阅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。请参阅主题第 291 页的『重命名或过滤 IBM SPSS Statistics 的字段』，了解更多信息。

语法参考。有关特定的 IBM SPSS Statistics 过程的详细信息，请参阅《*IBM SPSS Statistics 命令语法参考*》指南，该文档包含在您的 IBM SPSS Statistics 软件副本中。要从“语法”选项卡上查看该指南，请选择**语法编辑器**选项，并单击“启动 IBM SPSS 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 IBM SPSS Statistics 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。

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。

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（例如得分）写入活动数据集。

要创建语法，请执行下列操作：

1. 单击**选择对话框**按钮。
2. 选择其中一个选项：

- **分析。**列出 IBM SPSS Statistics 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 IBM SPSS Statistics 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 IBM SPSS Statistics 对话框。如果没有适用的对话框，那么将不显示此选项。

注意：不会显示“自动数据准备”对话框。

如果某个 IBM SPSS Statistics 自定义对话框创建有新字段，则这些字段不能在 IBM 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（默认位置是 IBM 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（默认位置是 IBM SPSS Modeler 安装目录）。

单击**运行时**，结果会显示在 IBM SPSS Statistics 输出查看器中。有关查看器的更多信息，请参阅 IBM SPSS Statistics 文档。

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**自定义**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 Collaboration and Deployment Services 存储库，随后在 Web 浏览器中查看。请参阅主题第 229 页的『发布到 Web』，了解更多信息。

注意：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，脚本编制 API（基本和 *Python SpssClient* 模块）将在 IBM SPSS Modeler 中不可用。

Statistics 导出节点

Statistics 导出节点可以以 IBM SPSS Statistics *.sav* 格式导出数据。IBM SPSS Statistics Base 和其他模块能够读取 IBM SPSS Statistics *.sav* 文件。这也是 IBM SPSS Modeler 缓存文件所使用的格式。

有时候，将 IBM SPSS Modeler 字段名称映射到 IBM SPSS Statistics 变量名称时会出错，这是因为 IBM SPSS Statistics 变量名称被限制为 64 个字符且不能包含某些字符，比如空格、美元符号 (\$) 和 划线 (-) 等。有两种方法可以规避上述限制：

- 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。请参阅主题第 291 页的『重命名或过滤 IBM SPSS Statistics 的字段』，了解更多信息。
- 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。请参阅主题第 103 页的『编辑多重响应集』，了解更多信息。

Statistics 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。

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。

使用密码对文件进行加密。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 IBM SPSS Modeler V16 或更高版本，或者由 IBM SPSS Statistics 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 IBM SPSS Modeler 导出至 IBM SPSS Statistics *.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 IBM SPSS Modeler 字段名称和字段标签。名称将以 IBM SPSS Statistics 变量名方式导出，而标签则以 IBM SPSS Statistics 变量标签方式导出。

- **将名称用作变量标签。**选中此选项可将 IBM SPSS Modeler 字段名称用作 IBM SPSS Statistics 中的变量标签。IBM SPSS Modeler 允许在字段名称中包含不适用于 IBM SPSS Statistics 变量名称的字符。为了防止创建无效的 IBM SPSS Statistics 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 IBM SPSS Statistics，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。请参阅主题第 259 页的『IBM SPSS Statistics 帮助应用程序』，了解更多信息。要在不打开外部程序的情况下直接创建 IBM SPSS Statistics .sav 或 .zsav 文件，请取消选择此选项。

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。请参阅主题第 284 页的『Statistics 文件节点』，了解更多信息。

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。请参阅主题第 101 页的『设置过滤选项』以获取更多信息。本主题介绍几个将数据读入 IBM SPSS Statistics 的技巧。

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 的可视化编程界面易于学习的原因之一是，每个节点都有着明确定义的功能。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便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。请参阅主题第 299 页的『超节点和脚本编写』，了解更多信息。

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。注意：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（形状像星星）。
 - 右键单击超节点，然后从上下文菜单中选择：

创建超节点 > 从所选项

- 从“超节点”菜单中，选择：

创建超节点 > 从所选项

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。

注意：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好的参数输入值，且不输入密码也可执行锁定的超节点。

注意：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**；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。

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。请参阅主题『[锁定和解锁超节点](#)』，了解更多信息。

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。请参阅主题第 295 页的『编辑锁定的超节点』，了解更多信息。

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 57. 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，您能够设置用户定义的变量，如 *Minvalue*，当在脚本编写或 CLEM 表达式中使用这些变量时可以指定其值。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。

存储类型。 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、日期及时间戳记。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。注意对于日期参数，其值必须用 ISO 标准符号（即 YYYY-MM-DD）指定。系统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，请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**。

请注意，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 58. 要创建的参数

参数	值(V)	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 IBM 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。

为终端超节点打开“脚本编写”对话框

- 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**。

注意：如果在对话框中选择了**运行此脚本**，那么将仅针对流和超节点执行超节点脚本。

IBM SPSS Modeler DVD 中提供的《脚本和自动化指南》中讨论了脚本编写的具体选项以及其在 IBM SPSS Modeler 内的使用。

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。保存和加载超节点时，请注意，它们使用 *.slb* 扩展名。

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*）或浏览到其他超节点文件。
3. 单击 **加载**。

注意：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。

声明

这些信息开发用于在全球提供的产品和服务。

IBM 可能在其他国家或地区不提供本文中讨论的产品、服务或功能特性。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、程序或服务的操作，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。您可以用书面形式将许可查询寄往：

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

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，或用书面方式将查询寄往：

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

以下段落对于英国和与当地法律有不同规定的其他国家或地区均不适用：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、适销和适用于某特定用途的保证。某些国家或地区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。此处的信息会定期进行更改；这些更改会体现在本出版物的新版本中。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。

在本信息材料中对任何非 IBM 网站的引用仅为了方便用户，并不以任何方式表明对这些网站的认可。那些 Web 站点中的资料不是 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) 使其能够在独立创建的程序和其它程序（包括本程序）之间进行信息交换，以及 (ii) 使其能够对已经交换的信息进行相互使用，请与下列地址联系：

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

此类信息的提供应遵照相关条款和条件，其中包括在某些情况下支付适当费用。

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。

此处所含的性能数据均在受控环境下决定。因此，在其他操作环境中获得的结果可能差异较大。有些测量可能在开发级的系统中进行，不保证这些测量结果与常用系统上的测量结果相同。此外，有些测量结果可能通过推断来估计得出。实际结果可能有所差异。此文档的用户应针对其具体环境验证适用的数据。

涉及非 IBM 产品的信息可从这些产品的供应商、其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。

有关 IBM 未来方向或意向的所有声明均可能未经通知即变更或撤销，并且仅代表目标和目的。

本信息包含日常业务运营中使用的数据和报告的示例。为了尽可能详尽地对其进行说明，示例中包含了人员的姓名、公司、品牌和产品的名称。所有这些名称均为虚构，与真实商业企业使用的名称和地址的任何雷同纯属巧合。

如果您正在查阅此信息的软拷贝，照片和彩色插图可能不会显示。

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp.，在全球许多管辖区域的商标或注册商标。其他产品和服务名称可能是 IBM 或其他公司的商标。当前的 IBM 商标列表，可从 Web 站点 www.ibm.com/legal/copytrade.shtml 上『版权和商标信息』部分获取。

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 或其子公司在美国和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 以及 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/或其他国家或地区的注册商标。

Java 和所有基于 Java 的商标及徽标皆为 Oracle 和/或其附属公司的商标或注册商标。

其他产品和服务名称可能是 IBM 或其他公司的商标。

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): 观测值聚类围绕中心点的程度的一种测度。对于正态分布，峰度统计量的值为零。正峰度值表示相对于正态分布，观察值在分布中心的聚集更多，同时尾部更薄，直到分布极值。在这一点，尖峰态分布的尾部比正态分布的尾部要厚。负峰度值表示相对于正态分布，观察值聚集得少并且尾部较厚，直到分布极值。在这一点，低峰态分布的尾部比正态分布的尾部要薄。

M

最大值 (Maximum): 数字变量的最大值。

平均值 (Mean): 集中趋势的测量值。算术平均值，等于总和除以观测值数。

中位数 (Median): 第 50 个百分位，大于该值和小于该值的个案数各占一半。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）。

最小值 (Minimum): 数字变量的最小值。

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 (Range): 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): 分布的不对称度量值。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：取大于其标准误差差两倍的偏度值指示离开对称的距离。

标准差 (standard deviation): 均数离差的度量值，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 (Standard Deviation): 均数离差的度量值。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): 因样本变化而变化的检验统计量值的个数的度量值。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检验（即如果比率小于 -2 或大于 +2，则可以拒绝正态）。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）。

平均值标准误差 (Standard Error of Mean): 来自同一分布的各样本之间均值差异的测量值。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): 偏态与其标准误差的比率可用作正态检验（即，如果该比率小于 -2 或大于 +2，则可以拒绝正态）。偏度正值越大表示长尾向右越长；负值表示向左的长尾。

Sum: 所有具有非缺失值的观测值的和或总计。

U

唯一 (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 (Valid): 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 (Variance): 平均值离散度的测量值，等于均值的平方差除以观测值数减一的差。方差按单元计量，即变量自身单元数的平方。

索引

[A]

案例数据

IBM SPSS Data Collection 源节点 20, 21

[B]

百分位数分级 116

帮助应用程序 259

保存

输出 228

输出对象 228, 231

报告

保存输出 231

报告节点 250

“模板”选项卡 250

“输出”选项卡 231

报告浏览器 251

编辑图形

图形元素的大小 217

编辑直观表示 213

边距 217

尺度 218

点高宽比 216

点形状 216

点旋转 216

对类别排序 219

规则 214

合并类别 219

划线 216

类别 219

面板 221

排除类别 219

拼并类别 219

数字格式 217

添加三维效果 221

填充 217

透明度 216

图例位置 224

选的上下文菜单创建超节点 214

颜色和样式 216

轴 218

转换坐标系 221

转置 221

自动设置 214

text 215

变换节点 243

变量标签

Statistics 导出节点 290

Statistics 文件节点 284

变量类型

在直观表示中 142

变量名

数据导出 262, 274, 279, 290

变量文件节点 16

设置选项 16

自动日期识别 16

标记 61, 65

标记元素 208, 210

标签 98

导出 279, 290

导入 32, 284

指定 96, 97, 98

标签类型

IBM SPSS Data Collection 源节点 23

标签字段

标注输出中的记录 99

标志类型 93, 98

标准差

分级节点 117

设置全局量节点 251

统计量输出 246

标准化连续目标 84, 92

标准化值

图形节点 183, 193

表

保存输出 231

保存为文本 231

连接 61

表达式构建器 51

表格浏览器

重新为列排序 231, 232

搜索 232

选择单元格 231, 232

“生成”菜单 232

表格输出

重新为列排序 231

选择单元格 231

表节点 231

输出设置 231

“设置”选项卡 231

“输出”选项卡 231

表面图 145

饼图 145

三维散点图 145

使用计数 145

示例 156

在地图上 145

并行处理

合并 65

聚合节点 58

并行处理 (续)

排序 61

不平衡的数据 56

不完整记录 63

部分连接 61, 64

[C]

财政年度

时间区间节点 129

残差

矩阵节点 234

测量级别

定义 93

在直观表示中 142

直观表示中的更改 141

测试样本

分区数据 121

差分变换

ARIMA 模型 74

差异节点

组合设置 69, 70

查看

HTML 输出在浏览器中 230

查询

数据库源节点 10, 11

查询编辑器

数据库源节点 15

超节点 293

保存 299

编辑 296

创建 294

创建缓存 298

放大 296

过程超节点 293

加载 299

脚本编写 299

解锁 295

密码保护 295

嵌套 295

设置参数 297

使用注释, 带有 296

锁定 295

以下的类型 293

源超节点 293

终端超节点 294

超节点参数 297, 298

成员 (SAS 导入)

设置 32

持续时间计算

自动数据准备 83

尺度因子 57
 冲突修饰符 221
 重复
 记录 67
 字段 61, 101
 重命名
 导出字段 291
 地图文件 163
 直观表示模板 163
 直观表示样式表 163
 重命名输出对象 228
 重新编码 112, 113, 114
 重新分类节点 113, 114
 概述 112, 114
 由条形图生成 176
 重新结构化节点 122, 123
 与汇总节点 122
 重要性
 比较平均值 248
 平均值节点 249
 重组数据 122
 抽样框 53
 抽样数据 56
 处理缺失值 79
 传输文件
 SAS 源节点 32
 创建
 新字段 105, 106
 创新离群值
 时间序列建模器 75
 从图形中生成节点 211
 重新分类节点 211
 过滤节点 211
 派生节点 211
 平衡节点 211
 选择节点 211
 存储格式 19

[D]

打开
 输出对象 228
 打印输出 228
 大小图形交叠 138
 大型数据库 51
 执行数据审核 237
 代码变量
 IBM SPSS Data Collection 源节点 21
 带状图 145
 单因素 ANOVA
 平均值节点 248
 单元格范围
 Excel 文件 32
 导出
 超节点 299
 地图文件 163

导出 (续)
 输出 230
 直观表示模板 163
 直观表示样式表 163
 IBM Cognos TM1 中的数据 278
 导出节点 261
 Analytic Server 导出 276
 导出数据
 到 Excel 279
 到 IBM SPSS Statistics 290
 平面文件格式 274
 至数据库 262
 DAT 文件 279
 IBM Cognos BI 导出节点 30, 276, 277
 IBM Cognos TM1 导出节点 278
 SAS 格式 279
 text 279
 XML 格式 280
 导出小数位 100
 导航 256
 导入
 超节点 299
 地图文件 163
 直观表示模板 163
 直观表示样式表 163
 IBM Cognos BI 中的报告 30
 IBM Cognos BI 中的数据 29
 IBM Cognos TM1 中的数据 31
 等值线图 145
 示例 159
 地图
 分发 169
 合并特征 167
 平滑 165, 166
 删除个别元素 168
 删除特征 168
 特征标签 166
 投影 168
 细化 165, 166
 移动特征 168
 转换 ESRI shapefile 164
 地图文件
 重命名 163
 导出 163
 导入 163
 删除 163
 位置 162
 在图形画板模板选择器中选择 144
 地图直观表示
 创建 150
 地图转换实用程序 164, 165
 地图 shapefile
 编辑预先安装的 SMZ 地图 164
 概念 164
 类型 164

地图 shapefile (续)
 与图形板模板选择器配合使用 164
 点图 145
 示例 154
 2-D 145
 调查数据
 导入 21, 23, 24
 IBM SPSS Data Collection 源节点 20
 调整后的倾向得分
 平衡数据 57
 丢弃
 字段 101
 抖动 173
 逗号分隔文件
 保存 231
 导出 230, 279
 堆积 213, 221
 对模型中使用的数据进行掩饰 110
 对数变换
 时间序列建模器 75
 对数据进行排序 60, 135
 多重派生 107
 多重散点图节点 183
 使用图形 185
 “散点图”选项卡 183
 “外观”选项卡 185
 多重响应集
 定义 103
 多二分集 103
 多类别集 103
 删除 103
 在直观表示中 142
 IBM SPSS Data Collection 源节点 180, 21, 23, 24
 IBM SPSS Statistics 源节点 284
 多二分集 103
 多个输入 61
 多个字段
 选择 107
 多类别集 103

[E]

二十分位数分级 116

[F]

发布到网络 229
 反连接 61
 范围 93
 缺失值 96
 统计量输出 246
 方案 6
 方差
 统计量输出 246

- 非随机样本 53
- 分布 179
- 分布节点 175
 - 使用表 176
 - 使用图形 176
 - “散点图”选项卡 175
 - “外观”选项卡 176
- 分层样本 53, 55, 56
- 分隔的文本数据 16
- 分隔符 16, 267
- 分级节点
 - 概述 114
 - 固定宽度分级 115
 - 均数/标准差分级 117
 - 排序 117
 - 设置选项 115
 - 相等计数 116
 - 相等总和 116
 - 预览分级 119
 - 最优 118
- 分级散点图 145
 - 十六进制分级 145
- 分类数据 95
- 分区节点 121
- 分区数据 121
 - 评估图表 199
 - “分析”节点 235
- 分区字段 99, 121
- 分数排秩 117
- 分位数
 - 分级节点 116
- 分析浏览器
 - 解释 236
- 分析数据视图 9
- 分钟增量
 - 时间区间节点 133
- 分组符号
 - 数字显示格式 100
- 封装节点 294
- 符合矩阵
 - “分析”节点 235
- 覆盖数据库表 262
- 复式条形图
 - 示例 152
- 复制类型属性 99
- 复制直观表示 224

[G]

- 干预
 - 创建 193
- 高低图 145
- 高速缓存
 - 超节点 298
- 高速缓存文件节点 284
- 高-低-闭合图 145

- 割点
 - 分级节点 114
- 格式
 - 数据 19
- 格式化文件 32
- 工作表
 - 从 Excel 导入 32
- 估计期 127
- 固定文件节点
 - 概述 18
 - 设置选项 18
 - 自动日期识别 18
- 固定字段文本数据 18
- 关键字段 58, 122
- 关键字段方法 61
- 关联绘制 186
- 管理器
 - “输出”选项卡 228
- 过滤节点
 - 多重响应集 103
 - 概述 101
 - 设置选项 101
- 过滤字段 64, 101
 - 用于 IBM SPSS Statistics 291

[H]

- 合并节点 61
 - 设置选项 63, 64
 - 优化设置 65
- 合并选项, 数据库导出 263
- 合计
 - 设置全局量节点 251
 - 统计量输出 246
- 合计值 58
- 回避 213, 221
- 绘制关联 186
- 汇总记录 122
- 汇总时间序列数据 126
- 汇总数据 57
- 货币显示格式 100

[J]

- 积分
 - ARIMA 模型 74
- 基线
 - 评估图选项 199
- 集合
 - 转换 113, 114
 - 转换为标志 122
- 集合类型 93
- 极坐标 221
- 季度数据
 - 时间区间节点 129

- 季节性差分变换
 - ARIMA 模型 74
- 季节性加性离群值
 - 时间序列建模器 75
- 季节性阶
 - ARIMA 模型 74
- 记录
 - 标签 99
 - 合并 61
 - 计数 58
 - 转置 124
 - length 18
- 记录操作节点 51
 - 时间区间节点 125
- 记录的平均值 57
- 计数
 - 分级节点 116
 - 统计量输出 246
- 计数字段
 - 时间区间节点 126
 - 填充或汇总时间序列 126
- 计算持续时间
 - 自动数据准备 83
- 加权样本 55
- 加性离群值
 - 时间序列建模器 75
- 假值 98
- 检查类型 98
- 减少数据 52, 53
- 将集合转换为标志 122
- 将值分组 176
- 降序 60
- 交叉列表
 - 矩阵节点 233, 234
- 交叠地图 145
- 脚本编制
 - 超节点 299
- 角色
 - 为字段指定 99
- 角色建模
 - 为字段指定 99
- 结
 - 分级节点 116
- 节点属性 298
- 截断字段名 101, 102
- 解锁超节点 295
- 局部加权最小二乘回归 (LOESS)
 - 散点图节点 172
- 局部趋势离群值
 - 时间序列建模器 75
- 矩阵表中的
 - 空白值 233
- 矩阵节点 233
 - 行百分比 234
 - 交叉列表 234
 - 列百分比 234

- 矩阵节点 (续)
 - 输出浏览器 234
 - 突出显示 234
 - 为行和列排序 234
 - “设置”选项卡 233
 - “输出”选项卡 231
 - “外观”选项卡 234
- 矩阵浏览器
 - “生成”菜单 234
- 矩阵输出
 - 保存为文本 231
- 聚合节点
 - 并行处理 58
 - performance 58
- 聚类 213, 221
- 聚类样本 53, 55
- 均数
 - 分级节点 117
- 均数/标准差
 - 用于分级字段 117
- 均值
 - 设置全局量节点 251
 - 统计量输出 246

[K]

- 卡方统计量
 - 矩阵节点 234
- 科学表示法显示格式 100
- 空白值
 - 空白值 233
- 空行
 - Excel 文件 32
- 空间时间限制中的密度定义 77
- 空值 96
 - 空白值 233

[L]

- 累计时间序列数据 126
- 类型 19
- 类型节点
 - 标志字段类型 98
 - 复制类型 99
 - 概述 92
 - 连续数据 97
 - 两步聚类中的 96
 - 名义数据 98
 - 设置建模角色 99
 - 设置选项 93, 95
 - 有序数据 98
- 类型属性 99
- 离群值
 - 在时间序列模型中 75
 - ARIMA 模型 75

- 利润图 195, 201
- 历史记录节点 135
 - 概述 134
- 连接 61, 63
 - 部分外部 64
 - 到 IBM SPSS Collaboration and Deployment Services 存储库 6
- 连接记录 66
- 连接数据集 66
- 连续关键值 58
- 连续数据 95, 97
- 连续数据抽样 53
- 链接
 - Web 节点 188
- 两步聚类中的 96
 - 分级节点 115
 - 填充值 109
- 列顺序
 - 表格浏览器 231, 232
- 流参数 15
- 流程图 145
- 路径图 145

[M]

- 密度
 - 三维散点图 145
- 面板 138
- 面板图形交叠 138
- 面积图 145
 - 三维散点图 145
- 秒钟增量
 - 时间区间节点 134
- 名义数据 98
- 模板
 - 报告节点 250
 - 重命名 163
 - 导出 163
 - 导入 163
 - 删除 163
- 模拟数据
 - “模拟生成”节点 35
- 模式
 - 数据库导出节点 264
- 模型
 - 匿名化数据 110
 - ARIMA 74
- 模型评估 195
- 模型视图
 - 自动数据准备过程中 86
- 模型选项
 - Statistics 模型节点 287

[N]

- 内部连接 61
- 匿名化节点
 - 创建匿名化值 112
 - 概述 110
 - 设置选项 111
- 匿名化字段名 103
- 年数据
 - 时间区间节点 129

[P]

- 排序
 - 差异节点 67
 - 记录 60
 - 预排序字段 61, 68
 - 字段 135
- 排序观测值 117
- 排序节点
 - 概述 60
 - 优化设置 61
- 派生节点
 - 标志 107
 - 从图形中生成 211
 - 从网络图形链接生成 190
 - 从自动数据准备生成 92
 - 对值重新编码 109
 - 多重派生 107
 - 概述 105
 - 计数 109
 - 名义 108
 - 设置选项 106
 - 条件 109
 - 通过分级节点生成 119
 - 通过分级生成 114
 - 转换字段存储类型 109
 - 状态 108
 - formula 107
- 批量加载 267, 268
- 匹配项
 - 评估图选项 200
- 偏倚数据 56
- 频率
 - 分级节点 116
- 平方根变换
 - 时间序列建模器 75
- 评分
 - 评估图选项 200
- 评估节点 195
 - 得分表达式 200
 - 读取结果 201
 - 匹配条件 200
 - 使用图形 202
 - 业务规则 200
 - “散点图”选项卡 199

评估节点 (续)
 “外观”选项卡 201
 “选项”选项卡 200
 评估模型 235
 平行坐标图 145
 平衡节点
 从图形中生成 211
 平衡因子 57
 平均值
 比较 247, 248
 平均值的标准误差
 统计量输出 246
 平均值节点 247
 成对字段 248
 重要性 248
 独立组 248
 输出浏览器 248, 249
 “输出”选项卡 231
 平面文件 16
 平面文件导出节点 274
 “导出”选项卡 274

[Q]

期望值
 矩阵节点 234
 气泡图 145
 强制转换值 98
 清除值 48
 倾向得分
 平衡数据 57
 区分节点
 概述 67
 排序记录 67
 优化设置 68
 区间
 时间序列数据 125
 权重
 评估图表 199
 全局值 251
 缺失值 79, 96, 98
 空白值 233
 在汇总节点中 57

[R]

热图 145
 示例 157
 日测量
 时间区间节点 131
 日期
 设置格式 100
 日期识别 16, 18
 日期/时间 93

如果任何函数都为真, 则为真
 时间序列汇总 126

[S]

散点图 145, 169, 183
 分级 145
 三维散点图 145
 十六进制分级 145
 散点图节点 169
 使用图形 174
 “散点图”选项卡 172
 “外观”选项卡 174
 “选项”选项卡 173
 散点图矩阵
 示例 158, 160
 散点图矩阵 (SPLOM) 145
 删除
 地图文件 163
 输出对象 228
 直观表示模板 163
 直观表示样式表 163
 设为标志节点 122
 设置全局量节点 251
 “设置”选项卡 251
 设置随机种子
 抽样记录 121
 审核
 初始数据审核 237
 “数据审核”节点 237
 生成标志 122, 123
 升序 60
 时测量
 时间区间节点 132
 十分位数分级 116
 时间戳 93
 时间格式 100
 时间区间节点 126, 127
 概述 125
 时间散点图节点 193
 使用图形 194
 “散点图”选项卡 193
 “外观”选项卡 194
 时间序列 134
 时间序列模型
 离群值 75
 序列变换 75
 指数平滑法标准 73
 周期性 75
 专家建模器标准 73
 转换函数 75
 ARIMA 标准 74
 时间序列数据
 标注 125, 126, 127
 定义 125, 126, 127
 根据数据构建 126

时间序列数据 (续)
 估计期 127
 汇总 125, 126
 区间 126
 填充 125, 126
 holdout 127
 实例化 93, 95, 96
 源节点 49
 十六进制分级散点图 145
 实数范围 97
 使用类型 19, 93
 市场调查数据
 导入 21, 24
 IBM SPSS Data Collection 源节点 20, 23
 示例
 概述 4
 应用程序指南 3
 收集节点 180
 使用图形 182
 “外观”选项卡 181
 “选项”选项卡 181
 收入
 评估图表 199
 收益图表 195, 201
 受监督的离散化 118
 输出 256, 257
 保存 228
 打印 228
 导出 230
 生成新节点 228
 HTML 230
 输出格式 231
 输出管理器 228
 输出节点 227, 231, 233, 235, 237, 246, 250, 251, 252, 253, 254, 256, 257, 258, 288
 发布到网络 229
 “输出”选项卡 231
 输出文件
 保存 231
 输出元素 257
 输入数据的顺序 65
 数据
 存储类型 96
 汇总 57
 理解 51
 匿名化 110
 审核 237
 探索 237
 准备 51
 storage 109, 110
 数据访问方案 9
 数据库
 批量加载 267, 268
 数据库导出节点 262

- 数据库导出节点 (续)
 - 表名称 262
 - 合并选项 263
 - 模式 264
 - 数据源 262
 - 索引表 266
 - 映射源数据字段到数据库列 263
 - “导出”选项卡 262
- 数据库连接
 - 定义 11
 - 预设值 12
- 数据库源节点 10
 - 查询编辑器 15
 - 选择表和视图 14
 - SQL 查询 11
- 数据类型 18, 79, 93
 - 实例化 95
- 数据审核浏览器
 - 编辑菜单 239
 - 生成节点 243
 - 生成图形 243
 - 文件菜单 239
- 数据提供者定义 6
- 数据源
 - 数据库连接 11
- 数据质量
 - “数据审核”浏览器 241
- 属性
 - 节点 298
 - 在地图中 164
- 数字显示格式 100
- 水平变动离群值
 - 时间序列建模器 75
- 瞬态离群值
 - 时间序列建模器 75
- 顺序合并 61
- 四分位数分级 116
- 搜索
 - 表格浏览器 232
- 随机种子值
 - 抽样记录 121
- 缩放 296
- 锁定超节点 295
- 索引数据库表 266

[T]

- 探索数据
 - “数据审核”节点 237
- 探索图形 203
 - 标记元素 210
 - 魔棒 210
 - 区域 208
 - 图形带状区域 204
- 特征
 - 在地图中 164

- 提交大小 267
- 提升图 195, 201
- 替换字段值 109
- 添加
 - 记录 57
- 填充节点
 - 概述 109
- 填充时间序列数据 126
- 条件
 - 为合并指定 64
 - 指定序列 109
- 条形图 145
 - 计数 145
 - 三维散点图 145
 - 示例 151, 152
 - 在地图上 145
- 统计量
 - 矩阵节点 233
 - 在直观表示中编辑 221
 - “数据审核”节点 237
- 统计量节点 246
 - 统计量 246
 - 相关 246
 - 相关标签 246
 - “设置”选项卡 246
 - “输出”选项卡 231
- 统计量浏览器
 - 解释 246
 - 生成过滤节点 247
 - “生成”菜单 246
- 投资回报率
 - 图表 195, 201
- 图标, IBM Cognos BI 28
- 图表
 - 保存输出 231
- 图表输出 257
- 图表选项 258
- 图形
 - 保存 226
 - 保存编辑的布局 225
 - 保存布局更改 225
 - 保存输出 231
 - 打印 226
 - 带状区域 204
 - 导出 226
 - 多重散点图 183
 - 复制 226
 - 脚注 224
 - 来自图形板 141
 - 默认颜色图示 225
 - 评估图表 195
 - 区域 208
 - 三维散点图 140
 - 删除区域 208
 - 生成节点 211
 - 时间序列 193

- 图形 (续)
 - 收集 180
 - 探索 203
 - 条形图 175
 - 图 169
 - 图形元素的大小 217
 - 网络 186
 - 旋转 3D 图像 140
 - 样式表 225
 - 由数据审核生成 243
 - 直方图 179
 - 轴标签 224
 - 注解选项卡 140
 - title 224
 - “输出”选项卡 139
- 图形板
 - 图形类型 145
- 图形板节点 141
 - “外观”选项卡 161
- 图形的交叠 138
- 图形节点 137
 - 动画 138
 - 多重散点图 183
 - 分布 175
 - 集合 180
 - 交叠 138
 - 面板 138
 - 评估(E) 195
 - 散点图 169
 - 时间散点图 193
 - 图形板 141
 - 直方图 179
 - Web 186
- 图形可视化
 - 示例 159
- 图形类型
 - 图形板 145
- 图形元素
 - 冲突修饰符 221
 - 更改 221
 - 转换 221
- 图形中的带状区域 204
- 图形中的动画 138
- 图形中的魔棒 210
- 图形中的区域 208
- 图形中的透明度 138
- 图注
 - 位置 224

[W]

- 外部连接 61
- 网络图形的布局 188
- 网络图形的定向布局 188
- 唯一性记录 67
- 未定义值 63

未使用字段排除
 自动数据准备 82
位图索引
 数据库表 266
文本文件 16
 导出 279
文档 3
无偏倚数据 56
五分位数分级 116

[X]

系统变量
 IBM SPSS Data Collection 源节点 21
系统化样本 53
显示格式
 分组符号 100
 货币 100
 科学表示法 100
 数字 100
 小数位 100
显著水平
 相关强度 246
线散点图 169, 183
相等计数
 分级节点 116
相关 246
 概率 246
 绝对值 246
 描述性标签 246
 平均值节点 249
 统计量输出 246
 显著水平 246
箱线图 145
 示例 155
响应图 195, 201
小数符号 16
 平面文件导出节点 274
 数字显示格式 100
小数位
 显示格式 100
形状图形交叠 138
性能评估统计量 235
修改数据值 105
虚拟编码 122
旋转 3D 图形 140
选项
 IBM SPSS Statistics 259
选用板
 显示 214
 移动 214
 隐藏 214
选择行 (观测值) 52
选择节点
 从图形中生成 211
 从网络图形链接生成 190

选择值 204, 208, 210
循环时间元素
 自动数据准备 83
循环周期
 时间区间节点 129
训练样本
 分区数据 121
 平衡 57

[Y]

颜色图 145
 示例 159
颜色图形交叠 138
验证样本
 分区数据 121
样本节点
 层的样本大小 56
 抽样框 53
 非随机样本 53
 分层样本 53, 55, 56
 加权样本 55
 聚类样本 53, 55
 随机样本 53
 系统化样本 53
样式表
 重命名 163
 导出 163
 导入 163
 删除 163
业务规则
 评估图选项 200
移动平均值
 ARIMA 模型 74
引导
 导入文本文件 16
 用于数据库导出 262
应用程序示例 3
映射
 导出到 IBM Cognos TM1 的数据 278
映射字段 263
用户缺失值
 空白值 233
用户输入节点
 概述 45
 设置选项 45
用于汇总的标准差 58
用于汇总的方差值 58
用于汇总的关键值 58
用于汇总的计数值 58
用于汇总的平均值 58
用于汇总的四分位数值 58
用于汇总的中位数 58
用于汇总的最大值 58
用于汇总的最小值 58
有序数据 98

预设值, 数据库连接 12
阈值
 查看分级阈值 119
源节点
 变量文件节点 16
 概述 5
 固定文件节点 18
 实例化类型 49
 数据库源节点 10
 用户输入节点 45
 Analytic Server 源 24
 Excel 源节点 32
 IBM Cognos BI 源节点 28, 30, 31
 IBM Cognos TM1 源节点 31
 SAS 源节点 32
 Statistics 文件节点 284
 XML 源节点 33
 “模拟生成”节点 35, 36
 “企业视图”节点 6
元数据 96
 IBM SPSS Data Collection 源节点 20, 21
源文件变量
 IBM SPSS Data Collection 源节点 21
月数据
 时间区间节点 130

[Z]

摘要统计
 “数据审核”节点 237
折线图 145
 在地图上 145
真值 98
整数范围 97
整体节点
 输出字段 104
 综合得分 104
指定数据类型 79
指数平滑法
 时间序列模型标准 73
值
 字段和值标签 96
值标签
 Statistics 文件节点 284
直方图 145
 三维散点图 145
 示例 153
直方图节点 179
 使用图形 180
 “散点图”选项卡 179
 “外观”选项卡 179
直观表示
 边距 217
 编辑 213
 编辑模式 213

- 直观表示 (续)
 - 尺度 218
 - 点高宽比 216
 - 点形状 216
 - 点旋转 216
 - 复制 224
 - 划线 216
 - 类别 219
 - 面板 219, 221
 - 数字格式 217
 - 填充 217
 - 透明度 216
 - 图例位置 224
 - 图形和图表 137
 - 颜色和样式 216
 - 轴 218
 - 转换坐标系 221
 - 转置 219, 221
 - text 215
- 直观表示模板
 - 重命名 163
 - 导出 163
 - 导入 163
 - 删除 163
 - 位置 162
- 直观表示样式表
 - 重命名 163
 - 导出 163
 - 导入 163
 - 删除 163
 - 位置 162
 - 应用 225
- 执行
 - 指定顺序 299
- 执行顺序
 - 指定 299
- 滞后数据 134
- 质量报告
 - “数据审核”浏览器 241
- 质量浏览器
 - 生成过滤节点 242
 - 生成选择节点 242
- 置信度区间
 - 平均值节点 249
- 中位数
 - 统计量输出 246
- 种子值
 - 抽样与记录 121
- 周期
 - 时间区间节点 128
- 周期性
 - 时间序列建模器 75
 - 时间序列数据 125
- 周数据
 - 时间区间节点 131
- 逐行绑定 267
- 逐列绑定 267
- 主数据集 66
- 主要关键字段
 - 数据库导出节点 264
- 注解字符
 - 在变量文件中 16
- 注释
 - 用于超节点 296
- 专家建模器
 - 时间序列模型标准 73
- 转换
 - 重新编码 112, 114
 - reclassify 112, 114
- 转换测量级别 95
- 转换函数 75
 - 差分阶数 75
 - 分母阶数 75
 - 分子阶数 75
 - 季节性阶 75
 - 延迟 75
- 转置节点 124
 - 数字字段 124
 - 字段名 124
 - 字符串字段 124
- 转置数据 124
- 自动重新编码 112, 113
- 自动归类 93, 96
- 自动日期识别 16, 18
- 自动设置 214
- 自动数据准备
 - 标准化连续目标 84, 92
 - 操作详细信息 89
 - 操作摘要 88
 - 重置视图 86
 - 构建 85
 - 命名字段 85
 - 模型视图 86
 - 目标 81
 - 目标准备 84
 - 排除未使用字段 82
 - 排除字段 83
 - 派生节点生成 92
 - 视图间链接 86
 - 输入准备 84
 - 特征选择 85
 - 未使用字段排除 82
 - 预测能力 88
 - 准备目标 84
 - 准备日期和时间 83
 - 准备输入 84
 - 字段 82
 - 字段表 88
 - 字段处理摘要 86
 - 字段分析 87
 - 字段设置 82
 - 字段详细信息 89
- 自动数据准备节点 81
- 字段
 - 重新排序 135
 - 匿名化数据 110
 - 派生多个字段 107
 - 选择多个 107
 - 转置 124
 - 字段和值标签 96
- 字段操作节点 79
 - 由数据审核生成 243
- 字段重排节点 135
 - 设置选项 135
 - 自定义排序 135
 - 自动排序 135
- 字段存储
 - 转换 109
- 字段的方向 99
- 字段类型
 - 在直观表示中 142
- 字段名 102
 - 匿名化 103
 - 数据导出 262, 274, 279, 290
- 字段派生公式 107
- 字段属性 99
- 自回归
 - ARIMA 模型 74
- 自然对数变换
 - 时间序列建模器 75
- 自然顺序
 - 更改 135
- 自由度
 - 矩阵节点 234
 - 平均值节点 249
- 自由字段文本数据 16
- 综合数据
 - 用户输入节点 45
- 组合记录 69
 - 定制设置 70
- 组合数据 66
 - 来自多个文件 61
- 最大值
 - 设置全局量节点 251
 - 统计量输出 246
- 最佳线
 - 评估图选项 199
- 最小值
 - 设置全局量节点 251
 - 统计量输出 246
- 最优分级 118
- 坐标图 145
- 坐标系统
 - 转换 221

[数字]

- 2-D 点图 145
- 3D 图形 140
- 3-D 饼图 145
- 3-D 密度 145
- 3-D 面积图
 - description 145
- 3-D 散点图 145
- 3-D 条形图 145
- 3-D 直方图 145

A

- ADO 数据库
 - 导入 21
- Aggregate 节点
 - 概述 57
 - 设置选项 58
- Analytic Server 导出 276
- Analytic Server 源 24
- ANOVA
 - 平均值节点 248
- Append 节点
 - 概述 66
 - 设置选项 66
 - 字段匹配 66
- ARIMA 模型
 - 差分阶数 74
 - 常量 74
 - 季节性阶 74
 - 离群值 75
 - 时间序列模型标准 74
 - 移动平均值阶数 74
 - 转换函数 75
 - 自回归阶数 74

B

- Balance 节点
 - 概述 56
 - 设置选项 57
- Blank 函数
 - 填充时间序列 126

C

- CLEM 表达式 51
- Cognos, 请参阅 IBM Cognos BI 30
- costs
 - 评估图表 199
- CREATE INDEX 命令 266
- CRISP-DM
 - 数据理解 5

- CRISP-DM 过程模型
 - 数据准备 79
- CSV 数据
 - 导入 21

D

- DAT 文件
 - 保存 231
 - 导出 230, 279
- DPD 6

E

- employee_data.sav 数据文件 285
- EOL 字符 16
- ESRI 文件 164
- events
 - 创建 193
- Excel
 - 从 IBM SPSS Modeler 启动 279
- Excel 导出节点 279
- Excel 导入节点
 - 通过输出生成 279
- Excel 文件
 - 导出 279
- Excel 源节点 32
- extension
 - 派生的字段 107

F

- F 统计量
 - 平均值节点 249
- FILLFACTOR 关键字
 - 索引数据库表 266
- First 函数
 - 时间序列汇总 126

H

- hassubstring 函数 107
- HDATA 格式
 - IBM SPSS Data Collection 源节点 20
- holdout
 - 时间序列建模 127
- HTML
 - 保存输出 231
- HTML 输出
 - 报告节点 250
 - 在浏览器中查看 230

I

- IBM Cognos BI 导出节点 30, 276, 277
 - IBM Cognos BI 源节点 28, 30, 31
 - 导入报告 30
 - 导入数据 29
 - 图标 28
 - IBM Cognos TM1 导出节点 278
 - 导出数据 278
 - 映射导出数据 278
 - IBM Cognos TM1 源节点 31
 - 导入数据 31
 - IBM SPSS Collaboration and Deployment Services 存储库
 - 连接到 6
 - 用作直观表示模板、样式表和地图的位置 163
 - IBM SPSS Data Collection 导出节点 275
 - IBM SPSS Data Collection 调查数据
 - 导入 20, 21
 - IBM SPSS Data Collection 源节点 20, 21, 24
 - 标签类型 23
 - 多重响应集 23
 - 日志文件 21
 - 数据库连接设置 23
 - 元数据文件 21
 - language 23
 - IBM SPSS Modeler 1
 - 文档 3
 - IBM SPSS Modeler Server 1
 - IBM SPSS Statistics
 - 从 IBM SPSS Modeler 启动 259, 288, 290
 - 许可证位置 259
 - 有效字段名 291
 - IBM SPSS Statistics 节点 283
 - IBM SPSS Statistics 模型 287
 - 高级块详细信息 288
 - 关于 287
 - 模型块 288
 - 模型选项 287
 - IBM SPSS Statistics 输出节点
 - “输出”选项卡 289
 - IBM SPSS Statistics 数据文件
 - 导入调查数据 21
 - if-then-else 语句 109
 - In2data 数据库
 - 导入 21
- ## J
- jitter 213, 221

L

language

IBM SPSS Data Collection 源节点 23

Last 函数

时间序列汇总 126

LOESS 光滑线

散点图节点 172

lowess 光滑线 请参阅 LOESS 光滑线

散点图节点 172

M

map

带有点 145

带有箭头 145

交叠字段 145

具有饼图 145

具有条形图 145

具有折线图 145

颜色 145

Max 函数

时间序列汇总 126

MDD 文档

导入 21

Mean 函数

时间序列汇总 126

Microsoft Excel 源节点 32

Min 函数

时间序列汇总 126

mode

统计量输出 246

Mode 函数

时间序列汇总 126

N

n 中取 1 抽样 53

O

ODBC

批量加载方式 267, 268

数据库源节点 10

IBM Cognos BI 导出节点的连接 277

ODBC 导出节点。请参阅数据库导出节点
262

Oracle 10

P

P 值

重要性 248

parameters

超节点 297, 298

parameters (续)

超节点设置 297

节点属性 298

在 IBM Cognos BI 中 31

Pearson 卡方

矩阵节点 234

Pearson 相关性

平均值节点 249

统计量输出 246

performance

抽样数据 53

分级节点 119

合并 65

聚合节点 58

排序 61

派生节点 119

Python

批量加载脚本 267, 268

Q

Quancept 数据

导入 21

Quantum 数据

导入 21

Quanvert 数据库

导入 21

R

recency

设置相关日期 60

RFM 分析节点

分级值 120

概述 119

设置 119

RFM 汇总节点

概述 59

设置选项 60

S

SAS

设置导入选项 32

SAS 导出节点 279

SAS 源节点

传输文件 32

.sd2 (SAS) 文件 32

.ssd (SAS) 文件 32

.tpt (SAS) 文件 32

shapefile 164

smoother

散点图节点 172

SMZ 文件

编辑预先安装的 SMZ 文件 164

SMZ 文件 (续)

重命名 163

创建 164

导出 163

导入 163

概述 164

删除 163

预先安装 164

SPLOM 145

示例 158, 160

SQL 查询

数据库源节点 10, 11, 15

Statistics 导出节点 290

“导出”选项卡 290

Statistics 输出节点 288

“语法”选项卡 288

Statistics 文件节点 284

Statistics 转换节点 285

设置选项 285

允许的语法 285

“语法”选项卡 285

storage 96

转换 109, 110

Sum 函数

时间序列汇总 126

Surveycraft 数据

导入 21

T

t 检验

成对样本 248

独立样本 248

平均值节点 248, 249

text

分隔 16

数据 16, 18

TimeIndex 字段

时间区间节点 126

TimeLabel 字段

时间区间节点 126

Triple-S 数据

导入 21

U

UNIQUE 关键字

索引数据库表 266

V

values

读取 96

指定 96

VDATA 格式
IBM SPSS Data Collection 源节点 20

W

Web 节点 186
调整点 190
调整阈值 192
定义链接 188
更改布局 190
滑块 190
链接滑块 190
使用图形 190
网络汇总 193
“散点图”选项卡 187
“外观”选项卡 189
“选项”选项卡 188

X

XLS 文件
导出 279
XML 导出节点 280
XML 输出
报告节点 250
XML 源节点 33
XPath 语法 33

“模拟拟合”节点 (续)
输出设置 253
“设置”选项卡 253
“模拟评估”节点 254, 256, 257, 258
输出设置 254
“设置”选项卡 254
“模拟生成”节点
概述 35
设置选项 36
“企业视图”节点 6
“数据审核”节点 237
“设置”选项卡 238
“输出”选项卡 231
“数据视图”节点 9
设置选项 9
“选择”节点
概述 52
“语法”选项卡
Statistics 输出节点 288
“追加”节点
标记字段 65
“最近均值”函数
填充时间序列 126
“最近”函数
填充时间序列 126

[特别字符]

.sav 文件 284
.sd2 (SAS) 文件 32
.slb 文件 299
.ssd (SAS) 文件 32
.tpt (SAS) 文件 32
.zsav 文件 284
“分析”节点 235
分析选项卡 235
“输出”选项卡 231
“合并”节点
标记字段 65
概述 61
过滤字段 64
“空间时间限制”节点
定义密度 77
概述 76
“类型”节点
清除值 48
“流式 TS”节点
部署选项 76
概述 71
模型选项 72
字段选项 71
“模拟拟合”节点 252
分布拟合 252



Printed in China