Узлы источников, обработки и вывода IBM SPSS Modeler 18.0

Примечание

Прежде чем использовать эту информацию и продукт, описанный в ней, прочтите сведения в разделе “Уведомления” на стр. 371.

Информация о продукте

Это издание применимо к версии 18, выпуск 0, модификация 0 IBM SPSS Modeler и ко всем последующим версиям и модификациям до тех пор, пока в новых изданиях не будет указано иное.

Содержание

Предисловие

IBM® SPSS Modeler - это инструментальная среда исследования данных IBM Corp., рассчитанная на работу с предприятием. SPSS Modeler помогает организациям улучшить взаимосвязи с клиентами и отдельными лицами, обеспечивая глубокое понимание данных. Организации используют приобретенные с помощью SPSS Modeler глубокие знания для сохранения выгодных заказчиков, обнаружения возможностей дополнительных покупок, привлечения новых клиентов, обнаружения ошибок, сокращения рисков и улучшений в обеспечении государственных служб.

Наглядный интерфейс SPSS Modeler дает пользователям возможность применить свой конкретный опыт в бизнесе, что способствует разработке более мощных предсказывающих моделей и сокращает время принятия решения. SPSS Modeler предлагает много способов моделирования, таких как алгоритмы предсказания, классификации, сегментации и ассоциативного обнаружения. Когда моделей IBM SPSS Modeler Solution Publisher поддерживает их распространение на уровне организации для принимающих решение сотрудников или для применения к базе данных.

О бизнес аналитике IBM

Программное обеспечение IBM для бизнес аналитики предоставляет полную, последовательную и точную информацию, которая повышает эффективность ведения бизнеса. Полный набор программного обеспечения для business intelligence, прогностической аналитики, управления финансовой эффективностью и стратегией и аналитических приложений позволяет ясно видеть текущую ситуацию, а также делать прогнозы, позволяющие предпринимать практические действия. В сочетании с решениями для конкретных отраслей, проверенной практикой и услугами бизнес аналитика IBM позволяет организациям любых размеров достигать наивысшей производительности, уверенно автоматизировать процессы принятия решений и добиться лучших результатов.

Как составная часть этого набора, программное обеспечение IBM SPSS Predictive Analytics помогает организациям предсказывать будущие события и предпринимать практические действия непосредственно на основе этих предсказаний. Коммерческие, правительственные и научные организации всего мира, полагаются на технологию IBM SPSS, обеспечивающую конкурентное преимущество в привлечении, удержании клиентов и повышения отдачи от них при уменьшении доли ошибочных решений и сокращении рисков. Включая программное обеспечение IBM SPSS в свои ежедневные операции, организации могут прогнозировать будущие события, направлять и автоматизировать решения для соответствия бизнес-целям и достигать ощутимых конкурентных преимуществ. Чтобы получить дальнейшую информацию или связаться с представителем, зайдите на http://www.ibm.com/spss.

Техническая поддержка

Техническая поддержка предоставляется клиентам, оплачивающим обновительные взносы. Пользователи могут обращаться в службу технической поддержки, если у них возникают какие-либо проблемы с использованием или установкой программного обеспечения IBM Corp.. К службе технической поддержки можно вызывать через сайт IBM Corp. по адресу http://www.ibm.com/support. При обращении за поддержкой будьте готовы назвать себя и организацию, в которой вы работаете.

Глава 1. О программе IBM SPSS Modeler

IBM SPSS Modeler - это комплект инструментов исследования данных, при помощи которого можно быстро разрабатывать прогнозные модели, использующие деловые знания и опыт, и внедрять их в деловые операции для усовершенствования процесса принятия решений. Разработанный на основе модели промышленного стандарта CRISP-DM, IBM SPSS Modeler поддерживает весь процесс исследования данных, от обработки исходных данных до получения лучших деловых результатов.

IBM SPSS Modeler предлагает ряд методов моделирования, взятых из таких областей, как обучение машин, искусственный интеллект и статистика. При помощи методов, доступных на палитре Моделирование, можно извлечь новую информацию из данных и разработать прогнозные модели. У каждого из методов есть свои сильные стороны и типы задач, для решения которых он лучше всего подходит.

SPSS Modeler можно приобрести как отдельный продукт или использовать как клиент в сочетании с SPSS Modeler Server. Кроме того, доступен ряд дополнительных возможностей, сводка которых дается в следующих разделах. Дополнительную информацию смотрите по адресу http://www.ibm.com/software/analytics/spss/products/modeler/.

Продукты IBM SPSS Modeler

В семейство продуктов IBM SPSS Modeler и связанные с этим семейством программы входят следующие продукты:

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- Адаптеры IBM SPSS Modeler Server для IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler - это полнофункциональная версия продукта, устанавливаемая и запускаемая на персональном компьютере. SPSS Modeler можно запустить в локальном режиме, как автономный продукт, или в распределенном режиме вместе с IBM SPSS Modeler Server, чтобы повысить производительность на больших наборах данных.

Используя SPSS Modeler, можно быстро и интуитивно строить точные прогнозные модели, не прибегая к программированию. Используя уникальный визуальный интерфейс, можно легко визуализировать процесс анализа данных. В продукт встроены расширенные функции аналитики, при поддержке которых можно обнаруживать в данных скрытые структуры и тенденции. Можно моделировать результаты и выяснять, какие факторы на них влияют, чтобы полностью использовать деловые возможности и ограничивать риски.

SPSS Modeler доступен в двух версиях: SPSS Modeler Professional и SPSS Modeler Premium. Дополнительную информацию смотрите в разделе “Выпуски IBM SPSS Modeler” на стр. 2.

IBM SPSS Modeler Server

SPSS Modeler пользуется архитектурой клиент - сервер, чтобы распределять требования ресурсоемких операций по мощным серверным программам, что повышает производительность для больших наборов данных.

SPSS Modeler Server - это отдельно лицензируемый продукт, который непрерывно работает в режиме распределенного анализа на хосте сервера совместно с одной или несколькими установками IBM SPSS Modeler. При этом SPSS Modeler Server обеспечивает высокую производительность для больших наборов данных, поскольку ресурсоемкие операции можно выполнять на сервере без скачивания данных на компьютер клиента. Кроме того, IBM SPSS Modeler Server обеспечивает поддержку для возможностей оптимизации SQL и моделирования в базе данных, что дает дополнительный выигрыш в производительности и автоматизации.

IBM SPSS Modeler Administration Console

Modeler Administration Console - это графическая программа для управления многочисленными опциями конфигурации SPSS Modeler Server, который также можно конфигурировать посредством файла опций. Эта прикладная программа содержит консольный пользовательский интерфейс для отслеживания и конфигурирования установок SPSS Modeler Server installations, and is available free-of-charge SPSS Modeler Server. Эту прикладную программу можно установить только на компьютерах Windows; однако она может управлять сервером на любой поддерживаемой платформе.

IBM SPSS Modeler Batch

Хотя обычно исследование данных - интерактивный процесс, можно также запустить SPSS Modeler из командной строки, не открывая графический интерфейс. Например, у вас могут быть продолжительные или повторяющиеся задачи, которые желательно выполнить без участия пользователя. SPSS Modeler Batch - это особая версия продукта, предоставляющая поддержку всех аналитических возможностей SPSS Modeler без вызова обычного пользовательского интерфейса. SPSS Modeler Server необходим для использования SPSS Modeler Batch.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher - это инструмент, при помощи которого можно создать пакетную версию потока SPSS Modeler; такую версию можно запускать внешним механизмом времени выполнения или встроить во внешнюю прикладную программу. Этим способом можно публиковать и внедрять полные потоки SPSS Modeler для использования в средах, где SPSS Modeler не установлен. SPSS Modeler Solution Publisher распространяется в составе службы IBM SPSS Collaboration and Deployment Services - Scoring, для которой требуется отдельная лицензия. С этой лицензией вы получаете SPSS Modeler Solution Publisher Runtime, при помощи которого можете запускать опубликованные потоки.

Дополнительную информацию о SPSS Modeler Solution Publisher смотрите в документации IBM SPSS Collaboration and Deployment Services. Центр знаний IBM SPSS Collaboration and Deployment Services содержит разделы "IBM SPSS Modeler Solution Publisher" и "IBM SPSS Analytics Toolkit".

Адаптеры IBM SPSS Modeler Server для IBM SPSS Collaboration and Deployment Services

Для IBM SPSS Collaboration and Deployment Services доступен ряд адаптеров, при посредстве которых SPSS Modeler и SPSS Modeler Server могут взаимодействовать с репозиторием IBM SPSS Collaboration and Deployment Services. При этом поток SPSS Modeler, внедренный в репозиторий, доступен для совместного использования несколькими пользователями или для обращения из прикладной программы IBM SPSS Modeler Advantage тонкого клиента. Адаптер устанавливается в той системе, в которой находится репозиторий.

Выпуски IBM SPSS Modeler

SPSS Modeler доступен в следующих выпусках.

SPSS Modeler Professional

SPSS Modeler Professional содержит все инструменты, необходимые для работы с большинством типов структурированных данных, таких как трассировка поведения и взаимодействия в системах CRM,

демографии, поведения покупателей и данных о продажах.

SPSS Modeler Premium

SPSS Modeler Premium - это отдельно лицензируемый продукт, расширяющий SPSS Modeler Professional для работы с такими специальными данными, как данные в аналитике объектов или социальных сетях, и с неструктурированными текстовыми данными. SPSS Modeler Premium состоит из следующих компонентов.

IBM SPSS Modeler Entity Analytics добавляет дополнительное измерение к прогностической аналитике IBM SPSS Modeler. Прогностическая аналитика пытается предсказать будущее поведение данных из прошлого, а объектная аналитика направлена на улучшение связности и согласованности текущих данных посредством устранения конфликтов идентичности в самих записях. Идентичность может относиться к индивидууму, организации, а также к любому другому объекту, для которого возможна неоднозначность. Разрешение идентичности может оказаться крайне необходимым для ряда полей, в том числе для управления отношениями с клиентами, обнаружения мошенничества, противодействия отмыванию денег или для национальной и международной безопасности.

IBM SPSS Modeler Social Network Analysis преобразует информацию о взаимосвязях в поля, характеризующие социальное поведение отдельных лиц и групп. При помощи данных, описывающих взаимосвязи, в основе которых лежат социальные сети, IBM SPSS Modeler Social Network Analysis определяет социальных лидеров, влияющих на поведение других участников сети. Кроме того, вы можете определить, какие люди наиболее подвержены влиянию других участников сети. Сочетая полученные результаты с результатами других измерений, можно создать исчерпывающие профили отдельных лиц, на которых будут основаны ваши прогнозные модели. Модели, содержащие эту социальную информацию, выполняются лучше моделей, которые ее не содержат.

IBM SPSS Modeler Text Analytics использует новейшие лингвистические технологии и обработку естественного языка (NLP) для быстрой обработки самых разнообразных неструктурированных текстовых данных, для извлечения и организации ключевых понятий и группирования этих понятий в категории. Извлеченные понятия и категории можно сочетать с существующими структурированными данными, такими как демографические, и применять к моделированию при помощи полного комплекта инструментов исследования данных IBM SPSS Modeler для получения более качественных и специализированных решений.

Документация IBM SPSS Modeler

Документация в формате электронной справки доступна через меню Справка SPSS Modeler. Сюда входит документация по SPSS Modeler, SPSS Modeler Server, а также Руководство по прикладным программам (Учебник) и другие сопроводительные материалы.

Полная документация по каждому продукту (включая указания по установке) доступна в формате PDF в отдельной сжатой папке как часть скачиваемого образа продукта. Документы в формате PDF также можно скачать с сайта по адресу http://www.ibm.com/support/docview.wss?uid=swg27046871.

Кроме того, документация в обоих этих форматах доступна в Центре знаний SPSS Modeler по адресу http://www-01.ibm.com/support/knowledgecenter/SS3RA7_18.1.0.

Документация SPSS Modeler Professional

В комплект документации SPSS Modeler Professional (включая указания по установке) входят:

- **IBM SPSS Modeler Руководство пользователя.** Общее введение в использование SPSS Modeler, в том числе о создании потоков данных, обработке пропущенных значений, построению выражений CLEM, работе с проектами и отчетами и составлению пакетов потоков для внедрения в IBM SPSS Collaboration and Deployment Services или IBM SPSS Modeler Advantage.
- **Узлы источников, обработки и вывода IBM SPSS Modeler.** Описания всех узлов, служащих для чтения, обработки и вывода данных в различных форматах. По существу это все узлы, кроме узлов моделирования.

- **Узлы моделирования IBM SPSS Modeler.** Описания всех узлов, служащих для создания моделей исследования данных. IBM SPSS Modeler предлагает ряд методов моделирования, взятых из таких областей, как обучение машин, искусственный интеллект и статистика.
- **Руководство по алгоритмам IBM SPSS Modeler.** Описание математических основ методов моделирования, используемых в IBM SPSS Modeler. Это руководство доступно только в формате PDF.
- **Руководство по прикладным программам IBM SPSS Modeler.** Примеры в этом руководстве служат кратким специализированным введением к тем или иным методам и технологиям моделирования. Это руководство доступно также в электронном виде в меню Справка. Дополнительную информацию смотрите в разделе “Примеры прикладных программ”.
- **Сценарии и автоматизация Python IBM SPSS Modeler.** Информация об автоматизации системы путем создания сценариев Python, включая сценарии свойств, которые могут использоваться для работы с узлами и потоками.
- **Руководство по внедрению IBM SPSS Modeler .** Информация о выполнении IBM SPSS Modeler потоков как шагов обработки заданий под управлением IBM SPSS Collaboration and Deployment Services Deployment Manager.
- **Руководство разработчика IBM SPSS Modeler CLEF .** CLEF предоставляет возможности интеграции с программами других производителей, таких как подпрограммы обработки данных или алгоритмы моделирования, как с узлами в IBM SPSS Modeler.
- **Руководство по исследованию данных в базе данных IBM SPSS Modeler.** Информация о том, как использовать мощности вашей базы данных для повышения производительности и расширения диапазона возможностей анализа с привлечением алгоритмов от сторонних производителей.
- **Руководство администратора и руководство по производительности IBM SPSS Modeler Server .** Информация о том, как сконфигурировать и администрировать IBM SPSS Modeler Server.
- **Руководство пользователя по консоли администратора IBM SPSS Modeler .** Информация об установке и использовании пользовательского интерфейса консоли для мониторинга и конфигурирования IBM SPSS Modeler Server. Консоль реализована как подключаемый модуль прикладной программы Deployment Manager.
- **Руководство по CRISP-DM IBM SPSS Modeler.** Пошаговое руководство к использованию методологии CRISP-DM для исследования данных SPSS Modeler.
- **IBM SPSS Modeler Batch Руководство пользователя.** Полное руководство по использованию IBM SPSS Modeler в пакетном режиме, включая подробности выполнения в пакетном режиме и аргументы командной строки. Это руководство доступно только в формате PDF.

Документация SPSS Modeler Premium

В комплект документации SPSS Modeler Premium (включая указания по установке) входят:

- **IBM SPSS Modeler Entity Analytics Руководство пользователя.** Информация об использовании аналитики объектов совместно с SPSS Modeler, в том числе по установке и конфигурированию репозитория, узлам аналитики объектов и задачам управления.
- **IBM SPSS Modeler Social Network Analysis Руководство пользователя.** Руководство по выполнению анализа социальной сети совместно с SPSS Modeler, включая анализ групп и анализ распространения.
- **SPSS Modeler Text Analytics Руководство пользователя.** Информация об использовании аналитики текстов совместно с SPSS Modeler, в том числе по узлам исследования текстов, интерактивной инструментальной среде, шаблонам и другим ресурсам.

Примеры прикладных программ

Инструменты исследования данных в SPSS Modeler помогают разрешить широкий спектр деловых и организационных проблем, а примеры прикладных программ предоставляют краткие, целевые введения в конкретные методы и способы моделирования. Используемые здесь наборы данных намного меньше огромных складов данных, которыми управляют некоторые исследователи данных, но применяемые понятия и методы должны масштабироваться до реальных прикладных программ.

Чтобы обратиться к примерам, выберите **Примеры прикладных программ** в меню Справка в SPSS Modeler.

Файлы данных и потоки примеров устанавливаются в папке Demos в каталоге установки продукта. Дополнительную информацию смотрите в разделе “Папка demos”.

Примеры моделирования баз данных. Смотрите эти примеры в руководстве *IBM SPSS Modeler In-Database Mining Guide*.

Примеры сценариев. Смотрите эти примеры в руководстве *IBM SPSS Modeler Scripting and Automation Guide*.

Папка demos

Файлы данных и примеры потоков, используемые с примерами прикладных программ, устанавливаются в папке Demos в каталоге установки продукта (например: C:\Program Files\IBM\SPSS\Modeler\<версия>\Demos). К этой папке можно также обратиться из группы программ IBM SPSS Modeler в меню Пуск Windows или, щелкнув по Demos в списке недавно использовавшихся каталогов в диалоговом окне **Файл > Открыть поток**.

Отслеживание лицензий

При работе с SPSS Modeler использование лицензий отслеживается и записывается в журнал через регулярные интервалы времени. В журнал записываются показатели лицензирования *AUTHORIZED_USER* и *CONCURRENT_USER*; тип записываемого в журнал показателя зависит от типа лицензии, которая у вас есть для SPSS Modeler.

Генерируемые файлы журналов могут обрабатываться инструментом IBM License Metric Tool, из которого вы можете сгенерировать отчеты об использовании лицензий.

Файлы журналов лицензирования создаются в том же каталоге, куда записываются и файлы журналов клиента SPSS Modeler (по умолчанию %ALLUSERSPROFILE%/9IBM/SPSS/Modeler/<версия>/log).

Глава 2. Узлы источников

Обзор

Узлы источников позволяют импортировать данные, хранящиеся в разных форматах, включая плоские файлы, IBM SPSS Statistics (*.sav*), SAS, Microsoft Excel и совместимые с ODBC реляционные базы данных. Можно также сгенерировать синтетические данные при помощи узла пользовательского ввода.

Палитра Источники содержит следующие узлы:

При помощи источника Analytic Server поток можно выполнить в файловой системе HDFS (Hadoop Distributed File System). Информация в источник данных Analytic Server может поступать из разнообразных мест, таких как текстовые файлы и базы данных. Дополнительную информацию смотрите в разделе “Исходный узел Analytic Server” на стр. 12.

Узел базы данных можно использовать для импорта данных из ряда других пакетов при помощи ODBC (Open Database Connectivity), включая Microsoft SQL Server, DB2, Oracle и другие. Дополнительную информацию смотрите в разделе “Узел источника базы данных” на стр. 17.

Узел файлов переменных читает данные из текстовых файлов в формате со свободными полями, то есть, файлов, записи которых содержат постоянное число полей, но изменяющееся число символов. Этот узел полезен также для файлов с текстом заголовков фиксированной длины и содержащих определенные типы аннотаций. Дополнительную информацию смотрите в разделе “Узел файла переменных” на стр. 26.

Узел фиксированных файлов импортирует данные из текстовых файлов в формате с полями фиксированной ширины, то есть, с полями без разделителей, но с началом в одной и той же позиции и фиксированной длины. В формате с полями фиксированной ширины хранятся данные, генерируемые компьютером или представленные в устаревшем формате. Дополнительную информацию смотрите в разделе “Узел фиксированных файлов” на стр. 30.

Узел Файл статистики читает данные в формате файлов *.sav*, используемом IBM SPSS Statistics, а также файлы кэша, сохраненные IBM SPSS Modeler, которые также используют этот формат.

Узел Data Collection импортирует данные опросов из различных форматов, используемых программами маркетинговых исследований, соответствующими Data Collection Data Model. Для использования этого узла должна быть установлена библиотека разработчика Data Collection. Дополнительную информацию смотрите в разделе “Узел Data Collection” на стр. 32.

Узел источника IBM Cognos BI импортирует данные из баз данных Cognos BI.

Узел источника IBM Cognos TM1 импортирует данные из баз данных Cognos TM1.

Узел файлов SAS импортирует данные SAS в IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “Узел источника SAS” на стр. 41.

Узел Excel импортирует данные из Microsoft Excel в формате файлов .xlsx. Источник данных ODBC не требуется. Дополнительную информацию смотрите в разделе “Узел источника Excel” на стр. 42.

Узел источника XML импортирует данные в формате XML в поток. Вы можете импортировать в каталог один файл или все файлы. Дополнительно вы можете задать файл схемы, из которой можно прочесть структуру XML.

Узел пользовательского ввода предоставляет простой способ создания синтетических данных - либо с чистого листа, либо путем изменения существующих данных. Этот способ полезен, например, если вы хотите создать тестовый набор данных для моделирования. Дополнительную информацию смотрите в разделе “Узел пользовательского ввода” на стр. 45.

Узел генерирования имитации обеспечивает удобный путь сгенерировать имитационные данные - либо с нуля, используя указанные пользователем статистические распределения, либо автоматически, используя распределения, полученные при выполнении узла подгонки имитации для существующих данных хронологии. Это полезно, когда нужно оценить вывод прогнозной модели при наличии неопределенности во входных данных модели.

Узел Представление данных можно использовать для доступа к источникам данных, определенных в аналитических представлениях данных IBM SPSS Collaboration and Deployment Services. Аналитическое представление данных определяет стандартный интерфейс для доступа к данным и связывает несколько физических источников данных с этим интерфейсом. Дополнительную информацию смотрите в разделе “Узел Представление данных” на стр. 62.

Узел Геопространственный источник используется для переноса карты или пространственных данных в сеанс исследования данных. Дополнительную информацию смотрите в разделе “Геопространственный узел источника” на стр. 64.

Чтобы начать поток, добавьте на холст потока узел источника. Далее щелкните дважды по узлу, чтобы открыть его диалоговое окно. Различные вкладки в этом диалоговом окне позволяют считывать данные, просматривать поля и значения и задавать разнообразные опции, в том числе для фильтров, типов данных, ролей полей и проверки пропущенных значений.

Задание форматирования и системы хранения для полей

Опции на вкладке Данные для узлов Фиксированный файл, Файл переменных, Источник XML и Пользовательский ввод позволяют задать тип хранения для полей при их импорте или создании в IBM SPSS Modeler. Для узлов Фиксированный файл, Файл переменных и Пользовательский ввод вы можете задать также форматирование полей и другие метаданные.

Для прочитанных из других источников данных система хранения определяется автоматически, но ее можно изменить при помощи функции преобразований, такой как `to_integer`, на узле фильтрации или извлечения.

Поле Используйте столбец **Поле** для просмотра и выбора полей в текущем наборе данных.

Переопределить Включите переключатель в столбце **Переопределить**, чтобы активировать опции в столбцах **Система хранения** и **Формат ввода**.

Хранение данных

Хранение описывает способ хранения данных в поле. Например, в поле со значениями 1 и 0 хранятся целочисленные данные. В этом оно отличается от шкалы измерений, которая описывает использование данных и на хранение не влияет. Например, вы можете захотеть задать шкалу измерения для целочисленного поля со значениями 1 и 0 как для *флага*. Обычно это означает, что 1 = *True*, а 0 = *False*. Хранение должно быть определено на источнике, тогда как шкалу измерения можно изменить при помощи узла Тип в любой точке в потоке. Дополнительную информацию смотрите в разделе “Уровни измерения” на стр. 129.

Доступны следующие типы хранения:

- **Строка** Используется для полей, содержащих нечисловые данные, они называются также алфавитно-цифровыми. Строка может содержать любую последовательность символов, например: *поле*, *Класс 2* или *1234*. Имейте в виду, что числа в строках нельзя использовать в вычислениях.
- **Целое** Поле с целочисленными значениями.
- **Действительное** Значения представляют собой числа с дробной частью (не только целые). Формат вывода задается в диалоговом окне Свойства потока и может быть переопределен для отдельных полей на узле Тип (вкладка Формат).
- **Дата** Значения дат, задаваемые в стандартном формате, таком как год, месяц и день (например: `2007-09-26`). Конкретный формат задается в диалоговом окне Свойства потока.
- **Время** Время, измеряемое как продолжительность. Например, вызов службы, продолжающийся 1 час, 26 минут и 38 секунд, может быть представлен как `01:26:38`, в зависимости от текущего формата времени, заданного в диалоговом окне Свойства потока.
- **Отметка времени** Значения, содержащие составляющие даты и времени, например: `2007-09-26 09:04:00` (тоже в зависимости от текущих форматов даты и времени в диалоговом окне Свойства потока). Имейте в виду, что значения отметок времени может потребоваться заключить в двойные кавычки, чтобы они интерпретировались как одно значение, а не как отдельные значения даты и времени. (Это применяется, например, при вводе значений на узле пользовательского ввода.)
- **Список** Введенное в SPSS Modeler версии 17 вместе с новыми шкалами измерений, Геопространственная и Собрание, поле хранения Список содержит несколько значений для одной записи. Существуют списочные версии всех других типов хранения.

Таблица 1. Значки типов хранения списков

Значок	Тип хранения
[A]	Список строковых
[◇]	Список целочисленных
[◈]	Список действительных
[🕒]	Список переменных времени
[▦]	Список переменных даты

Таблица 1. Значки типов хранения списков (продолжение)

Значок	Тип хранения
[]	Список переменных отметки времени
[[]]	Введите значение большее нуля

Кроме этого, для использования со шкалой измерения Собрание существуют списочные версии следующих шкал измерения.

Таблица 2. Значки уровня измерения списков

Значок	Шкала измерения
[]	Список количественных
[]	Список категориальных
[]	Список флагов
[]	Список номинальных
[]	Список порядковых

Списки можно импортировать в SPSS Modeler на одном из трех узлов источников (Analytic Server, Геопространственный или Файл переменных) или создать в вашем потоке с помощью узлов операций с полями Извлечение или Заполнитель.

Дополнительную информацию о списках и их взаимодействии со шкалами измерений Собрание и Геопространственная смотрите в разделе “Система хранения списков и связанные шкалы измерений” на стр. 11

Преобразования хранения. Можно преобразовать хранение для поля при помощи разнообразных функций преобразования, таких как `to_string` и `to_integer`, на узле заполнения. Дополнительную информацию смотрите в разделе “Преобразование хранения при помощи узла заполнения” на стр. 152. Имейте в виду, что функции преобразования (и любые другие функции, которым требуется конкретный тип ввода, такой как значение даты и времени) зависят от текущих форматов, заданных в диалоговом окне Свойства потока. Например, если вы хотите преобразовать строковое поле со значениями *Jan 2003*, *Feb 2003* (и так далее) в хранение даты, в качестве формата дат по умолчанию для потока выберите **МЕС ГГГГ**. Функции преобразования доступны также с узла извлечения, для временного преобразования во время вычисления операции извлечения. При помощи узла извлечения можно выполнять и другие операции с данными, такие как перекодирование строковых полей с категориальными значениями. Дополнительную информацию смотрите в разделе “Перекодирование значений при помощи узла извлечения” на стр. 151.

Чтение данных смешанных типов. Имейте в виду, что при чтении значений в полях с числовым хранением (целых чисел, действительных чисел, времени, отметок времени или дат) все нечисловые значения задаются как пустые или системные пропущенные. Это связано с тем, что в отличие от некоторых прикладных программ IBM SPSS Modeler не разрешает смешанные типы хранения в поле. Во избежание этого все поля с данными смешанных типов следует считывать как строки, изменив тип хранения либо на узле источника, либо во внешней прикладной программе, как требуется.

Формат ввода в поле (только для узлов Фиксированный файл, Файл переменных и Пользовательский ввод)

Для всех типов хранения, кроме строкового и целочисленного, можно задать опции форматирования для выбранных полей, используя раскрывающийся список. Например, при слиянии данных от различных локалей вам может потребоваться задать точку (.) в качестве десятичного разделителя в одном поле, а в другом для разделителя потребуется запятая.

Опции ввода, заданные на узле источника, перезаписывают опции форматирования, заданные в диалоговом окне свойств потока; однако в дальнейшем они не сохранятся в потоке. Эти опции предназначены для правильного синтаксического анализа входных данных на основе вашей информации об этих данных. Заданные форматы используются как руководство для синтаксического анализа данных, когда они считываются в IBM SPSS Modeler, а не для определения их форматирования уже после их чтения в IBM SPSS Modeler. Чтобы задать форматирование по конкретным полям где-то еще в потоке, используйте вкладку Формат узла Тип. Дополнительную информацию смотрите в разделе “Вкладка Параметры форматов полей” на стр. 140.

Опции изменяются в зависимости от типа хранения. Например, для типа хранения Real можно выбрать **точку (.)** или **запятую (,)** в качестве десятичного разделителя. Для полей отметки времени открывается отдельное диалоговое окно, где из раскрывающегося списка выбирается пункт **Задать**. Дополнительную информацию смотрите в разделе “Задание опций форматирования полей” на стр. 141.

Для всех типов хранения можно выбрать также опцию **Значения по умолчанию потока**, чтобы использовать при импорте настройки потока по умолчанию. Настройки потока задаются в диалоговом окне свойств потока.

Дополнительные опции

На вкладке Данные можно задать несколько других опций:

- Чтобы просмотреть параметры хранения для данных, которые больше не соединены через текущий узел (например, данные обучения), выберите **Просмотр неиспользуемых параметров полей**. Очистить устаревшие поля можно нажатием кнопки **Очистить**.
- При работе в этом диалоговом окне в любой момент можно нажать кнопку **Обновить**, чтобы перезагрузить поля из источника данных. Это полезно при изменении соединений данных с узлом источника или при работе с несколькими вкладками в этом диалоговом окне.

Система хранения списков и связанные шкалы измерений

Введенное в SPSS Modeler версии 17 вместе с новыми шкалами измерений, Геопространственная и Собрание, поле хранения Список содержит несколько значений для одной записи. Списки заключаются в квадратные скобки ([]). Примерами списков могут быть: `[1,2,4,16]` и `["abc", "def"]`.

Списки можно импортировать в SPSS Modeler на одном из трех узлов источников (Analytic Server, Геопространственный или Файл переменных), создать в вашем потоке с помощью узлов операций с полями Извлечение или Заполнитель или сгенерировать на узле Слияние с использованием ранжированных условий слияния.

Для списков определена глубина; например, простой список с элементами внутри одной пары квадратных скобок в формате `[1,3]` записывается в IBM SPSS Modeler с глубиной, равной нулю. В дополнение к простым спискам нулевой глубины можно использовать вложенные списки, где каждое значение в списке само представляет собой список.

Глубина вложенного списка зависит от связанной шкалы измерений. Для данных типа Без типа ограничений на глубину списка нет, для Собрания глубина равна нулю, для Геопространственного типа глубина может быть от нуля до двух включительно в зависимости от числа вложенных элементов.

Для списков нулевой глубины шкалу измерений можно задать как Собрание или Геопространственная. Обе эти шкалы - это родительские шкалы измерений, и в диалоговом окне Значения задается информация о подуровнях шкал измерения. Подуровень шкалы измерения в списке Собрание определяет шкалу измерения элементов в этом списке. Все шкалы измерений (кроме типов Без типа и Геопространственный) доступны как подуровни для Собраний. У Геопространственной шкалы измерений есть шесть подуровней - Точка, Ломаная, Многоугольник, Несколько точек, Мультиломаная и Мультиполигон; дополнительную информацию смотрите в разделе “Подуровни геопространственных измерений” на стр. 131.

Примечание: Шкала измерений Собрание может использоваться только со списками глубина 0, шкала измерений Геопространственная - со списками глубины не более 2, а шкала Без типа - с любой глубиной списка.

В следующем примере показано различие между списком нулевой глубины и вложенным списком на примере структуры двух подуровней шкалы измерения Геопространственная - Точка и Ломаная:

- У подуровня Точка шкалы измерения Геопространственная глубина поля равна нулю:
  `[1,3]` - две координаты
  `[1,3,-1]` - три координаты
- У подуровня Ломаная шкалы измерения Геопространственная глубина поля равна единице:
  `[ [1,3], [5,0] ]` - две координаты
  `[ [1,3,-1], [5,0,8] ]` - три координаты

Поле Точка с нулевой глубиной - это обычный список, в котором каждое значение составляется из двух или трех координат. Поле Ломаная с глубиной единица - это список точек, где каждая точка состоит из внутреннего ряда значений списка.

Дополнительную информацию о создании списков смотрите в разделе “Извлечение геопространственного поля или поля списка” на стр. 148.

Неподдерживаемые управляющие символы

Некоторые из процессов в SPSS Modeler не могут обрабатывать данные, включающие определенные управляющие символы. Если в ваших данных используются такие символы, может появиться примерно такое сообщение об ошибке:

```
В значениях поля {0} обнаружены несовместимые управляющие символы
```

Не поддерживаются следующие символы: от `0x0` до `0x3F` включительно и `0x7F`; однако символы табуляции (`0x9(\t)`), новой строки (`0xA(\n)`) и возврата каретки (`0xD(\r)`) проблем не вызывают.

Если появится сообщение об ошибке, относящееся к неподдерживаемым символам, после узла Источник используйте узел Filler и выражение CLEM **`stripctrlchars`** для замены этих символов.

Исходный узел Analytic Server

При помощи источника Analytic Server поток можно выполнить в файловой системе HDFS (Hadoop Distributed File System). Информация в источник данных Analytic Server может поступать из разнообразных мест, в том числе:

- Текстовые файлы в HDFS
- Базы данных
- HCatalog

Обычно поток с источником Analytic Server будет выполняться в HDFS; но если поток содержит узел, не поддерживаемый для выполнения в HDFS, максимально большая часть потока будет "вытолкнута обратно"

в Analytic Server, и затем SPSS Modeler Server попытается обработать остаток потока. Вам потребуется произвести подвыборку из очень больших баз данных; например, разместив узел Выборка в потоке.

Источник данных. Предполагая, что ваш администратор SPSS Modeler Server установил соединение, вы выбираете источник данных, содержащий нужные для использования данные. Источник данных содержит файлы и метаданные, связанные с этим источником. Нажмите кнопку **Выбрать**, чтобы вывести список доступных источников данных. Дополнительную информацию смотрите в разделе “Выбор источника данных”.

Если нужно создать новый источник данных или изменить существующий, нажмите кнопку **Запустить редактор источников данных...**.

Выбор источника данных

В таблице Источники данных выводится список доступных источников данных. Выберите источник, который вы хотите использовать, и нажмите кнопку **OK**.

Нажмите кнопку **Показать владельца**, чтобы вывести владельца источника данных.

Опция **Фильтровать по** позволяет вам фильтровать список источников данных для **Ключевого слова**, которое проверяет критерии фильтра по имени источника данных, его описанию и **Владельцу**. В качестве критерия фильтра можно ввести комбинацию символов - строковых, численных или символов подстановки. Строка поиска зависит от регистра. Нажмите кнопку **Обновить**, чтобы изменить таблицу Источники данных.

_ Подчеркивание можно использовать для представления любого единичного символа в строке поиска.

% Знак процента можно использовать для представления последовательности из любого числа символов в строке поиска, в том числе и для отсутствия символов.

Исправление регистрационных данных

Если ваши регистрационные данные для доступа к Analytic Server отличаются от регистрационных данных для SPSS Modeler Server, при запуске потока для Analytic Server, надо вводить регистрационные данные Analytic Server. Если вам неизвестны регистрационные данные, обратитесь к администратору сервера.

Поддерживаемые узлы

Многие узлы SPSS Modeler поддерживаются для вызова в HDFS, но вызов некоторых узлов может чем-то отличаться, а некоторые узлы в настоящее время не поддерживаются. В этой теме подробно описан текущий уровень поддержки.

Общие свойства

- Некоторые символы, обычно допустимые в заключенном в кавычки имени поля Modeler, в Analytic Server будут неприемлемы.
- Чтобы поток Modeler был запущен в Analytic Server, он должен начинаться с одного или нескольких узлов источника Analytic Server и заканчиваться на одном узле моделирования или узле экспорта Analytic Server.
- Рекомендуется задать хранение непрерывных значений назначения не как целых, а как действительных чисел. Модели скоринга всегда записывают действительные значения в выходные файлы данных для непрерывных значений назначения, тогда как выходная модель данных использует для оценок хранение назначения. Поэтому, если у непрерывного назначения будет целочисленный тип хранения, возникнет несогласованность между записанными значениями и моделью данных для оценок, которое приведет к ошибкам при попытке прочитать оцененные данные.

Источник

- Поток, который начинается не с узла источника Analytic Server, будет выполняться локально.

Операции записи

Поддерживаются все операции записей, кроме потоковых временных рядов и узлов пространственно-временных диапазонов. Далее следуют дополнительные замечания о функциональных возможностях поддерживаемых узлов.

Выбрать

- Поддерживается один и тот же набор функций, поддерживаемых узлом извлечения.
- Если используется узел выбора с опцией отбрасывания, поля со значением null будут отброшены из набора результатов. Например, если критерий - отбрасывать строки, в которых OCCUPATION = "Retired", все строки с OCCUPATION = "Retired" и OCCUPATION = null будут отброшены. Необходимо изменить критерии отбора, чтобы добавить условие "not(field = undef)". Например, измените критерий отбора на следующий: ((OCCUPATION = "Retired) and not(OCCUPATION = undef)). При этом в наборе результатов останутся строки, для которых значение в поле OCCUPATION - null.

Выборка

- Выборка на уровне блоков не поддерживается.
- Методы сложной выборки не поддерживаются.

Агрегировать

- Смежные ключи не поддерживаются. Если вы повторно используете существующий поток, который сконфигурирован для сортировки данных, и затем используете этот параметр в узле Агрегировать, измените поток, удалив узел сортировки.
- Порядковые статистики (медиана, 1-й квартиль, 3-й квартиль) вычисляются приблизительно и поддерживаются на вкладке Оптимизация.

Сортировка

- Вкладка Оптимизация не поддерживается.

В распределенной среде существует ограниченное число операций, сохраняющих порядок записей, устанавливаемый узлом сортировки.

- Сортировка, после которой следует узел экспорта, генерирует источник отсортированных данных.
- Сортировка, после которой следует узел выборки, с выборкой записей **Первые** возвращает первые *N* записей.

В общем случае узел сортировки следует размещать как можно ближе к операциям, которым требуются отсортированные записи.

Слияние

- Слияние по порядку не поддерживается.
- Вкладка Оптимизация не поддерживается.
- Analytic Server не выполняет операцию объединения по ключам пустых строк; то есть если один из ключей, по которым выполняется слияние, содержит пустые строки, все записи, содержащие пустую строку, будут отброшены из полученного слиянием вывода.
- Операции слияния относительно медленны. При наличии доступного пространства в HDFS может получиться гораздо быстрее слить источники данных за один раз и использовать полученный слиянием источник в следующих потоках, чем выполнять слияние источников в каждом потоке.

R-преобразование

Синтаксис R в узле должен состоять из операций с записями по одной записи за раз.

Операции с полями

Поддерживаются все операции с полями, за исключением узлов Транспонирование, Временные интервалы и Хронология. Далее следуют дополнительные замечания о функциональных возможностях поддерживаемых узлов.

Автоматическая подготовка данных

- Обучение узла не поддерживается. Применение преобразований на узле автоматической подготовки данных к новым данным не поддерживается.

Тип

- Столбец проверки не поддерживается.
- Вкладка Формат не поддерживается.

Произвести от

- Поддерживаются все функции Произвести от (за исключением функций последовательности).
- Поля разбиения нельзя получить в том же потоке, где они используются для разбиений; необходимо создать два потока: один, получающий поле разбиения, и другой, использующий это поле для разбиений.
- Поле флага нельзя использовать само по себе для сравнения, то есть конструкция вида if (flagField) then ... endif приведет к ошибке; рекомендуемый прием - использовать конструкцию if (flagField=trueValue) then ... endif
- При использовании оператора ** рекомендуется задавать показатель степени в виде действительного числа (например, x**2,0 вместо x**2) для сопоставления результатов в Modeler.

Заполнитель

- Поддерживается один и тот же набор функций, поддерживаемых узлом извлечения.

Категоризация

Следующие функциональные возможности не поддерживаются.

- Оптимальная категоризация
- Ранги
- Квантили -> Разделение на квантили: Сумма значений
- Квантили -> Совпадающие наблюдения: Сохранять в текущем и Назначать произвольно
- Квантили -> Пользовательское N: Значения свыше 100 и любое значение N, где 100% N не равно нулю.

RFM-анализ (недавность, частота, деньги)

- Опция Сохранять в текущем для обработки совпадающих наблюдений не поддерживается. RFM-оценки недавности, частоты и денег не всегда будут совпадать с вычисляемыми Modeler по одним и тем же данным. Диапазоны оценок будут одинаковы, но число назначений оценок (количество интервалов) может отличаться на единицу.

Диаграммы

Поддерживаются все узлы диаграмм.

Моделирование

Поддерживаются следующие узлы моделирования: TCM, Дерево-AS, Дерево C&R, Quest, CHAID, Линейный, Линейный-AS, Нейросеть, GLE, LSVM, Двухшаговый-AS, Случайные деревья, STP и Правила связывания. Далее следует дополнительные замечания о функциональных возможностях этих узлов.

Линейное

При построении моделей на больших данных, скорее всего, вам захочется или изменить цель на цель Очень большие наборы данных, или задать расщепления.

- Непрерывное обучение существующих моделей PSM не поддерживается.
- Цель Построение стандартных моделей рекомендуется, только если поля расщепления определены так, что число записей в каждом расщеплении не слишком велико; здесь определение "слишком велико" зависит от мощности отдельных узлов в используемом

кластере Hadoop. И наоборот, нужно также соблюдать осторожность, чтобы расщепления не были определены настолько точно, чтобы существовало слишком мало записей для построения модели.

- Цель Бустинг не поддерживается.
- Цель Бэггинг не поддерживается.
- При наличии небольшого количества записей цель Очень большие наборы данных не рекомендуется; зачастую она либо не построит модель, либо построит ухудшенную модель.
- Автоматическая подготовка данных не поддерживается. Это может привести к ошибкам при попытке построения модели для данных с множеством пропущенных значений; как правило, их можно можно импутировать в составе автоматической подготовки данных. Как обходной прием, можно использовать модель дерева или нейросеть с дополнительным параметром для импутации выбранных пропущенных значений.
- Статистика точности для моделей расщепления не вычисляется.

Нейросеть

При построении моделей на больших данных, скорее всего, вам захочется или изменить цель на цель Очень большие наборы данных, или задать расщепления.

- Непрерывное обучение существующих стандартных моделей и моделей PSM не поддерживается.
- Цель Построение стандартных моделей рекомендуется, только если поля расщепления определены так, что число записей в каждом расщеплении не слишком велико; здесь определение "слишком велико" зависит от мощности отдельных узлов в используемом кластере Hadoop. И наоборот, нужно также соблюдать осторожность, чтобы расщепления не были определены настолько точно, чтобы существовало слишком мало записей для построения модели.
- Цель Бустинг не поддерживается.
- Цель Бэггинг не поддерживается.
- При наличии небольшого количества записей цель Очень большие наборы данных не рекомендуется; зачастую она либо не построит модель, либо построит ухудшенную модель.
- Если в данных много пропущенных значений, ипмпутируйте их при помощи дополнительного параметра.
- Статистика точности для моделей расщепления не вычисляется.

Дерево C&R, CHAID и Quest

При построении моделей на больших данных, скорее всего, вам захочется или изменить цель на цель Очень большие наборы данных, или задать расщепления.

- Непрерывное обучение существующих моделей PSM не поддерживается.
- Цель Построение стандартных моделей рекомендуется, только если поля расщепления определены так, что число записей в каждом расщеплении не слишком велико; здесь определение "слишком велико" зависит от мощности отдельных узлов в используемом кластере Hadoop. И наоборот, нужно также соблюдать осторожность, чтобы расщепления не были определены настолько точно, чтобы существовало слишком мало записей для построения модели.
- Цель Бустинг не поддерживается.
- Цель Бэггинг не поддерживается.
- При наличии небольшого количества записей цель Очень большие наборы данных не рекомендуется; зачастую она либо не построит модель, либо построит ухудшенную модель.
- Интерактивные сеансы не поддерживаются.
- Статистика точности для моделей расщепления не вычисляется.

- Когда есть поля расщепления, локально построенные в Modeler, модели деревьев немного отличаются от моделей деревьев, построенных Analytic Server, и поэтому приводят к отличающимся оценкам. В обоих случаях используются допустимые алгоритмы, но алгоритмы Analytic Server новее. С учетом того факта, что у алгоритмов деревьев обычно возникает много эвристических правил, различие между двумя компонентами вполне допустимо.

Скоринг модели

Все модели, поддерживаемые для моделирования, поддерживаются также для скоринга. Кроме этого, для скоринга поддерживаются локально встроенные слепки моделей для следующих узлов: C&RT, Quest, CHAID, Линейный и Нейросеть (Независимо от типа модели - стандартная, инкапсулированная с бустингом или для очень больших наборов данных), Регрессия, C5.0, Логистический, Genlin, GLMM, Кокса, SVM, Сети Байеса, Двухэтапный, KNN, Модели списка решений, Дискриминантный, Самообучающийся, выявление аномалий, Априорные значения, Carma, K-средние, Коонена, R и Текстовый анализ.

- Ни простые, ни скорректированные склонности оцениваться не будут. Один и тот же эффект можно получить обходным приемом, вычислив простую склонность вручную при помощи узла извлечения и введя следующее выражение: if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif
- Analytic Server при скоринге модели не проверяет, все ли используемые в модели поля представлены в наборе данных, чтобы убедиться в соблюдении этого условия перед запуском Analytic Server.

R Синтаксис R в слепке должен состоять из операций с записями по одной записи за раз.

Вывод Узлы матрицы, анализа, аудита данных, преобразования, статистики, средних и таблиц не поддерживаются. Далее следуют дополнительные замечания о функциональных возможностях поддерживаемых узлов.

Аудит данных

На узле Аудит данных не может быть установлен режим для непрерывных полей.

Средние

На узле Средние нельзя создать среднеквадратичную ошибку или 95%-ный доверительный интервал.

Таблица

Узел Таблица поддерживается записью временного источника данных Analytic Server, содержащего результаты операций более высокого уровня. Затем узел Таблица просматривает содержание страниц этого источника данных.

Экспортировать

Поток может начинаться с узла источника Analytic Server и заканчиваться узлом экспорта, иным чем узел экспорта Analytic Server, но данные будут перемещены из HDFS на SPSS Modeler Server и окончательно в положение экспорта.

Узел источника базы данных

Узел источника базы данных можно использовать для импорта данных из ряда других пакетов при помощи ODBC (Open Database Connectivity), включая Microsoft SQL Server, DB2, Oracle и другие.

Для чтения или записи данных из базы данных пользователь должен установить источник данных ODBC, настроить соответствующую базу данных и установить разрешения на чтение и запись. IBM SPSS Data Access Pack содержит набор драйверов ODBC, который может использоваться для этой цели; эти драйверы доступны на сайте скачивания. Для получения дополнительных сведений о создании и настройке разрешений для источников данных ODBC обратитесь к своему администратору базы данных.

Поддерживаемые драйверы ODBC

Наиболее свежие сведения о базах данных и драйверах ODBC, работоспособность которых проверена и поддерживается при использовании с IBM SPSS Modeler 18, находятся в матрице совместимости программного обеспечения на веб-сайте технической поддержки (http://www.ibm.com/support).

Установка драйверов на компьютеры

Примечание: Драйверы ODBC должны быть установлены и настроены на каждом из компьютеров, обрабатывающих данные.

- Если программа IBM SPSS Modeler используется в локальном (отдельном) режиме, драйверы должны быть установлены на локальный компьютер.
- Если программы IBM SPSS Modeler работают в распределенном режиме совместно с IBM SPSS Modeler Server, драйверы ODBC должны быть установлены на компьютере с IBM SPSS Modeler Server. Для IBM SPSS Modeler Server в системах UNIX смотрите также "Конфигурирование драйверов ODBC в системах UNIX" далее в этом разделе.
- При необходимости доступа к одним и тем же источникам данных как с IBM SPSS Modeler, так и с IBM SPSS Modeler Server, драйверы ODBC должны быть установлены на обоих компьютерах.
- При использовании IBM SPSS Modeler совместно со службами терминалов драйверы ODBC должны быть установлены на сервере служб терминалов, на котором установлена программа IBM SPSS Modeler.

Обращение к данным из базы данных

Чтобы обратиться к данным в базе данных, выполните следующие действия.

- Установите драйвер ODBC и сконфигурируйте источник данных для базы данных, которую вы хотите использовать.
- В диалоговом окне узла База данных соединитесь с базой данных, применив режим Таблица или режим Запрос SQL.
- Выберите таблицу из базы данных.
- При помощи вкладок в диалоговом окне узла База данных можно изменить типы использования и отфильтровать поля данных.

Дополнительные подробности о предшествующих шагов представлены в соответствующих темах документации.

Примечание: Если хранимые процедуры базы данных (Stored Procedure, SP) вызываются из SPSS Modeler, может появиться одно возвращенное выходное поле с именем `RowsAffected` вместо ожидаемого вывода хранимой процедуры. Это происходит, когда ODBC не возвращает достаточной информации для возможности определить выходную модель данных SP. У SPSS Modeler есть только ограниченная поддержка хранимых процедур, возвращающих выходные данные, и рекомендуется вместо использования SP извлекать из хранимой процедуры SELECT и использовать одно из следующих действий.

- Создать представление на основе оператора SELECT и выбрать это представление в узле источника базы данных
- Непосредственно использовать SELECT в узле источника базы данных.

Задание опций узла базы данных

Опции на вкладке Данные диалогового окна источника базы данных позволяют получить доступ к базе данных и прочитать данные из выбранной таблицы.

Режим. Выберите **Таблица**, чтобы соединиться с таблицей при помощи элементов управления этого диалогового окна.

Выберите **Запрос SQL**, чтобы запросить выбранную ниже базу данных при помощи SQL. Дополнительную информацию смотрите в разделе “Запросы к базе данных” на стр. 25.

Источник данных. И для моды Таблица, и для моды Запрос SQL можно ввести имя в поле Источник данных или выбрать **Добавить новое соединение с базой данных** в выпадающем списке.

Для соединения с базой данных и выбора таблицы при помощи этого диалогового окна используются следующие опции:

Имя таблицы. Если вам известно имя таблицы, к которой вы хотели бы обратиться, ведите его в поле Имя таблицы. В противном случае нажмите кнопку **Выбрать**, чтобы открыть диалоговое окно, предоставляющее список доступных таблиц.

Заключать в кавычки имена таблиц и столбцов. Укажите, хотите ли вы заключать имена таблиц и столбцов в кавычки при отправке запросов к базе данных (если, например, они будут содержать знаки пробелов или пунктуации).

- Опция **Когда требуется** будет заключать в кавычки имена таблиц и столбцов, *только* если они будут содержать нестандартные символы. К нестандартным символам относятся символы не из кодового набора ASCII, символы пробелов и все не алфавитно-цифровые символы, кроме точки (.).
- Выберите **Всегда**, если вы хотите заключать в кавычки *все* имена таблиц и столбцов.
- Выберите **Никогда**, если вы хотите *никогда* не заключать имена таблиц и столбцов в кавычки.

Отсекать ведущие и конечные пробелы. Выберите опции для отбрасывания начальных и конечных пробельных символов в строках.

Примечание. Сравнения между строками, использующими и не использующими SQL pushback, могут генерировать различные результаты, если существуют заключительные пробелы.

Чтение пустых строк из Oracle. При чтении из базы данных Oracle или записи в нее данных следует знать, что в отличие от IBM SPSS Modeler и от большинства других баз данных Oracle обрабатывает и хранит значения пустых строк, приравнивая их к пустым значениям (NULL). Это означает, что данные, извлеченные из базы данных Oracle, могут вести себя не так, как данные, извлеченные из файла или другой базы данных, и могут возвратить другие результаты.

Добавление соединения с базой данных

Чтобы открыть базу данных, сначала выберите источник данных, с которым вы хотите соединиться. На вкладке Данные в выпадающем списке Источник данных выберите **Добавить новое соединение с базой данных**.

Откроется диалоговое окно Соединения с базами данных.

Примечание: Другой способ открыть это диалоговое окне - воспользоваться пунктом главного меню: **Инструменты > Базы данных...**

Источники данных. Список доступных источников данных. Прокрутите его, если не видите нужной базы данных. После выбора источника данных и ввода нужных паролей нажмите кнопку **Соединить**. Нажмите кнопку **Обновить**, чтобы обновить список.

Имя пользователя и пароль Если источник данных защищен паролем, введите ваше имя пользователя.

Регистрационные данные Если регистрационные данные сконфигурированы в IBM SPSS Collaboration and Deployment Services, можно выбрать эту опцию, чтобы просмотреть их в репозитории. Имя пользователя и пароль из регистрационных данных должны соответствовать имени пользователя и паролю для доступа к базе данных.

Соединения Показывает подключенные в текущий момент базы данных.

- **По умолчанию** Одно соединение можно выбрать как соединение по умолчанию, но это необязательно. Это приведет к тому, что для узла источника данных или узла экспорта базы данных это соединение будет предопределено как их источник данных, хотя при желании его можно будет отредактировать.
- **Сохранить** Необязательно: выберите одно или несколько соединений, которые вы хотите выводить повторно в последующих сеансах.
- **Источник данных** Строки соединений для баз данных, подключенных в текущий момент.
- **Предустановка** Указывает (при помощи символа *), заданы ли для соединения с базой данных значения предустановок. Чтобы задать значения предустановок, щелкните в этом столбце в строке, соответствующей нужному соединению с базой данных, и выберите в списке опцию Задать. Дополнительную информацию смотрите в разделе “Задание значений предустановок для соединения с базой данных” на стр. 21.

Чтобы удалить соединения, выберите одно из них в списке и нажмите кнопку **Удалить**.

Завершив работу с выбранными опциями, нажмите кнопку **OK**.

Для чтения или записи данных из базы данных пользователь должен установить источник данных ODBC, настроить соответствующую базу данных и установить разрешения на чтение и запись. IBM SPSS Data Access Pack содержит набор драйверов ODBC, который может использоваться для этой цели; эти драйверы доступны на сайте скачивания. Для получения дополнительных сведений о создании и настройке разрешений для источников данных ODBC обратитесь к своему администратору базы данных.

Поддерживаемые драйверы ODBC

Наиболее свежие сведения о базах данных и драйверах ODBC, работоспособность которых проверена и поддерживается при использовании с IBM SPSS Modeler 18, находятся в матрице совместимости программного обеспечения на веб-сайте технической поддержки (http://www.ibm.com/support).

Установка драйверов на компьютеры

Примечание: Драйверы ODBC должны быть установлены и настроены на каждом из компьютеров, обрабатывающих данные.

- Если программа IBM SPSS Modeler используется в локальном (отдельном) режиме, драйверы должны быть установлены на локальный компьютер.
- Если программы IBM SPSS Modeler работают в распределенном режиме совместно с IBM SPSS Modeler Server, драйверы ODBC должны быть установлены на компьютере с IBM SPSS Modeler Server. Для IBM SPSS Modeler Server в системах UNIX смотрите также "Конфигурирование драйверов ODBC в системах UNIX" далее в этом разделе.
- При необходимости доступа к одним и тем же источникам данных как с IBM SPSS Modeler, так и с IBM SPSS Modeler Server, драйверы ODBC должны быть установлены на обоих компьютерах.
- При использовании IBM SPSS Modeler совместно со службами терминалов драйверы ODBC должны быть установлены на сервере служб терминалов, на котором установлена программа IBM SPSS Modeler.

Настройка драйверов ODBC в системах UNIX

По умолчанию менеджер драйверов DataDirect не настроен для систем IBM SPSS Modeler Server UNIX. Чтобы настроить UNIX для загрузки менеджера драйверов DataDirect, введите следующие команды:

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

В результате будет удалена ссылка по умолчанию и создана ссылка на диспетчер устройств DataDirect.

Примечание: Для некоторых баз данных использование драйверов SAP HANA или IBM DB2 CLI требует оболочки драйвера UTF16. Для DashDB требуется драйвер IBM DB2 CLI. Чтобы создать ссылку на оболочку драйвера UTF16, введите следующие команды:

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

Чтобы сконфигурировать SPSS Modeler Server:

1. Сконфигурируйте SPSS Modeler Server запускать сценарий `modelersrv.sh` для использования файла среды IBM SPSS Data Access Pack `odbc.sh`, добавив следующую строчку к файлу `modelersrv.sh`:

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

   Здесь `<путь_к_установке_SDAP>` - полный путь к вашей установке IBM SPSS Data Access Pack.
2. Перезапустите SPSS Modeler Server.

Кроме того, только для SAP HANA и IBM DB2, добавьте следующее определение параметра к DSN в вашем файле `odbc.ini`, чтобы избежать переполнения буфера во время соединения:

```
DriverUnicodeType=1
```

Примечание: Оболочка `libspssodbc_datadirect_utf16.so` совместима также с другими поддерживаемыми SPSS Modeler Server драйверами ODBC.

Задание значений предустановок для соединения с базой данных

Для некоторых баз данных можно задать ряд значений параметров по умолчанию для соединения с базой данных. Все эти параметры применяются к экспорту базы данных.

Эта возможность поддерживается базами данных следующих типов.

- IBM InfoSphere Warehouse. Дополнительную информацию смотрите в разделе “Параметры для IBM DB2 InfoSphere Warehouse”.
- SQL Server Enterprise edition и Developer edition. Дополнительную информацию смотрите в разделе “Параметры для SQL Server” на стр. 22.
- Oracle Enterprise edition или Personal edition. Дополнительную информацию смотрите в разделе “Параметры для Oracle” на стр. 22.
- IBM Netezza, IBM DB2 for z/OS и Teradata - все они соединяются с базой данных или схемой схожим образом. Дополнительную информацию смотрите в разделе “Параметры для IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW и Teradata” на стр. 22.

В случае соединения с базой данных или схемой, не поддерживающей эту возможность, появляется сообщение **Никакие предустановки для этого соединения с базой данных сконфигурировать невозможно**.

Параметры для IBM DB2 InfoSphere Warehouse

Эти параметры выводятся для IBM InfoSphere Warehouse.

Табличное пространство Табличное пространство, которое должно использоваться для экспорта. Администраторы баз данных могут создать и сконфигурировать табличные пространства как многораздельные. Мы рекомендуем, выбрав одно или несколько табличных пространств (вместо табличного пространства по умолчанию), использовать их для экспорта базы данных.

Использовать сжатие. Эта опция, если она выбрана, создает таблицы для экспорта со сжатием (например, эквивалентного `CREATE TABLE MYTABLE(...) COMPRESS YES;` в SQL).

Не записывать в журнал обновления Эта опция, если она выбрана, предотвращает запись в журнал при создании таблиц и вставке данных (эквивалент `CREATE TABLE MYTABLE(...) NOT LOGGED INITIALLY;` в SQL).

Параметры для SQL Server

Эти параметры выводятся для SQL Server редакций Enterprise и Developer.

Использовать сжатие. Эта опция, если она выбрана, создает таблицы для экспорта со сжатием.

Сжатие для. Выберите уровень сжатия.

- **Строка.** Эта опция включает поддержку сжатия на уровне строк (например, эквивалентного `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW);` в SQL).
- **Страница.** Эта опция включает поддержку сжатия на уровне страниц (например, `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);` в SQL).

Параметры для Oracle

Параметры Oracle - Базовая опция

Эти параметры выводятся для Oracle редакций Enterprise или Personal, если используется базовая опция.

Использовать сжатие. Эта опция, если она выбрана, создает таблицы для экспорта со сжатием.

Сжатие для. Выберите уровень сжатия.

- **По умолчанию.** Эта опция включает поддержку сжатия по умолчанию (например, `CREATE TABLE MYTABLE(...) COMPRESS;` в SQL). В этом она действует также, как и опция **Базовый**.
- **Тип.** Эта опция включает поддержку базового метода сжатия (например, `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` в SQL).

Параметры Oracle - Дополнительная опция

Эти параметры выводятся для Oracle редакций Enterprise или Personal, если используется дополнительная опция.

Использовать сжатие. Эта опция, если она выбрана, создает таблицы для экспорта со сжатием.

Сжатие для. Выберите уровень сжатия.

- **По умолчанию.** Эта опция включает поддержку сжатия по умолчанию (например, `CREATE TABLE MYTABLE(...) COMPRESS;` в SQL). В этом она действует также, как и опция **Базовый**.
- **Тип.** Эта опция включает поддержку базового метода сжатия (например, `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` в SQL).
- **OLTP.** Эта опция включает сжатие OLTP (например, `CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP;` в SQL).
- **Низкая/высокая для запросов.** (Только для серверов Exadata) Эта опция включает поддержку сжатия по столбцам для запросов, (например, `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW;` или `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH;` в SQL). Сжатие для столбцов полезно в средах хранилищ данных; опция `HIGH` обеспечивает более высокую степень сжатия, чем `LOW`.
- **Низкая/высокая для архивов.** (Только для серверов Exadata) Эта опция включает поддержку сжатия по столбцам для архива, (например, `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW;` или `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH;` в SQL). Сжатие для архивов полезно для сжатия данных, которые будут храниться длительные периоды времени; опция `HIGH` обеспечивает более высокую степень сжатия, чем `LOW`.

Параметры для IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW и Teradata

При задании предустановок для IBM Netezza, IBM DB2 for z/OS, IBM DB2 LUW или Teradata вам будет предложено выбрать следующие параметры:

Использовать базу данных адаптера скоринга сервера или **Использовать схему адаптера скоринга сервера**. Этот параметр, если он выбран, включает опцию **Использовать базу данных адаптера скоринга сервера** или **Использовать схему адаптера скоринга сервера**.

Использовать базу данных адаптера скоринга сервера или **Использовать схему адаптера скоринга сервера** Выберите нужное вам соединение из выпадающего списка.

Кроме этого, в Teradata можно задать подробности о мечении запросов, чтобы предоставить дополнительные метаданные, помогающие в таких вопросах, как управление рабочими нагрузками; упорядочение, идентификация и разрешение запросов; отслеживание использования базы данных.

Формулировка мечения запросов. Выберите, проводится ли определение мечения запросов на все время вашей работы с соединением базы данных Teradata (**На сеанс**), или будет повторяться каждый раз при запуске потока (**На транзакцию**).

Примечание: Если задать мечение запроса для потока, оно теряется при копировании этого потока на другой компьютер. Чтобы избежать этого, можно использовать для запуска потоков сценарии с ключевым словом *querybanding* для применения нужных параметров.

Требуемые разрешения базы данных

Чтобы правильно работали возможности баз данных SPSS Modeler, предоставьте всем используемым ID пользователей доступ к следующим элементам:

DB2 LUW

SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABCONSTPKC

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSCAT.TABLESPACES

SYSCAT.SCHEMATA

DB2/z SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABCONSTPKC

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSIBM.SYSDUMMYU

SYSIBM.SYSPACKSTMT

Netezza

_V_FUNCTION

_V_DATABASE

Teradata

DBC.Functions

DBC.USERS

Выбор таблицы базы данных

После подключения к источнику данных можно выбрать импорт полей из конкретной таблицы или представления. На вкладке Данные диалогового окна База данных можно либо ввести имя таблицы в поле Имя таблицы, либо нажать кнопку **Выбрать**, чтобы открыть диалоговое окно Выбрать таблицу/представление, содержащее список доступных таблиц и представлений.

Показать владельца таблицы . Выберите эту опцию, если источник данных требует, чтобы владелец таблицы был задан до того, как можно будет обратиться к этой таблице. Отключите эту опцию для источников данных, где указанное требование отсутствует.

Примечание: Базы данных SAS и Oracle обычно требуют, чтобы был представлен владелец таблицы.

Таблицы/представления. Выберите таблицу или представление для импорта.

Показать. Возвращает список столбцов в источнике данных, с которым вы в текущий момент соединены. Щелкните по одной из следующих опций, чтобы настроить представление доступных таблиц:

- Выберите опцию **Таблицы пользователей**, чтобы просмотреть обычные таблицы базы данных, созданные пользователями базы данных.
- Выберите **Системные таблицы**, чтобы просмотреть таблицы базы данных, которыми владеет система (например, таблицы, предоставляющие информацию о базе данных, такую как подробности индексов). При помощи этой опции можно просмотреть вкладки, используемых базами данных Excel. (Имейте в виду, что доступен также отдельный узел источника Excel. Дополнительную информацию смотрите в разделе “Узел источника Excel” на стр. 42.)

- Выберите **Представления**, чтобы просмотреть виртуальные таблицы на основе запросов, в которых используется одна или несколько обычных таблиц.
- Выберите **Синонимы**, чтобы просмотреть синонимы, созданные в базе данных для каких-либо существующих таблиц.

Фильтры имя/владелец. Эти поля позволяют отфильтровать список выводимых таблиц по имени или владельцу. Например, введите SYS, чтобы в списке остались только таблицы с этим владельцем. Для поисков с подстановкой один символ можно представлять знаком подчеркивания (_), а последовательность, состоящую из любого числа символов (включая 0) - знаком процента (%).

Задать как значения по умолчанию. Сохраняет текущие значения параметров в качестве значений по умолчанию для текущего пользователя. Эти значения параметров будут восстановлены в будущем при открытии пользователем нового диалогового окна выбора таблиц *только для этого же имени источника данных и входа пользователя в систему*.

Запросы к базе данных

После подключения к источнику данных можно выбрать импорт полей при помощи запросов SQL. В главном диалоговом окне выберите в качестве режима соединения **Запрос SQL**. Эта опция добавляет в это диалоговое окно окно редактора запросов. При помощи редактора запросов можно создать или загрузить один или насколько запросов SQL, набор результатов которых будет считан в поток данных.

В случае задания нескольких запросов SQL разделяйте их точкой с запятой (;) и убедитесь, что не используется оператор SELECT множественного выбора.

Чтобы отменить и закрыть окно редактора запросов, выберите в качестве режима соединения опцию **Таблица**.

В запрос SQL можно включить параметры потока SPSS Modeler (тип пользовательской переменной). Дополнительную информацию смотрите в разделе “Использование параметров потока в запросе SQL”.

Загрузить запрос. Нажмите эту кнопку, чтобы открыть браузер, при помощи которого можно загрузить сохраненный ранее запрос.

Сохранить запрос. Нажмите эту кнопку, чтобы открыть диалоговое окно Сохранить запрос, при помощи которого можно сохранить текущий запрос.

Импортировать по умолчанию. Нажмите эту кнопку, чтобы импортировать оператор SQL SELECT примера, построенный автоматически при помощи таблицы и столбцов, выбранных в этом диалоговом окне.

Очистить. Очистка содержания рабочей области. Используйте эту опцию, если хотите начать сначала.

Разделять текст. Опция по умолчанию **Никогда** означает, что запрос будет отправлен на базу данных как единое целое. Другой вариант - выбрать **Как требуется**, что означает, что SPSS Modeler пытается проанализировать запрос и идентифицировать, есть ли операторы SQL, которые можно было бы послать на базу данных один за другим.

Использование параметров потока в запросе SQL

При написании запроса SQL для импорта полей в него можно включить параметры потока SPSS Modeler, которые были определены заранее. Поддерживаются параметры потока всех типов.

В следующей таблице показано, как некоторые примеры параметров потока будут интерпретироваться в запросе SQL.

Таблица 3. Примеры параметров потока.

Имя параметра потока (пример)	Хранение	Значение параметра потока	Интерпретируется как
PString	Текстовое	ss	'ss'
PInt	Целое число	5	5
PReal	Действительное число	5.5	5.5
PTime	Время	23:05:01	t{'23:05:01'}
PDate	Дата	2011-03-02	d{'2011-03-02'}
PTimeStamp	TimeStamp	2011-03-02 23:05:01	ts{'2011-03-02 23:05:01'}
PColumn	Нет данных	IntValue	IntValue

В запросе SQL параметр потока задается тем же способом, что и в выражении CLEM, а именно, посредством `'$P-<имя_параметра>'`, где `<имя_параметра>` - имя, которое было определено для этого параметра потока.

При ссылке на поле тип хранения должен быть определен как Нет данных, а значение параметра при необходимости должно быть заключено в кавычки. Таким образом, использование показанных в таблице примеров, если вы ввели запрос SQL:

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

будет оценено как:

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

Если вы должны были сослаться на поле `IntValue` при помощи параметра `PColumn`, для получения того же результата нужно будет задать следующий запрос:

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

Узел файла переменных

Вы можете использовать узлы Файл переменных для чтения данных из текстовых файлов в формате со свободными полями, то есть, файлов, записи которых содержат постоянное число полей, но изменяющееся число символов. Этот узел полезен также для файлов с текстом заголовков фиксированной длины и содержащих определенные типы аннотаций. Записи читаются по одной и передаются через поток, пока не будет прочитан весь файл.

Замечания о чтении геопространственных данных

Если содержащий геопространственные данные узел был создан при экспорте из плоского файла, нужно выполнить следующие дополнительные действия для конфигурирования геопространственных метаданных. Дополнительную информацию смотрите в разделе “Импорт геопространственных данных на узел файла переменных” на стр. 29.

Замечания о чтении данных текста с разделителями

- Записи должны быть разделены символом новой строки в конце каждой строки. Символ новой строки не должен использоваться для каких-то других целей (например, в любом имени или значении поля). В идеальном случае все начальные и заключительные пробелы нужно удалять для экономии памяти, хотя это не критично. Если требуется, эти пробелы могут быть удалены узлом.
- Поля должны разделяться запятой или другим символом, который в идеале должен использоваться только как разделитель, то есть не появляется в именах или значениях полей. Если это невозможно, все текстовые поля можно заключить в двойные кавычки при условии, что никакие имена или текстовые значения полей не содержат двойных кавычек. Если имена или значения полей содержат двойные кавычки, есть вариант заключения текстовых полей в одинарные кавычки, снова в предположении, что одинарные

кавычки не используются где-либо еще в значениях. Если нельзя использовать ни одинарные, ни двойные кавычки, текстовые значения нужно исправить для удаления или замены или символа разделения, или одинарных/двойных кавычек.

- Каждая строка, в том числе строка заголовка, должна содержать одинаковое число полей.
- Первая строка должна содержать имена полей. Если это не так, отмените выбор **Читать имена полей из файла**, чтобы дать каждому полю свое имя, такое как Поле1, Поле2 и так далее.
- Вторая строка должна содержать первую запись данных. Здесь не должно быть пустых строк или комментариев.
- Числовые значения не должны содержать разделителя тысяч или знака группировки, например, запятой в числе 3,000.00. Десятичный разделитель (точка в стандартах США или Великобритании) должна использоваться только в подходящих местах.
- Значения даты и времени должны быть в одном из форматов, распознаваемых в диалоговом окне Опции потока, таком как `ДД/ММ/ГГГГ` или `ЧЧ:ММ:СС`. В идеале все поля даты и времени в файле должны следовать одному формату, и любое поле, содержащее дату, должно использовать одинаковый формат для всех значений в этом поле.

Задание опций для узла файла переменных

Опции задаются на вкладке Файл диалогового окна узла Файл переменных.

Файл Задайте имя файла. Можно ввести имя файла или нажать кнопку с многоточием (**...**), чтобы выбрать файл. После выбора файла выводится его путь, и содержимое с ограничителями появляется на панели ниже.

Образец выведенного с вашего источника данных текста можно скопировать и вставить в следующие элементы управления: символы комментариев EOL и заданные пользователем разделители. Для копирования и вставки используйте комбинации клавиш Ctrl-C и Ctrl-V.

Читать имена полей из файла Эта опция, выбираемая по умолчанию, рассматривает первую строку в файле данных как метки для столбца. Если ваша первая строка - это не заголовок, отмените выбор, чтобы автоматически дать каждому полю свое имя, такое как *Поле1, Поле2* по числу полей в наборе данных.

Задайте число полей. Задайте число полей в каждой записи. Число полей можно определить автоматически, если записи завершаются символа новой строки. Можно также задать число вручную.

Пропустить символы заголовка. Укажите, сколько символов вы хотите игнорировать в начале первой записи.

Символы комментариев конца строки. Задайте символы, такие как # или !, чтобы указать аннотации в данных. В любом месте файла, где бы ни встретился этот символ, все символы до следующего символа новой строки, но не включая его, будут игнорированы.

Отсекать ведущие и конечные пробелы. Выберите опции для отбрасывания начальных и конечных пробельных символов в строках при импорте.

Примечание. Сравнения между строками, использующими и не использующими SQL pushback, могут генерировать различные результаты, если существуют заключительные пробелы.

Недопустимые символы. Выберите **Отбросить**, чтобы удалить недопустимые символы из источника данных. Выберите **Заменить на**, чтобы заменить недопустимые символы на заданный специальный символ (всего один символ). Недопустимые символы - это пустые символы и все символы, не существующие в заданном методе кодировки.

Кодировка. Задает используемый метод кодирования текста. Можно выбрать по умолчанию для системы, по умолчанию для потока или UTF-8.

- По умолчанию для системы задается на панели управления Windows или (при работе в распределенном режиме) на компьютере сервера.

- По умолчанию для потока задается в диалоговом окне Свойства потока.

Десятичный разделитель Выберите тип десятичного разделителя, используемого в вашем источнике данных. **Значение по умолчанию потока** - это символ, выбранный на вкладке Опции диалогового окна свойств потока. Иначе выберите либо **Точка (.)**, либо **Запятая (,)**, чтобы все данные в этом диалоговом окне читались при помощи выбранного символа в качестве десятичного разделителя.

Разделитель строк - символ новой строки Выберите эту опцию, чтобы вместо разделителя полей использовать в качестве разделителя строк символ новой строки. Например, это может быть полезно, если в строке есть нечетное количество разделителей, что приводит к переносу строки. Обратите внимание, что выбор этой опции означает невозможность выбора опции **Новая строка** в списке разделителей.

Примечание: Если вы выбрали эту опцию, любые пробельные значения в конце строк данных будут усечены.

Разделители. При помощи переключателей, показанных для этого элемента управления, можно указать, какие символы будут определять границы полей в файле; например, знак запятой (,). Можно также задать несколько разделителей (таких как ", |") для записей, где используется несколько разделителей. Разделитель по умолчанию - запятая.

Примечание: Если запятая определена также и как разделитель десятичной части, значения параметров по умолчанию здесь работать не будут. В случаях, где запятая является и разделителем полей, и разделителем десятичной части, выберите в списке разделителей **Другой**. Затем задайте в поле вводя запятую вручную.

Выберите **Разрешить несколько пробельных разделителей**, чтобы несколько смежных символов пробельных разделителей обрабатывались как один разделитель. Например, если после значения данных следуют четыре пробельных символа, а затем другое значение, эта группа будет обработана как два поля вместо пяти.

Строки просмотра для столбца и типа Укажите, сколько строк и столбцов просматривать для заданных типов данных.

Автоматически распознавать даты и время Чтобы включить для IBM SPSS Modeler автоматическое распознавание записей данных как дат или времени, включите этот переключатель. Например, это подразумевает, что запись 07-11-1965 будет идентифицирована как дата, а запись 02:35:58 - как время; однако неоднозначные записи, такие как 07111965 или 023558, будут выводиться как целочисленные значения из-за отсутствия разделителей между числами.

Примечание: Во избежание возможных проблем с данными при использовании файлов данных из прежних версий IBM SPSS Modeler этот переключатель выключается по умолчанию для информации, сохраненной в версиях до версии 13.

Рассматривать квадратные скобки как списки Если включить этот переключатель, данные между открывающей и закрывающей квадратными скобками будут рассматриваться как одно значение, даже если в содержимое между скобками входят символы разделителей, такие как запятые или двойные кавычки. Например, так могут описываться двумерные или трехмерные геопространственные данные, координаты которых между квадратными скобками могут обрабатываться как один элемент списка. Дополнительную информацию смотрите в разделе “Импорт геопространственных данных на узел файла переменных” на стр. 29

Кавычки. При помощи выпадающих списков можно указать, как обрабатывать одинарные и двойные кавычки при импорте. Можно выбрать опцию **Отбрасывать** все кавычки, **Включать в виде текста** посредством включения их в значение поля, или **Составлять в пары и отбрасывать**, чтобы составлять знаки кавычек в пары и удалять их. В случае непарных знаков кавычек вы получите сообщение об ошибке. И опция **Отбрасывать**, и опция **Составлять в пары и отбрасывать** сохраняют значение поля (без кавычек).

Примечание: Если выбрана опция **Составлять в пары и отбрасывать**, пробелы сохраняются. Если выбрана опция **Отбрасывать**, концевые пробелы внутри и вне кавычек удаляются (например, `' " ab c" , "d ef " , " gh i " '` становится `'ab c, d ef, gh i'`). Если выбрана опция **Включать в виде текста**, кавычки обрабатываются как обычные символы, поэтому начальные и концевые пробелы будут сниматься естественным образом.

При работе в этом диалоговом окне в любой момент можно нажать кнопку **Обновить**, чтобы перезагрузить поля из источника данных. Это полезно при изменении соединений данных с узлом источника или при работе с несколькими вкладками в этом диалоговом окне.

Импорт геопространственных данных на узел файла переменных

Если содержащий геопространственные данные узел был создан при экспорте из плоского файла и используется в том же потоке, в котором он был создан, этот узел сохраняет геопространственные метаданные, и никакие действия по дальнейшему конфигурированию не требуются.

Однако если этот узел экспортирован и используется в другом потоке, геопространственные данные списка автоматически преобразуются в строковый формат; вам нужно выполнить некоторые дополнительные действия для восстановления типа хранения списка и связанных геопространственных метаданных.

Дополнительную информацию о списках смотрите в разделе “Система хранения списков и связанные шкалы измерений” на стр. 11.

Дополнительную информацию о подробностях, которые можно задать как геопространственные метаданные, смотрите в разделе “Подуровни геопространственных измерений” на стр. 131.

Чтобы сконфигурировать геопространственные метаданные, выполните следующие действия.

1. На вкладке Файл узла Файл переменных включите переключатель **Рассматривать квадратные скобки как списки**. Если включить этот переключатель, данные между открывающей и закрывающей квадратными скобками будут рассматриваться как одно значение, даже если в содержимое между скобками входят символы разделителей, такие как запятые или двойные кавычки. Если не включить этот переключатель, ваши данные будут читаться, как для строкового типа хранения, все запятые будут обрабатываться как разделители, и данные будут интерпретироваться неправильно.
2. Если ваши данные содержат одинарные или двойные кавычки, выберите подходящую опцию **Составлять в пары и отбрасывать** в полях **Одинарные кавычки** и **Двойные кавычки**.
3. На вкладке Данные узла Файл переменных включите для полей геопространственных данных переключатель **Перезаписать** и измените тип **хранения** со строки на список.
4. По умолчанию для типа **хранения** списка задано значение *Список действительных чисел*, а для внутреннего типа хранения значений в поле списка задано *Действительное число*. Для изменения внутреннего типа хранения значений или глубины нажмите кнопку **Задать...**, чтобы вывести внутреннее диалоговое окно Система хранения.
5. Во внутреннем диалоговом окне Система хранения можно изменить следующие параметры:
   - **Система хранения** Задайте общий тип хранения поля данных. По умолчанию для типа хранения задано значение Список; однако в выпадающем списке содержатся все остальные типы хранения (Строка, Целое, Действительное, Дата, Время и Отметка времени). Если выбрать тип хранения, отличный от Список, опции **Хранение значений** и **Глубина** будут недоступны.
   - **Система хранения значений** Задайте типы хранения элементов в списке (в отличие от задания типа в целом для поля). При импорте геопространственных полей единственными обоснованными типами хранения могут быть Действительное или Целое; параметр по умолчанию - Действительное.
   - **Глубина** Задайте глубину поля списка. Требуемая глубина зависит от типа геопространственного поля и определяется по следующим критериям:
     – Точка – 0
     – Ломаная – 1
     – Многоугольник – 1

– Несколько точек – 1
– Мультиломаная – 2
– Мультиполигон – 2

Прим.: Вы должны знать тип геопространственного поля, преобразуемого обратно в список, и нужную глубину поля такого сорта. Если эти сведения заданы неправильно, использовать данное поле нельзя.

6. На вкладке Типы узла Файл переменных убедитесь для полей геопространственных данных, что ячейка **Измерение** содержит правильную шкалу измерений. Чтобы изменить шкалу измерений, в ячейке **Измерение** нажмите кнопку **Задать...** для вывода диалогового окна Значения.
7. В диалоговом окне Значения появятся **Измерение**, **Система хранения** и **Глубина** для списка. Выберите опцию **Задать значения и метки** и из выпадающего списка **Тип** выберите правильный тип для **Измерения**. В зависимости от **Типа** у вас могут запросить другие подробности, например, о размерности данных - двумерные они или трехмерные, а также о используемой системе координат.

Узел фиксированных файлов

С помощью узла фиксированных файлов можно импортировать данные из текстовых файлов в формате с полями фиксированной ширины (то есть, с полями без разделителей, но с началом в одной и той же позиции и фиксированной длины). В формате с полями фиксированной ширины хранятся данные, генерируемые компьютером или представленные в устаревшем формате. При помощи вкладки Файл узла фиксированных файлов можно легко задать позицию и длину столбцов в данных.

Задание опций для узла фиксированных файлов

Вкладка Файл узла фиксированных файлов позволяет перенести данные в IBM SPSS Modeler и задать позицию столбцов и длину записей. Находясь на панели предварительного просмотра данных в центральной части диалогового окна, можно щелкнуть кнопкой мыши, чтобы добавить кнопки со стрелками, задающие точки разбиения между полями.

По порядку в файле данных. Задайте имя файла. Можно ввести имя файла или нажать кнопку с многоточием (**...**), чтобы выбрать файл. После выбора файла выводится его путь, и содержимое с ограничителями появляется на панели ниже.

Панель предварительного просмотра позволяет указать позицию и длину столбца. Линейка в верхней части окна предварительного просмотра помогает измерить длину переменных и задать точку разбиения между ними. Щелкнув кнопкой мыши в области на линейке над полями, можно задать строки точек разбиения. Точки разбиения можно перемещать перетаскиванием и отбросить, перетащив их за границы региона предварительного просмотра.

- С каждой строкой точек разбиения в таблицу полей ниже автоматически добавляется новое поле.
- Начальные позиции, указываемые кнопками со стрелками, автоматически добавляются в начальный столбец в таблице ниже.

С ориентацией по строкам. Выберите эту опцию, если хотите пропустить символ новой строки в конце каждой записи.

Пропустить строки заголовка. Укажите, сколько строк вы хотите игнорировать в начале первой записи. Эта опция полезна для игнорирования заголовков столбцов.

Длина записей. Задайте число символов в каждой записи.

Поле. Здесь выводятся все поля, определенные для этого файла данных. Определить поля можно двумя способами:

- Задать поля интерактивно при помощи панели предварительного просмотра данных выше.

- Задать поля вручную, добавив пустые строки полей в таблицу ниже. Чтобы добавить новые поля, нажмите кнопку справа от панели полей. Затем в пустом поле введите имя поля, его начальную позицию и длину. Эти опции автоматически добавляют кнопки со стрелками на панель предварительного просмотра данных, которую можно легко настроить.

Для удаления ранее определенного поля, выберите его в списке и нажмите красную кнопку удаления.

Начать. Задайте позицию первого символа в поле. Например, если второе поле записи начинается с шестнадцатого символа, введите в качестве начальной точки 16.

Длина. Укажите, сколько символов будет в самом длинном значении для каждого поля. Эта опция определяет точку отсечения для следующего поля.

Отсекать ведущие и конечные пробелы. Выберите эту опцию для отбрасывания начальных и конечных пробельных символов в строках при импорте.

Примечание. Сравнения между строками, использующими и не использующими SQL pushback, могут генерировать различные результаты, если существуют заключительные пробелы.

Недопустимые символы. Выберите **Отбросить**, чтобы удалить недопустимые символы из источника ввода данных. Выберите **Заменить на**, чтобы заменить недопустимые символы на заданный специальный символ (всего один символ). Недопустимые символы - это пустые символы (0) и все символы, не существующие в текущей кодировке.

Кодировка. Задает используемый метод кодирования текста. Можно выбрать по умолчанию для системы, по умолчанию для потока или UTF-8.

- По умолчанию для системы задается на панели управления Windows или (при работе в распределенном режиме) на компьютере сервера.
- По умолчанию для потока задается в диалоговом окне Свойства потока.

Десятичный разделитель. Выберите тип десятичного разделителя, используемого в вашем источнике данных. **Для потока по умолчанию** - это символ-разделитель, выбранный с вкладки Опции диалогового окна свойств потока. Иначе выберите либо **Точка (.)**, либо **Запятая (,)**, чтобы все данные в этом диалоговом окне читались при помощи выбранного символа в качестве десятичного разделителя.

Автоматически распознавать даты и время. Чтобы включить для IBM SPSS Modeler автоматическое распознавание записей данных как дат или времени, включите этот переключатель. Например, это подразумевает, что запись 07-11-1965 будет идентифицирована как дата, а запись 02:35:58 - как время; однако неоднозначные записи, такие как 07111965 или 023558, будут выводиться как целочисленные значения из-за отсутствия разделителей между числами.

Примечание: Во избежание возможных проблем с данными при использовании файлов данных из прежних версий IBM SPSS Modeler этот переключатель выключается по умолчанию для информации, сохраненной в версиях до версии 13.

Строки просмотра для типа. Укажите, сколько строк просматривать для заданных типов данных.

При работе в этом диалоговом окне в любой момент можно нажать кнопку **Обновить**, чтобы перезагрузить поля из источника данных. Это полезно при изменении соединений данных с узлом источника или при работе с несколькими вкладками в этом диалоговом окне.

Узел Data Collection

Узлы источников Data Collection импортируют данные опросов на основе Survey Reporter Developer Kit, предоставляемого с продуктом Data Collection. Этот формат позволяет отличить *данные наблюдений* (фактические ответы на вопросы, собираемые во время опроса) от *метаданных*, описывающих способы сбора и организации данных наблюдений. В метаданные включается такая информация, как текст вопросов, имена и описания переменных, определения переменных множественных ответов, варианты перевода текстовых строк и определение структуры данных наблюдений.

Примечание: Для этого узла требуется Survey Reporter Developer Kit, распространяемый с продуктом Data Collection. Помимо установки Developer Kit, никакого дополнительного конфигурирования не требуется.

Комментарии

- Данные опросов читаются из плоского табличного формата VDATA или из источников в иерархическом формате HDATA, если они включают в себя источник метаданных.
- Типы инстанциируются автоматически при помощи информации из метаданных.
- При импорте данных опросов в SPSS Modeler вывод вопросов обрабатывается подобно выводу полей, с ведением записи для каждого респондента.

Опции импорта файлов Data Collection

Вкладка Файл на узле Data Collection позволяет задать опции для метаданных и данных наблюдений, которые вы хотите импортировать.

Параметры метаданных

Примечание: Чтобы увидеть полный список доступных типов файлов провайдеров, нужно установить Survey Reporter Developer Kit, доступный с программным продуктом Data Collection.

Провайдер метаданных. Данные опросов можно импортировать из ряда форматов, поддерживаемых комплектом Data Collection Survey Reporter Developer Kit. В состав доступных провайдеров входят:

- **DataCollectionMDD.** Читает метаданные из файла определений опросного листа (*.mdd*). Это стандартный формат Data Collection Data Model.
- **База данных ADO.** Читает данные наблюдений и метаданные из файла ADO. Задайте имя и положение файла *.adoinfo*, содержащего метаданные. Внутреннее имя этого DSC - *mrADODsc*.
- **База данных In2data.** Читает данные наблюдений и метаданные In2data. Внутреннее имя этого DSC - *mrI2dDsc*.
- **Файл журнала Data Collection.** Читает метаданные из стандартного файла журнала Data Collection. Обычно для файлов журналов используется расширение имени файла *.tmp*. Однако у файлов журналов может быть и другое расширение имени файла. При необходимости этот файл можно переименовать, указав расширение имени файла *.tmp*. Внутреннее имя этого DSC - *mrLogDsc*.
- **Файл определений Quancept.** Преобразует метаданные в сценарий Quancept. Задайте имя файла Quancept *.qdi*. Внутреннее имя этого DSC - *mrQdiDrsDsc*.
- **База данных Quanvert.** Читает данные наблюдений и метаданные Quanvert. Задайте имя и положение файла *.qvinfo* или *.pkd*. Внутреннее имя этого DSC - *mrQvDsc*.
- **База данных-участник Data Collection.** Читает таблицы проекта Таблица выборки и хронологии и создает дифференциальные категориальные переменные, соответствующие столбцам в этих таблицах. Внутреннее имя этого DSC - *mrSampleReportingMDSC*.
- **Файл статистики.** Читает данные наблюдений и метаданные из файла IBM SPSS Statistics *.sav*. Записывает данные наблюдений в файл IBM SPSS Statistics *.sav* для анализа в IBM SPSS Statistics. Записывает метаданные из файла IBM SPSS Statistics *.sav* в файл *.mdd*. Внутреннее имя этого DSC - *mrSavDsc*.
- **Файл Surveycraft.** Читает данные наблюдений и метаданные SurveyCraft. Задайте имя файла SurveyCraft *.vq*. Внутреннее имя этого DSC - *mrSCDsc*.

- **Файл сценариев Data Collection.** Читает метаданные в файле mrScriptMetadata. Обычно для этих файлов используется расширение имени файла *.mdd* или *.dms*. Внутреннее имя этого DSC - *mrScriptMDSC*.
- **Файл XML Triple-S.** Читает метаданные из файла Triple-S в формате XML. Внутреннее имя этого DSC - *mrTripleSDsc*.

Свойства метаданных. Необязательно: выберите **Свойства**, чтобы задать версию опроса для импорта, а также язык, контекст и тип меток для использования. Дополнительную информацию смотрите в разделе “Свойства импорта метаданных Data Collection” на стр. 34.

Параметры данных наблюдений

Примечание: Чтобы увидеть полный список доступных типов файлов провайдеров, нужно установить Survey Reporter Developer Kit, доступный с программным продуктом Data Collection.

Получить параметры данных наблюдений. Только при чтении метаданных из файлов *.mdd*; выберите **Получить параметры данных наблюдений**, чтобы определить, какие источники данных наблюдений связаны с выбранными метаданными, наряду с конкретными значениями параметров, необходимыми для обращения к данному источнику. Эта опция доступна только для файла *.mdd*.

Провайдер файла наблюдений. Поддерживаются следующие типы провайдеров:

- **База данных ADO.** Читает данные при помощи интерфейса Microsoft ADO. Выберите `UDL OLE-DB` для указанного типа данных наблюдений и задайте строку соединения в поле UDL данных наблюдений. Дополнительную информацию смотрите в разделе “Строка соединения с базой данных” на стр. 35. Внутреннее имя этого компонента - *mrADODsc*.
- **Текстовый файл с разделителями (Excel).** Читает данные наблюдений из файла с разделителями-запятыми (.CSV), который может выводиться в Excel. Внутреннее имя - *mrCsvDsc*.
- **Файл данных Data Collection.** Читает данные наблюдений из файла собственного формата данных Data Collection. Внутреннее имя - *mrDataFileDsc*.
- **База данных In2data.** Читает данные наблюдений и метаданные из файла базы данных In2data (*.i2d*). Внутреннее имя - *mrI2dDsc*.
- **Файл журнала Data Collection.** Читает данные наблюдений из стандартного файла журнала Data Collection. Обычно для файлов журналов используется расширение имени файла *.tmp*. Однако у файлов журналов может быть и другое расширение имени файла. При необходимости этот файл можно переименовать, указав расширение имени файла *.tmp*. Внутреннее имя - *mrLogDsc*.
- **Файл данных Quantum.** Читает данные наблюдений из любого файла ASCII формата Quantum (*.dat*). Внутреннее имя - *mrPunchDsc*.
- **Файл данных Quancept.** Читает данные наблюдений из файла Quancept *.drs*, *.drz* или *.dru*. Внутреннее имя - *mrQdiDrsDsc*.
- **База данных Quanvert.** Читает данные наблюдений из файла Quanvert *qvinfo* или *.pkd*. Внутреннее имя - *mrQvDsc*.
- **База данных Data Collection (MS SQL Server).** Считывает данные наблюдений в базу данных Microsoft SQL Server. Дополнительную информацию смотрите в разделе “Строка соединения с базой данных” на стр. 35. Внутреннее имя - *mrRdbDsc2*.
- **Файл статистики.** Читает данные наблюдений из файла IBM SPSS Statistics *.sav*. Внутреннее имя - *mrSavDsc*.
- **Файл Surveycraft.** Читает данные наблюдений из файла SurveyCraft *.qdt*. И файл *.vq*, и файл *.qdt* должны находиться в одном и том же каталоге, с доступом чтения и записи для обоих файлов. Это не соответствует способу их создания по умолчанию при помощи SurveyCraft, поэтому для импорта данных SurveyCraft один из этих файлов необходимо переместить. Внутреннее имя - *mrScDsc*.
- **Файл данных Triple-S.** Читает данные наблюдений из файла данных Triple-S либо в формате фиксированной длины, либо в формате с разделителями-запятыми. Внутреннее имя - *mr TripleDsc*.

- **XML Data Collection.** Читает данные наблюдений из файла данных XML Data Collection. Обычно этот формат можно использовать для передачи данных наблюдений из одного положения в другое. Внутреннее имя - *mrXmlDsc*.

Тип данных наблюдений. Указывает, читаются ли данные наблюдений из файла, папки, UDL OLE-DB или DSN ODBC, и изменяет опции диалогового окна соответственно. Допустимые опции зависят от типа провайдера. Для провайдеров баз данных можно задать опции для соединения OLE-DB или ODBC. Дополнительную информацию смотрите в разделе “Строка соединения с базой данных” на стр. 35.

Проект данных наблюдений. При чтении данных наблюдений из базы данных Data Collection вы можете ввести имя проекта. Для всех остальных типов данных наблюдений значение этого параметра следует оставить пустым.

Импортировать переменные

Импортировать системные переменные. Указывает, будут ли импортироваться системные переменные, включая переменные, указывающие на состояние интервью (продолжается, завершено, дата завершения и так далее). Можно выбрать **Нет**, **Все** или **Обычные**.

Импортировать переменные "кодов". Управляет импортом переменных, которые представляют коды, используемые для кодирования "открытых" вопросов в категориальные переменные.

Импортировать переменные "источника". Управляет импортом переменных, содержащих имена файлов образов сканированных анкет.

Импортировать переменные множественных ответов как. Переменные множественных ответов можно импортировать как несколько полей флагов (набор множественных дихотомий), что является методом, используемым по умолчанию для новых потоков. Потоки, созданные в выпусках IBM SPSS Modeler до выпуска 12.0, импортировали множественные ответы в одно поле, со значениями, разделяемыми запятыми. Этот более старый метод все еще поддерживается, что позволяет обрабатывать существующие потоки прежним способом, но старые потоки рекомендуется обновить, чтобы использовать новый метод. Дополнительную информацию смотрите в разделе “Импорт наборов множественных ответов” на стр. 35.

Свойства импорта метаданных Data Collection

При импорте данных опроса Data Collection в диалоговом окне Свойства метаданных можно задать версию опроса для импорта, а также язык, контекст и тип меток для использования. Имейте в виду, что за один раз можно импортировать только один язык, контекст и тип меток.

Версия. Каждая версия опроса может считаться снимком метаданных, используемых для сбора конкретного набора данных наблюдений. По мере внесения изменений в опросный лист может быть создано несколько его версий. Возможен импорт последней версии, всех версий или конкретной версии.

- **Все версии.** Выберите эту опцию, если вы хотите использовать сочетание (расширенный набор) всех доступных версий. (Иногда его называют суперверсией.) В случае конфликта между версиями приоритет самых свежих версий в общем случае будет выше приоритета более старых версий. Например, если метка категории в каких-либо версиях отличается, будет использоваться текст в самой последней версии.
- **Последняя версия.** Выберите эту опцию, если хотите использовать самую свежую версию.
- **Задать версию.** Выберите эту опцию, если хотите использовать конкретную версию опроса.

Выбор всех версий полезен, если вы хотите, например, экспортировать данные наблюдений для нескольких версий, а в определения переменных и категорий были внесены изменения, подразумевающие, что данные наблюдений, собранные с одной версией, будут недопустимы в другой. Выбор всех версий, для которых вы хотите экспортировать данные наблюдений, подразумевает, что в общем случае одновременно можно экспортировать данные наблюдений, собранные с разными версиями, не сталкиваясь с ошибками допустимости из-за различий между версиями. Однако, в зависимости от изменений в версиях, некоторые ошибки допустимости все же могут встретиться.

Язык. Вопросы и связанный текст могут храниться в метаданных на нескольких языках. Для опроса можно использовать язык по умолчанию или задать конкретный язык. Если позиция окажется недоступной на заданном языке, будет использоваться язык по умолчанию.

Контекст. Выберите пользовательский контекст, который вы хотите использовать. Пользовательский контекст управляет тем, какие тексты будут выводиться. Например, выберите **Вопрос**, чтобы выводились тексты вопросов, или **Анализ**, чтобы выводились более краткие текстовые описания, вывод которых целесообразен при анализе данных.

Тип метки. Список уже определенных типов меток. Тип по умолчанию - **метка**, используемый для текстов вопросов в пользовательском контексте Вопрос и для описаний переменных в пользовательском контексте Анализ. Другие типы меток могут быть определены для инструкций, описаний и так далее.

Строка соединения с базой данных

При использовании узла Data Collection для импорта данных наблюдений из базы данных через OLE-DB или ODBC выберите на вкладке Файл **Правка**, чтобы открыть диалоговое окно Соединение, позволяющее настроить передаваемую провайдеру строку соединения, и точно настройте соединение.

Дополнительные свойства

При использовании узла Data Collection для импорта данных наблюдений из базы данных, где требуется вход в систему явным образом, выберите **Дополнительные**, чтобы ввести ID пользователя и пароль для доступа к источнику данных.

Импорт наборов множественных ответов

Наборы множественных ответов можно импортировать из Data Collection как наборы множественных дихотомий с отдельным полем флага для каждого возможного значения переменной. Например, если респондентов просят выбрать в списке, какие музеи они посетили, в этот набор должно входить отдельное поле флага для каждого указанного в списке музея.

После импорта данных наборы множественных ответов можно добавить или отредактировать с любого узла, содержащего вкладку Фильтр. Дополнительную информацию смотрите в разделе “Редактирование наборов множественных ответов” на стр. 143.

Импорт множественных ответов в одно поле (из потоков, созданных в прежних выпусках)

Потоки, созданные в прежних выпусках SPSS Modeler множественные ответы импортировались не описанным выше способом, а в одно поле, со значениями, разделяемыми запятыми. Этот метод все еще применяется для поддержки существующих потоков, но все такие потоки рекомендуется обновить, чтобы использовать новый метод.

Примечания к импорту столбцов Data Collection

Столбцы из данных Data Collection читаются в SPSS Modeler, как показано в следующей таблице.

Таблица 4. Сводка импорта столбцов Data Collection

Тип столбца Data Collection	Хранение SPSS Modeler	Шкала измерений
Логический флаг (да/нет)	Текстовое	Флаг (значения 0 и 1)
Категориальная	Текстовое	Номинальное
Date или Timestamp	Метка даты/времени	Непрерывное
Double (значение с плавающей запятой в заданном диапазоне)	Действительное число	Непрерывное

Таблица 4. Сводка импорта столбцов Data Collection (продолжение)

Тип столбца Data Collection	Хранение SPSS Modeler	Шкала измерений
Long (целочисленное значение в заданном диапазоне)	Целое число	Непрерывное
Text (произвольное текстовое описание)	Текстовое	Без типа
Level (показывает сетки или циклы в вопросе)	В VDATA не встречается и в SPSS Modeler не импортируется	
Object (двоичные данные, такие факсимильное изображение рукописного текста или звукозапись речи)	В SPSS Modeler не импортируется	
Нет (неизвестный тип)	В SPSS Modeler не импортируется	
Столбец Respondent.Serial (связывает уникальный ID с каждым респондентом)	Целое число	Без типа

Во избежание возможных несогласованностей между метками значений, читаемых из метаданных, и фактическими значениями все значения метаданных преобразуются в нижний регистр. Например, метка значения *E1720_years* будет преобразована в *e1720_years*.

Узел источника IBM Cognos BI

Узел источника IBM Cognos BI позволяет ввести данные базы данных Cognos BI или отчеты одного списка в сеанс исследования данных. Таким образом можно объединить возможности бизнес-аналитики Cognos с возможностями прогностической аналитики IBM SPSS Modeler. Можно импортировать реляционные данные, реляционные данные, моделируемые по измерению (dimensionally-modeled relational, DMR) и данные OLAP.

Находясь в сеансе соединения с сервером Cognos, сначала вы выбираете положение, из которого будут импортироваться данные или отчеты. Положение содержит модель Cognos и все папки, запросы, отчеты, представления, ярлыки, URL и определения заданий, связанные с данной моделью. Модель Cognos определяет правила, описания данных, взаимосвязи, направления и иерархические структуры бизнеса, а также другие административные задачи.

Если выполняется импорт данных, выберите объекты, которые вы хотите импортировать из выбранного пакета. В состав объектов, которые можно импортировать, входят темы запросов (представляющие таблицы базы данных) и отдельные позиции запросов (представляющие столбцы таблиц). Дополнительную информацию смотрите в разделе “Значки объектов Cognos” на стр. 37.

Если для пакета определены фильтры, можно импортировать один или несколько из них. Если импортируемый фильтр связан с импортируемыми данными, он применяется перед импортом данных. *Примечание*: Данные, подлежащие импорту, должны быть в формате UTF-8.

Если импортируются отчет, надо выбрать пакет или папку в пакете, содержащие один или несколько отчетов. Затем надо выбрать отдельный отчет, который вы хотите импортировать. *Примечание*: Можно импортировать отчеты только одного списка; несколько списков не поддерживаются.

Если параметры определены (как для объекта данных, так и для отчета), можно указать значения для этих параметров перед импортом объекта или отчета.

Значки объектов Cognos

Разнообразные типы объектов, которые можно импортировать из базы данных Cognos BI, представлены различными значками, что иллюстрирует следующая таблица.

Таблица 5. Значки объектов Cognos.

Значок	Объект
	Пакет
	Пространство имен
	Тема запроса
	Элемент запроса
	Измерение показателя
	Мера
	Измерение
	Уровень иерархии
	Уровень
	Фильтр
	Отчет
	Автономное вычисление

Импорт данных Cognos

Чтобы импортировать данные из базы данных IBM Cognos BI, при помощи вкладки Данные диалогового окна IBM Cognos BI убедитесь, что задан **Режим Данные**.

Подключение. Нажмите кнопку **Правка**, чтобы вывести диалоговое окно, где можно определить подробности нового соединения с Cognos, с помощью которого следует импортировать данные и отчеты. Если вы уже вошли в систему на сервер Cognos через IBM SPSS Modeler, то сможете также отредактировать и подробности текущего соединения. Дополнительную информацию смотрите в разделе “Подключения Cognos” на стр. 39.

Местоположение. Если уже установлено соединение с Cognos, нажмите кнопку **Правка** рядом с этим полем, чтобы вывести список доступных пакетов, из которых следует импортировать содержимое. Дополнительную информацию смотрите в разделе “Выбор положения Cognos” на стр. 39.

Содержимое. Содержит имя выбранного пакета вместе с пространствами имен, связанными с этим пакетом. Щелкните дважды по пространству имен, чтобы вывести объекты, которые вы можете импортировать. Объекты разных типов обозначаются различными значками. Дополнительную информацию смотрите в разделе “Значки объектов Cognos”.

Чтобы выбрать объекты для импорта, выберите объект и щелкните по верхней из двух стрелок вправо, чтобы переместить объект на панель **Поля для импорта**. При выборе предмета запроса импортируются все элементы запроса. Двойной щелчок по предмету запроса раскрывает его, так что вы можете выбрать один или несколько отдельных элементов запроса. Можно выбрать несколько элементов при помощи щелчка при нажатой клавише Ctrl (выбор отдельных элементов) или Shift (выбор блока элементов) и Ctrl-A (выбор всех элементов).

Чтобы выбрать фильтр для применения (если для пакета определены фильтры), перейдите к этому фильтру на панели Содержимое, выберите фильтр и щелкните по нижней из двух стрелок вправо для перемещения фильтра на панель **Применяемые фильтры**. Можно выбрать несколько фильтров при помощи щелчка при нажатой клавише Ctrl (выбор отдельных фильтров) или Shift (выбор блока фильтров).

Поля для импорта. Список объектов базы данных, выбранной вами для импорта в IBM SPSS Modeler. Если вам больше не нужен конкретный объект, выберите его и нажмите кнопку со стрелкой влево, чтобы вернуть его на панель **Содержимое**. Можно выбрать несколько элементов аналогично выбору **Содержимого**.

Фильтры для применения. Список фильтров, которые вы выбрали для применения к данным перед их импортом. Если вам больше не нужен конкретный фильтр, выберите его и нажмите кнопку со стрелкой влево, чтобы вернуть его на панель **Содержимое**. Можно выбрать несколько элементов аналогично выбору **Содержимого**.

Анализ важности независимых переменных. Если эта кнопка доступна, то для выбранного объекта параметры определены. Можно воспользоваться параметрами для внесения изменений (например, выполнения параметризированного отчета) перед импортом данных. Если параметры определены, но ни один из них не задан по умолчанию, но на кнопке показывается треугольник с предупреждением. Нажмите эту кнопку, чтобы вывести параметры и (необязательно) отредактировать их. Если кнопка недоступна, параметры для отчета не определены.

Объединить данные перед импортом. Включите этот переключатель, если вы хотите импортировать сводные данные вместо необработанных.

Импорт отчетов Cognos

Чтобы импортировать заранее определенный отчет из базы данных IBM Cognos BI, при помощи вкладки Данные диалогового окна IBM Cognos BI убедитесь, что задана **Режим Отчет**. *Примечание*: Можно импортировать отчеты только одного списка; несколько списков не поддерживаются.

Подключение. Нажмите кнопку **Правка**, чтобы вывести диалоговое окно, где можно определить подробности нового соединения с Cognos, с помощью которого следует импортировать данные и отчеты. Если вы уже вошли в систему на сервер Cognos через IBM SPSS Modeler, то сможете также отредактировать и подробности текущего соединения. Дополнительную информацию смотрите в разделе “Подключения Cognos” на стр. 39.

Местоположение. Если уже установлено соединение с Cognos, нажмите кнопку **Правка** рядом с этим полем, чтобы вывести список доступных пакетов, из которых следует импортировать содержимое. Дополнительную информацию смотрите в разделе “Выбор положения Cognos” на стр. 39.

Содержание. Выводит имя выбранного пакета или папки, содержащей отчеты. Перейдите к конкретному отчету, выберите его и нажмите кнопку со стрелкой вправо, чтобы переместить его в поле **Отчет для импорта**.

Отчет для импорта. Показывает отчет, выбранный вами для импорта в IBM SPSS Modeler. Если отчет вам больше не нужен, выберите его и нажмите кнопку со стрелкой влево, чтобы вернуть его на панель **Содержимое**, либо переместите в это поле другой отчет.

Анализ важности независимых переменных. Если эта кнопка доступна, то для выбранного отчета параметры определены. При помощи параметров перед импортом отчета в него можно внести корректировки (например, задать дату начала и окончания для данных отчета). Если параметры определены, но ни один из них не задан по умолчанию, но на кнопке показывается треугольник с предупреждением. Нажмите эту кнопку, чтобы вывести параметры и (необязательно) отредактировать их. Если кнопка недоступна, параметры для отчета не определены.

Подключения Cognos

Диалоговое окно Соединения Cognos позволяет выбрать сервер Cognos BI для импорта или экспорта объектов баз данных.

URL сервера Cognos Введите URL сервера Cognos BI для импорта или экспорта. Это значение свойства среды "URI внешнего диспетчера" конфигурации IBM Cognos на сервере Cognos BI. Обратитесь к администратору системы Cognos, если вы не знаете, какой URL использовать.

Режим Если вы хотите войти в систему с конкретным пространством имен Cognos, именем пользователя и паролем (например, вход в качестве администратора), выберите **Задать регистрационные данные**. Чтобы войти без ввода учетной записи, выберите **Использовать анонимное подключение**. В этом случае другие поля не заполняются.

Другой вариант - если у вас есть регистрационные данные IBM Cognos, хранимые в репозитории IBM SPSS Collaboration and Deployment Services, их можно использовать вместо ввода имени пользователя и пароля или создания анонимного соединения. Чтобы использовать существующие регистрационные данные, выберите опцию **Хранимые регистрационные данные** и введите нужное **Имя регистрационных данных** или выберите его.

Пространство имен Cognos моделируется доменом в IBM SPSS Collaboration and Deployment Services.

ID пространства имен Укажите провайдер аутентификации защиты Cognos, используемый для входа в систему сервера. Провайдер аутентификации используется для определения и поддержки пользователей, групп и ролей, а также для управления процессом аутентификации. Заметим, что это ID пространства имен, а не имя пространства имен (ID не всегда совпадает с именем).

Имя пользователя Введите имя пользователя Cognos для входа на сервер.

Пароль Введите пароль, связанный с указанным именем пользователя.

Сохранить как умолчания Нажмите эту кнопку, чтобы сохранить эти параметры в качестве параметров по умолчанию, чтобы не вводить их каждый раз заново.

Выбор положения Cognos

Диалоговое окно Указать расположение позволяет выбрать пакет Cognos, с которого необходимо импортировать данные или пакет/папку, из которых необходимо импортировать отчеты.

Общедоступные папки. Если выполняется импорт данных, эта опция возвращает список пакетов и папок, доступных на выбранном сервере. Выберите пакет, который вы хотите использовать, и нажмите кнопку **OK**. Для одного узла источника Cognos BI можно выбрать только один пакет.

Если выполняется импорт отчетов, эта опция возвращает список папок и пакетов, содержащих отчеты, которые доступны на выбранном сервере. Выберите папку пакета или отчета и нажмите кнопку **OK**. Для одного узла источника Cognos BI можно выбрать только один пакет или отчет, хотя папки отчетов могут содержать другие папки отчетов и отдельные отчеты.

Указание параметров для данных или отчетов

Если параметры определены в Cognos BI (как для объекта данных, так и для отчета), можно указать значения для этих параметров перед импортом объекта или отчета. В качестве примеров параметров для отчета можно привести начальные и конечные даты для содержимого отчета.

Имя. Имя параметра, как оно указано в базе данных Cognos BI.

Тип. Описание параметра.

Значение. Значение для назначения параметру. Чтобы ввести или изменить значение, дважды щелкните его ячейку в таблице. Проверка значений здесь не выполняется, поэтому любые недопустимые значения выявляются во время выполнения.

Автоматически удалить недействительные параметры из таблицы. Этот параметр выбирается по умолчанию и заменит любые недействительные параметры, которые обнаружатся в объекте данных или отчете.

Узел источника IBM Cognos TM1

Узел источника IBM Cognos TM1 позволяет перенести данные Cognos TM1 в сеанс исследования данных. Таким образом можно объединить возможности планирования на уровне предприятия Cognos с возможностями прогностической аналитики IBM SPSS Modeler. Вы можете импортировать версию данных многомерного куба OLAP меньшей размерности.

Примечание: У пользователя TM1 должны быть следующие разрешения: привилегия записи для кубов, привилегия чтения для измерений и привилегия записи для элементов измерений.

Необходимо изменить данные в TM1 до импорта этих данных.

Примечание: Импортируемы данные должны быть в формате UTF-8.

Из соединений хоста администрирования IBM Cognos TM1 сначала выбирается сервер TM1, с которого будут импортироваться данные; сервер содержит один или несколько кубов TM1. Затем выбирается нужный куб, а в этом кубе выбираются столбцы и строки для импорта.

Примечание: Прежде чем использовать узлы источник TM1 или узлы экспорта в SPSS Modeler, необходимо верифицировать некоторые параметры в файле `tm1s.cfg`, то есть в файле конфигурации сервера TM1 в корневом каталоге сервера TM1.

- HTTPPortNumber - задать допустимый номер порта; обычно от 1 до 65535.
- UseSSL - если задать для этого параметра значение *True*, в качестве транспортного протокола будет использоваться HTTPS. В этом случае необходимо импортировать сертификацию TM1 в SPSS Modeler Server JRE.

Импорт данных IBM Cognos TM1

Чтобы импортировать данные в базу данных IBM Cognos TM1, на вкладке Данные диалогового окна IBM Cognos TM1 выберите соответствующий хост администрирования TM1 и связанный сервер, куб и подробности о данных.

Примечание: Перед импортом данных необходимо произвести некоторую предварительную обработку в TM1, чтобы обеспечить формат данных, распознаваемый в IBM SPSS Modeler. Сюда входит фильтрование ваших данных с помощью редактора поднаборов, чтобы привести представление к правильному размеру и нужному для импорта виду.

Заметим, что нулевые (0) значения, импортированные из TM1, будут рассматриваться как значения "null" (TM1 не различает пустые и нулевые значения). Кроме того, обратите внимание на то, что нечисловые

данные (или метаданные) из *регулярных измерений* можно импортировать в IBM SPSS Modeler. Однако импорта нечисловых *показателей* в данный момент не поддерживается.

Хост администрирования Введите URL хоста администрирования, где установлен сервер TM1, с которым вы хотите соединиться. Хост администрирования определяется как один URL для всех серверов TM1. Из положения с этим URL можно обнаружить все серверы IBM Cognos TM1, установленные и запущенные в вашей среде, и связаться с ними.

Сервер TM1 Когда соединение с хостом администрирования будет установлено, выберите сервер, содержащий данные, которые вы хотите импортировать, и нажмите кнопку **Вход в систему**. Если прежде вы не соединялись с этим сервером, появится предложение ввести **Имя пользователя** и **Пароль**; другой вариант - найти ранее введенные сведения входа в систему, сохраненные как **Хранимые регистрационные данные**.

Выберите представление кубов TM1 для импорта Содержит имена кубов на сервере TM1, откуда можно импортировать данные. Дважды щелкните по кубу для просмотра данных, которые можно импортировать.

Примечание: В IBM SPSS Modeler можно импортировать только кубы с измерением.

Чтобы определить данные для импорта, выберите представление и щелкните по стрелке вправо, чтобы перенести его на панель **Представление для импорта**. Если нужное представление не видно, дважды щелкните по кубу, чтобы раскрыть его список представлений.

Измерения строк. Содержит имя измерения строк в данных, которые вы выбрали для импорта. Прокрутите список уровней и выберите нужный.

Измерение столбцов Содержит имя измерения столбцов в данных, которые вы выбрали для импорта. Прокрутите список уровней и выберите нужный.

Контекстные измерения Только для вывода. Содержит контекстные измерения, относящиеся к выбранным столбцам и строкам.

Узел источника SAS

Примечание: Эта возможность доступна в SPSS Modeler Professional и SPSS Modeler Premium.

Узел источника SAS позволяет перенести данные SAS в сеанс исследования данных. Возможен импорт файлов четырех типов:

- SAS for Windows/OS2 (*.sd2*)
- SAS for UNIX (*.ssd*)
- Транспортный файл SAS (*.tpt*)
- SAS версии 7/8/9 (*.sas7bdat*)

При импорте данных все переменные сохраняются и никакие типы переменных не изменяются. Выбираются все наблюдения.

Задание опций для узла источника SAS

Импорт. Выберите, какой тип файла SAS транспортировать. Можно выбрать **SAS for Windows/OS2 (.sd2)**, **SAS for UNIX (.SSD)**, **Транспортный файл SAS (.tpt)** или **SAS Версии 7/8/9 (.sas7bdat)**.

Файл импорта. Задайте имя файла. Можно ввести имя файла или нажать кнопку с многоточием (**...**), чтобы найти положение файла.

Элемент. Выберите элемент для импорта из выбранного выше транспортного файла SAS. Можно ввести имя элемента или, нажав кнопку **Выбрать**, просмотреть все элементы в файле.

Читать пользовательские форматы из файла данных SAS. Выберите эту опцию для чтения пользовательских форматов. В файлах SAS хранятся данные и форматы данных (таких как метки переменных) в различных файлах. Чаще всего предпочтителен и импорт форматов. Однако при большом наборе данных эту опцию желательно отключить, чтобы сэкономить память.

Файл формата. Если требуется файл формата, это текстовое поле активируется. Можно ввести имя файла или нажать кнопку с многоточием (**...**), чтобы найти положение файла.

Имена переменных. Выберите метод обработки имен и меток переменных после импорта из файла SAS. Метаданные, выбираемые вами здесь для включения, сохраняются в течение всей вашей работы в IBM SPSS Modeler и могут быть снова экспортированы для использования в SAS.

- **Читать имена и метки.** Выберите эту опцию для чтения имен и меток переменных в IBM SPSS Modeler. По умолчанию эта опция включена, а имена переменных вводятся на узле типа. Метки могут выводиться на диаграммах Построителя выражений, в браузерах моделей и выводе других типов, в зависимости от опций, заданных в диалоговом окне свойств потока.
- **Читать метки как имена.** Выберите эту опцию для чтения описательных меток переменных из файла SAS вместо кратких имен полей и для использования этих меток в качестве имен переменных в IBM SPSS Modeler.

Узел источника Excel

Узел Excel позволяет импортировать данные из Microsoft Excel в формате файлов .xlsx.

Тип файла. Выберите тип файла Excel, который вы импортируете.

Файл импорта. Задает имя и положение файла электронной таблицы для импорта.

Использовать именованный диапазон. Позволяет задать именованный диапазон ячеек в качестве определяемого в рабочей таблице Excel. Нажмите кнопку с многоточием (**...**), чтобы выбрать диапазон из списка доступных диапазонов. При использовании именованного диапазона другие параметры рабочих таблиц и диапазонов данных больше не применяются и поэтому становятся, недоступны.

Выбрать рабочую таблицу. Задает рабочую таблицу для импорта либо по индексу, либо имени.

- **По индексу.** Задайте значение индекса для рабочей таблицы, которую вы хотите импортировать, начиная с 0 для первой рабочей таблицы, 1 для второй рабочей таблицы и так далее.
- **По именам.** Задайте имя рабочей таблицы, которую вы хотите импортировать. Нажмите кнопку с многоточием (**...**), чтобы выбрать рабочую таблицу в списке доступных.

Диапазон для рабочей таблицы. Можно импортировать данные, начиная с первой непустой строки или явного диапазона ячеек.

- **Диапазон начинается с первой непустой строки.** Находит первую непустую ячейку и использует ее в качестве верхнего левого угла для диапазона данных.
- **Явный диапазон ячеек.** Позволяет задать явный диапазон в строках и столбцах. Например, чтобы задать диапазон Excel `A1:D5`, в первом поле можно ввести `A1`, а во втором поле - `D5` (или, как вариант, `R1C1` и `R5C4`). Будут возвращены все строки в заданном диапазоне, включая пустые строки.

Включить пустые строки. Для обработки нескольких вхождений пустых строк можно выбрать **Остановить чтение** или **Возвратить пустые строки**, чтобы продолжить чтение всех данных до конца рабочей таблицы, включая пустые строки.

Первая строка содержит имена столбцов. Указывает, чтоб первую строку в заданном диапазоне следует использовать в качестве имен полей (столбцов). Если эта опция не выбрана, имена полей будут генерироваться автоматически.

Хранение полей и шкала измерений

При чтении значений из Excel поля с числовым хранением считываются со шкалой измерений про умолчанию *Непрерывная*, а строковые поля считываются как *Номинальные*. Шкалу измерений (непрерывную сопоставительно с номинальной) можно изменить вручную на вкладке Тип, но хранение определяется автоматически (хотя при необходимости его можно изменить при помощи функции преобразований, такой как `to_integer`, на узле фильтрации или извлечения). Дополнительную информацию смотрите в разделе “Задание форматирования и системы хранения для полей” на стр. 8.

По умолчанию поля, содержащие смесь числовых и строковых значений, считываются как числовые, из-за чего все строковые значения преобразуются в IBM SPSS Modeler в пустые (системные пропущенные) значения. Это связано с тем, что, в отличие от Excel, IBM SPSS Modeler не разрешает смешанные типы хранения в поле. Чтобы избежать этого, в электронной таблице Excel можно задать вручную формат ячейки как `Текст`, что приведет к считыванию значений (включая числовые) как строковых значений.

Узел источника XML

Примечание: Эта возможность доступна в SPSS Modeler Professional и SPSS Modeler Premium.

Используйте узел источника XML для импорта данных из файла формата XML в поток IBM SPSS Modeler. XML - это стандартный язык для обмена данными, и многие организации выбрали для этого именно данный формат. Например, правительственной налоговой службе может потребоваться проанализировать налоговые декларации, поданные через интернет в формате XML (смотрите http://www.w3.org/standards/xml/).

Импорт данных XML в поток IBM SPSS Modeler позволяет выполнить многие прогнозирующие аналитические функции для источника. Данные XML анализируются и преобразуются в табличный формат, в котором столбцы соответствуют различным уровням вложенности элементов и атрибутов XML. Элементы XML выводятся в формате XPath (смотрите http://www.w3.org/TR/xpath20/).

Прочесть отдельный файл По умолчанию SPSS Modeler читает один файл, который вы задали в поле **Источник данных XML**.

Прочесть все файлы XML в каталоге Выберите эту опцию, если вы хотите прочесть все файлы XML в конкретном каталоге. Задайте положение в появившемся поле **Каталог**. Включите переключатель **Включить подкаталоги**, чтобы дополнительно читать файлы XML из всех подкаталогов заданного каталога.

Источник данных XML Введите полный путь и имя файла источника XML, который вы хотите импортировать, или используйте кнопку Просмотр, чтобы найти этот файл.

Схема XML (Необязательно) Задайте полный путь и имя файла XSD или DTD, из которого будет читаться структура XML, или используйте кнопку Просмотр, чтобы найти этот файл. Если оставить это поле пустым, структура читается из файла источника XML. У файла XSD или DTD может быть несколько корневых элементов. В этом случае, когда фокус переходит на другое поле, выводится диалоговое окно, в котором можно выбрать корневой элемент для использования. Дополнительную информацию смотрите в разделе “Выбор из нескольких корневых элементов” на стр. 44.

Примечание: SPSS Modeler игнорирует индикаторы XSD.

Структура XML Иерархическое дерево, показывающее структуру файла источника XML (или схему, если вы ее задали в поле **Схема XML**). Для определения границы записи выберите элемент и нажмите кнопку с правой стрелкой, чтобы скопировать его в поле **Записи**.

Вывести атрибуты Выводит или скрывает атрибуты элементов XML в поле **Структура XML**.

Записи (выражение XPath) Показывает синтаксис XPath для элемента, скопированного из поля структуры XML. Затем этот элемент выделяется в структуре XML и определяет границу записи. Каждый раз, когда этот элемент встречается в файле источника, создается новая запись. Если это поле пустое, первый дочерний элемент корня используется как граница записи.

Читать все данные По умолчанию все данные в файле источника читаются в поток.

Задать данные для чтения Выберите эту опцию, если вы хотите импортировать отдельные элементы и/или атрибуты. Выбор этой опции включает таблицу Поля, где можно задать данные, которые вы хотите импортировать.

Поля В этой таблице перечисляются элементы и атрибуты, выбранные для импорта, если вы выбрали опцию **Задать данные для чтения**. Вы можете ввести синтаксис XPath элемента или атрибута непосредственно в столбец XPath или выбрать элемент или атрибут в структуре XML и нажать кнопку с правой кнопкой для копирования элемента в таблицу. Чтобы скопировать все дочерние элементы и атрибуты элемента, выберите этот элемент в структуре XML и нажмите кнопку с двойной стрелкой.

- **XPath** Синтаксис XPath элементов для импорта.
- **Положение** Положение структуры XML элементов, которые будут импортироваться. **Фиксированный путь** показывает путь элемента относительно элемента, выделенного в структуре XML (или первого дочернего элемента от корня, если никакой элемент не выделен). **Любое положение** обозначает элемент с данным именем в любом положении структуры XML. Выводится **Пользовательское**, если вы вводите положение непосредственно в столбец XPath.

Выбор из нескольких корневых элементов

Тогда как правильно сформированный файл XML может содержать только один корневой элемент, у файла XSD или DTD может быть несколько корневых элементов. Если один из таких корневых элементов совпадает с корневым элементом файла источника XML, он и используется, в противном случае вам нужно выбрать один из них для использования.

Выберите корневой элемент для вывода. Выберите корневой элемент, который вы хотите использовать. По умолчанию это первый корневой элемент в структуре XSD или DTD.

Удаление нежелательных пробелов из данных источника XML

Разрывы строки в исходных данных XML можно реализовать комбинацией символов `[CR][LF]`. В некоторых случаях эти разрывы строки могут встречаться в середине текстовой строки, например:

```
<описание> Подробное изучение создания прикладных программ[CR][LF]
с XML.</описание>
```

В некоторых прикладных программах при открытии файла эти разрывы строки не видны, например, в Web-браузере. Однако когда эти данные читаются в поток через узел источника XML, разрывы строки преобразуются в ряд символов пробела.

Это можно исправить, используя узел заполнителя для удаления этих нежелательных пробелов:

Ниже приведен пример, как этого можно достигнуть:

1. Присоедините узел Заполнитель к узлу источника XML.
2. Откройте узел Заполнитель и используйте инструмент выбора полей для выбора поля с нежелательными пробелами.
3. Задайте для опции **Замена** значение **На основе условия**, а для **Условия - true**.
4. В поле **Заменить на** введите `replace(" ","",@FIELD)` и нажмите кнопку OK.
5. Присоедините узел Таблица к узлу Заполнитель и запустите поток.

В выходе узла Таблица теперь появится текст без дополнительных пробелов.

Узел пользовательского ввода

Узел пользовательского ввода предоставляет простой способ создания синтетических данных - либо с нуля, либо путем изменения существующих данных. Этот способ полезен, например, если вы хотите создать тестовый набор данных для моделирования.

Создание данных с нуля

Узел пользовательского ввода доступен на палитре источников и может быть непосредственно добавлен на холст потока.

1. Перейдите на вкладку **Источники** палитры узлов.
2. Перетащите узел пользовательского ввода на холст источников или дважды щелкните по нему для этого.
3. При двойном щелчке откроется диалоговое окно, в котором можно задать поля и значения.

Примечание: Узлы пользовательского ввода, выбранные на палитре источников, будут пустыми, без информации о полях и данных. Это позволяет создавать синтетические данные с нуля.

Генерирование данных из существующего источника данных

Узел пользовательского ввода можно сгенерировать также из любого не конечного узла в потоке:

1. Решите, в какой точке потока вы хотите заменить узел.
2. Щелкните правой кнопкой мыши по узлу, из которого будут взяты данные на узел пользовательского ввода, и выберите в меню **Генерировать узел пользовательского ввода**.
3. Появится узел пользовательского ввода со всеми присоединенными к нему процессами нижележащего уровня, заменяющий существующий узел в данной точке вашего потока данных. При генерировании этот узел наследует всю структуру данных и информацию о типах полей (если она доступна) из метаданных.

Примечание: Если данные не были пропущены по всем узлам потока, узлы еще не полностью инстанциированы, что означает возможную недоступность значений хранения и данных при замене узла на узел пользовательского ввода.

Задание опций для узла пользовательского ввода

Диалоговое окно для узла пользовательского ввода содержит несколько инструментов, которые можно использовать для ввода значений и определения структуры для синтетических данных. Для сгенерированного узла таблица на вкладке Данные содержит имена полей из исходного источника данных. Для узла, добавленного с палитры источника, эта таблица пустая. Используя опции таблицы, можно выполнить следующие задачи:

- Добавить новые поля, используя кнопку Добавить новое поле справа в таблице.
- Переименовать существующие поля.
- Задать систему хранения данных для каждого поля.
- Задать значения.
- Изменить порядок полей на экране.

Ввод данных

Для каждого поля можно определить значения или вставить значения из исходного набора данных с помощью кнопки средства выбора значения справа от таблицы. Более подробную информацию о задании значений смотрите в описанных ниже правилах. По выбору можно оставить поле пустым, такие поля заполняются системным значением null (`$null$`).

Чтобы задать строковые значения, просто введите их в столбце Значения, разделяя пробелами:

```
Fred Ethel Martin
```

Строки, содержащие пробелы, можно заключить в двойные кавычки:

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

Для числовых полей можно или ввести несколько значений, так же перечисляя их через пробел...:

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

... или задать тот же ряд чисел, указав его пределы (10, 20) и шаг между соседними значениями (2). Этим способом можно ввести следующее:

```
10,20,2
```

Эти два способа можно объединять, включая один в состав другого, например:

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

Эта запись создает следующие значения:

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

Значения даты и времени можно ввести, используя текущий формат по умолчанию, выбранный в диалоговом окне Свойства потока, например:

```
11:04:00 11:05:00 11:06:00
2007-03-14 2007-03-15 2007-03-16
```

Для значений отметки времени, у которых есть компоненты и даты, и времени, должны использоваться двойные кавычки:

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

Дополнительные подробности смотрите в комментариях о системе хранения данных ниже.

Сгенерировать данные. Позволяет задать, как будут генерироваться записи при запуске потока.

- **Все комбинации**. Генерирует записи, содержащие все возможные комбинации значений полей, так что значение каждого поля появится в нескольких записях. Иногда это приводит к генерированию большего количества данных, чем нужно, поэтому часто после этого узла запускают узел выборки.
- **По порядку**. Генерирует записи в том порядке, в котором задаются значения полей данных. Каждое значение поля появляется только в одной записи. Общее количество записей равно максимальному количеству значений для одного поля. Когда у поля меньше значений, чем максимальное число, вставляется не определенное значение ($null$).

Показать пример

Например, следующие элементы данных сгенерируют записи, перечисленные в двух следующих примерах таблиц.

- **Возраст.** `30,60,10`
- **BP.** `НИЗКИЙ`
- **Холестерол.** `НОРМАЛЬНЫЙ ВЫСОКИЙ`
- **Препарат.** (оставлено пустым)

Таблица 6. Для поля генерирования данных задано Все комбинации.

Возраст	BP	Cholesterol	Препарат
30	НИЗКИЙ	NORMAL	$null$
30	НИЗКИЙ	HIGH	$null$
40	НИЗКИЙ	NORMAL	$null$
40	НИЗКИЙ	HIGH	$null$

Таблица 6. Для поля генерирования данных задано Все комбинации (продолжение).

Возраст	BP	Cholesterol	Препарат
50	НИЗКИЙ	NORMAL	$null$
50	НИЗКИЙ	HIGH	$null$
60	НИЗКИЙ	NORMAL	$null$
60	НИЗКИЙ	HIGH	$null$

Таблица 7. Для поля генерирования данных задано По порядку.

Возраст	BP	Cholesterol	Препарат
30	НИЗКИЙ	NORMAL	$null$
40	$null$	HIGH	$null$
50	$null$	$null$	$null$
60	$null$	$null$	$null$

Хранение данных

Хранение описывает способ хранения данных в поле. Например, в поле со значениями 1 и 0 хранятся целочисленные данные. В этом оно отличается от шкалы измерений, которая описывает использование данных и на хранение не влияет. Например, вы можете захотеть задать шкалу измерения для целочисленного поля со значениями 1 и 0 как для *флага*. Обычно это означает, что 1 = *True*, а 0 = *False*. Хранение должно быть определено на источнике, тогда как шкалу измерения можно изменить при помощи узла Тип в любой точке в потоке. Дополнительную информацию смотрите в разделе “Уровни измерения” на стр. 129.

Доступны следующие типы хранения:

- **Строка** Используется для полей, содержащих нечисловые данные, они называются также алфавитно-цифровыми. Строка может содержать любую последовательность символов, например: *поле*, *Класс 2* или *1234*. Имейте в виду, что числа в строках нельзя использовать в вычислениях.
- **Целое** Поле с целочисленными значениями.
- **Действительное** Значения представляют собой числа с дробной частью (не только целые). Формат вывода задается в диалоговом окне Свойства потока и может быть переопределен для отдельных полей на узле Тип (вкладка Формат).
- **Дата** Значения дат, задаваемые в стандартном формате, таком как год, месяц и день (например: `2007-09-26`). Конкретный формат задается в диалоговом окне Свойства потока.
- **Время** Время, измеряемое как продолжительность. Например, вызов службы, продолжающийся 1 час, 26 минут и 38 секунд, может быть представлен как `01:26:38`, в зависимости от текущего формата времени, заданного в диалоговом окне Свойства потока.
- **Отметка времени** Значения, содержащие составляющие даты и времени, например: `2007-09-26 09:04:00` (тоже в зависимости от текущих форматов даты и времени в диалоговом окне Свойства потока). Имейте в виду, что значения отметок времени может потребоваться заключить в двойные кавычки, чтобы они интерпретировались как одно значение, а не как отдельные значения даты и времени. (Это применяется, например, при вводе значений на узле пользовательского ввода.)
- **Список** Введенное в SPSS Modeler версии 17 вместе с новыми шкалами измерений, Геопространственная и Собрание, поле хранения Список содержит несколько значений для одной записи. Существуют списочные версии всех других типов хранения.

Таблица 8. Значки типов хранения списков

Значок	Тип хранения
[A]	Список строковых
[ ]	Список целочисленных
[ ]	Список действительных
[ ]	Список переменных времени
[ ]	Список переменных даты
[ ]	Список переменных отметки времени
[[]]	Введите значение большее нуля

Кроме этого, для использования со шкалой измерения Собрание существуют списочные версии следующих шкал измерения.

Таблица 9. Значки уровня измерения списков

Значок	Шкала измерения
[ ]	Список количественных
[ ]	Список категориальных
[ ]	Список флагов
[ ]	Список номинальных
[ ]	Список порядковых

Списки можно импортировать в SPSS Modeler на одном из трех узлов источников (Analytic Server, Геопространственный или Файл переменных) или создать в вашем потоке с помощью узлов операций с полями Извлечение или Заполнитель.

Дополнительную информацию о списках и их взаимодействии со шкалами измерений Собрание и Геопространственная смотрите в разделе “Система хранения списков и связанные шкалы измерений” на стр. 11

Преобразования хранения. Можно преобразовать хранение для поля при помощи разнообразных функций преобразования, таких как `to_string` и `to_integer`, на узле заполнения. Дополнительную информацию смотрите в разделе “Преобразование хранения при помощи узла заполнения” на стр. 152. Имейте в виду, что функции преобразования (и любые другие функции, которым требуется конкретный тип ввода, такой как значение даты и времени) зависят от текущих форматов, заданных в диалоговом окне Свойства потока. Например, если вы хотите преобразовать строковое поле со значениями *Jan 2003*, *Feb 2003* (и так далее) в хранение даты, в качестве формата дат по умолчанию для потока выберите **МЕС ГГГГ**. Функции преобразования доступны также с узла извлечения, для временного преобразования во время вычисления операции извлечения. При помощи узла извлечения можно выполнять и другие операции с данными, такие как перекодирование строковых полей с категориальными значениями. Дополнительную информацию смотрите в разделе “Перекодирование значений при помощи узла извлечения” на стр. 151.

Чтение данных смешанных типов. Имейте в виду, что при чтении значений в полях с числовым хранением (целых чисел, действительных чисел, времени, отметок времени или дат) все нечисловые значения задаются как пустые или системные пропущенные. Это связано с тем, что в отличие от некоторых прикладных программ IBM SPSS Modeler не разрешает смешанные типы хранения в поле. Во избежание этого все поля с данными смешанных типов следует считывать как строки, изменив тип хранения либо на узле источника, либо во внешней прикладной программе, как требуется.

Примечание: Сгенерированные узлы пользовательского ввода могут уже содержать информацию хранения, полученную от узла источника, если он инстанциирован. Не инстанциированный узел не содержит информацию о типе хранения или использования.

Правила для задания значений

Для символических полей между несколькими значениями нужно оставлять пробелы, например:

```
ВЫСОКИЙ СРЕДНИЙ НИЗКИЙ
```

Для числовых полей можно или ввести несколько значений, так же перечисляя их через пробел...:

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

... или задать тот же ряд чисел, указав его пределы (10, 20) и шаг между соседними значениями (2). Этим способом можно ввести следующее:

```
10,20,2
```

Эти два способа можно объединять, включая один в состав другого, например:

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

Эта запись создает следующие значения:

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

Узел Генерирование имитации

Узел Генерирование имитации предоставляет простой способ сгенерировать имитированные данные, или без исторических данных, с использованием статистических распределений, заданных пользователем, или автоматически с использованием распределений, полученных в результате запуска узла Имитация Подгонка на существующих хронологических данных. Генерирование имитированных данных полезно, если вам нужно оценить результат работы предиктивной модели при наличии неопределенности во входных полях модели.

Создание данных без хронологических данных

Узел Генерирование имитации доступен на палитре Источники и может быть непосредственно добавлен на холст потока.

1. Перейдите на вкладку **Источники** палитры узлов.
2. Перетащите узел Генерирование имитации на холст потока или дважды щелкните по нему для этого.
3. При двойном щелчке откроется диалоговое окно, в котором можно задать поля, типы хранения, статистические распределения и параметры распределений.

Примечание: Узлы Генерирование имитации, выбранные на палитре Источники, будут пустыми, без информации о полях и распределениях. Это позволяет полностью создать имитированные данные без исторических данных.

Генерирование имитированных данных с использованием существующих исторических данных

Узел Генерирование имитации можно также создать путем выполнения конечного узла Подгонка имитации:

1. Щелкните правой кнопкой мыши по узлу Подгонка имитации и выберите из меню **Выполнить**.
2. Узел Генерирование имитации появится на холсте потока со связью обновления с узлом Подгонка имитации.
3. Сгенерированный узел Генерирование имитации наследует всю информацию о полях, типах хранения и статистических распределениях от узла Подгонка имитации.

Определение связи обновления с узлом подгонки имитации

Вы можете создать связь между узлом Генерирование имитации и узлом Подгонка имитации. Это полезно, если вы хотите обновить одно или несколько полей информацией о наиболее точно соответствующем распределении, определенном в ходе подгонки под хронологические данные.

1. Щелкните правой кнопкой мыши по узлу Генерирование имитации.
2. В меню выберите **Определить связь обновления**. Указатель изменится на указатель Ссылка.
3. Щелкните по другому узлу. Если этот узел - Подгонка имитации, ссылка будет создана. Если это другой узел (не Подгонка имитации), ссылка не будет создана, и указатель изменит форму на нормальную.

Если поля в узле Подгонка имитации отличаются от полей в узле Генерирование имитации, появится сообщение с информацией об отличиях.

Когда узел Подгонка имитации используется для обновления связанного с ним узла Генерирование имитации, результат зависит от того, присутствуют ли одни и те же поля в обоих узлах и разблокированы ли эти поля в узле Генерирование имитации. Результаты обновления узла Подгонка имитации показаны в следующей таблице.

Таблица 10. Результаты обновления узла Генерирование имитации

	Поле в имитации	**Узел подгонки**
Поле в узле Генерирование имитации	**Наличие**	**Пропущенные**
Присутствует и разблокировано.	Поле перезаписывается.	Поле удаляется.
Пропущенные.	Поле добавляется.	Без изменений.
Присутствует и заблокировано.	Распределение поля не перезаписывается. Информация в диалоговом поле Подробности подгонки обновляется, как и корреляции.	Поле не перезаписывается. Для корреляций задаются нулевые значения.
Переключатель **Не очищать Мин/Макс при повторной подгонке** включен.	Поле перезаписывается, кроме	значений в столбце Мин, Макс.
Переключатель **Не пересчитывать корреляции при повторной подгонке** включен.	Если поле разблокировано, оно перезаписывается.	Корреляции не перезаписываются.

Удаление связи обновления с узлом подгонки имитации

Чтобы удалить связь между узлом Генерирование имитации и узлом Подгонка имитации, выполните следующие действия:

1. Щелкните правой кнопкой мыши по узлу Генерирование имитации.
2. В меню выберите **Удалить связь обновления**. Связь будет удалена.

Задание опций для узла Генерирование имитации

Опции на вкладке Данные диалогового окна узла Генерирование имитации позволяют:

- Просматривать, задавать и редактировать информацию статистического распределения для полей.
- Просматривать, задавать и редактировать корреляции между полями.
- Задать число итераций и наблюдений для имитации.

Выбрать элемент. Позволяет переключаться между тремя представлениями узла Генерирование имитации: Имитированные поля, Корреляции и Дополнительные опции.

Представление Имитированные поля

Если узел Генерирование имитации сгенерирован или изменен из узла Подгонка имитации с использованием хронологических данных, в представлении Имитированные поля можно просматривать и редактировать информацию статистического распределения для каждого поля. Следующая информация о каждом поле копируется на вкладку **Типы** узла Генерирование имитации из узла Подгонка имитации:

- Шкала измерения
- Значения
- Пропущенные
- Проверить
- Роль

Если у вас нет хронологических данных, можно определить поля и задать их распределения, выбрав тип хранения, а затем выбрав тип распределения и задав требуемые параметры. Генерирование данных этим способом означает, что информация об уровне измерения каждого поля не будет доступна до создания экземпляра данных, например, на вкладке **Типы** или в узле Типы.

Представление Имитированные поля содержит несколько инструментов, которые можно использовать для выполнения следующих задач:

- Добавлять и удалять поля.
- Изменить порядок полей на экране.
- Задать тип хранения для каждого поля.
- Задать статистическое распределение для каждого поля.
- Задать значения параметров для статистического распределения каждого поля.

Имитированные поля. Эта таблица содержит одну пустую строку, если узел Генерирование имитации добавлен на холст потока из палитры Источники. Если отредактировать эту строку, в конец таблицы будет добавлена новая пустая строка. Если узел Генерирование имитации создан из узла Подгонка имитации, в этой таблице будет по одной строке для каждого поля хронологических данных. Чтобы добавить в таблицу дополнительные строки, щелкните по значку **Добавить новое поле**.

Таблица Имитированные поля состоит из следующих столбцов:

- **Поле.** Содержит имена полей. Имена полей можно редактировать путем ввода значений в ячейки.
- **Хранение.** Ячейки в этом столбце содержат выпадающие списки типов хранения. Доступные типы хранения - **Строка**, **Целое**, **Действительное число**, **Время**, **Дата** и **Отметка времени**. Выбор типа хранения определяет, какие распределения будут доступны в столбце Распределение. Если узел Генерирование имитации создан из узла Подгонка имитации, тип хранения копируется из узла Подгонка имитации.

  Примечание: Для полей с типами хранения даты/времени следует указывать параметры распределения в виде целых чисел. Например, чтобы задать в качестве средней даты 1 января 1970 года, укажите целое число 0. Это целое число со знаком соответствует числу секунд после (или до) полуночи 1 января 1970 года.

- **Статус.** Значки в столбце Статус отражают статус согласия для каждого поля.

Для поля не задано распределение, или же не хватает одного или нескольких параметров распределения. Чтобы запустить имитацию, необходимо задать для этого поля распределение и ввести допустимые значения для параметров.

Для поля задано наиболее точно соответствующее распределение.
Примечание: Этот значок может выводиться, только если узел Генерирование имитации создан из узла Подгонка имитации.

Наиболее точно соответствующее распределение заменено на альтернативное распределение из диалогового подокна Детали подгонки. Дополнительную информацию смотрите в разделе “Детали подгонки” на стр. 56.

Распределение указано вручную или отредактировано, и может включать параметр, заданный на нескольких уровнях.

- **Заблокировано.** Блокировка имитированного поля путем включения переключателя в столбце со значком замка исключает это поле из автоматического обновления связанным с ним узлом Подгонка имитации. Это наиболее полезно, если вы указали распределение вручную и хотите, чтобы на него не влияла автоматическая подгонка распределения при выполнении узла Подгонка имитации.
- **Распределение.** Ячейки в этом столбце содержат выпадающие списки статистических распределений. Выбор типа хранения определяет, какие распределения будут доступны в этом столбце для конкретного поля. Дополнительную информацию смотрите в разделе “Распределения” на стр. 59.

  Примечание: Нельзя указать Фиксированное распределение для всех полей. Чтобы все поля в сгенерированных данных были фиксированными, можно использовать узел Пользовательский ввод, за которым следует узел Баланс.
- **Параметры.** В этом столбце выводятся параметры распределения, связанные с каждым подгоняемым распределением. Несколько значений для одного параметра указываются через запятую. Указание нескольких значений для параметра генерирует несколько итераций для имитации. Дополнительную информацию смотрите в разделе “Итерации” на стр. 59. Если параметры отсутствуют, это отражается видом значка в столбце Состояние. Чтобы указать значения для параметров, щелкните мышью по этому столбцу в строке, соответствующей интересующему вас полю, и выберите в списке опцию **Задать**. Откроется диалоговое подокно Указать параметры. Дополнительную информацию смотрите в разделе “Указать параметры” на стр. 57. Если в столбце Распределение выбрана опция Эмпирическое, этот столбец недоступен.
- **Мин, Макс.** В этом столбце для некоторых распределений можно указать минимальное значение, максимальное значение или оба этих значения для имитированных данных. Имитированные данные меньше минимального значения или больше максимального значения будут отвергнуты, даже если они будут допустимы для заданного распределения. Чтобы указать минимальное и максимальное значения, щелкните мышью по этому столбцу в строке, соответствующей интересующему вас полю, и выберите в списке опцию **Задать**. Откроется диалоговое подокно Указать параметры. Дополнительную информацию смотрите в разделе “Указать параметры” на стр. 57. Если в столбце Распределение выбрана опция Эмпирическое, этот столбец недоступен.

Использовать наиболее точное соответствие. Доступно, только если узел Генерирование имитации автоматически создан из узла Подгона имитации с использованием хронологических данных, а в таблице Имитированные поля выбрана одна строка. Заменяет информацию для поля в выбранной строке на

информацию наиболее точно соответствующего распределения для этого поля. Если информация в выбранной строке была отредактирована, нажатие этой кнопки вернет эту информацию к наиболее точно соответствующему распределению, определенному из узла Подгонка имитации.

Детали подгонки. Доступно, только если узел Генерирование имитации автоматически создан из узла Подгона имитации. Открывает диалоговое подокно Детали подгонки. Дополнительную информацию смотрите в разделе “Детали подгонки” на стр. 56.

Некоторые нужные задачи можно выполнить при помощи значков в правой части представления Имитированные поля. Эти значки описаны в следующей таблице.

Таблица 11. Значки в представлении Имитированные поля.

Значок	Всплывающая подсказка	Описание
	**Редактировать параметры распределения**	Доступно, только если в таблице Имитированные поля выбрана одна строка. Открывает диалоговое подокно Указать параметры для выбранного поля. Дополнительную информацию смотрите в разделе “Указать параметры” на стр. 57.
	**Добавить новое поле**	Доступно, только если в таблице Имитированные поля выбрана одна строка. Добавляет новую пустую строку в конец таблицы Имитированные поля.
	**Создать несколько копий**	Доступно, только если в таблице Имитированные поля выбрана одна строка. Открывает диалоговое подокно Клонировать поле. Дополнительную информацию смотрите в разделе “Клонировать поле” на стр. 56.
	**Удалить выбранное поле**	Удаляет выбранную строку из таблицы Имитированные поля.
	**Переместить наверх**	Доступно, только если выбранная строка еще не является верхней строкой таблицы Имитированные поля. Перемещает выбранную строку в начало таблицы Имитированные поля. Это действие влияет на порядок полей в имитированных данных.
	**Переместить вверх**	Доступно, только если выбранная строка не является верхней строкой таблицы Имитированные поля. Перемещает выбранную строку на одну позицию вверх в таблице Имитированные поля. Это действие влияет на порядок полей в имитированных данных.

Таблица 11. Значки в представлении Имитированные поля (продолжение).

Значок	Всплывающая подсказка	Описание
	**Переместить вниз**	Доступно, только если выбранная строка не является последней строкой таблицы Имитированные поля. Перемещает выбранную строку на одну позицию вниз в таблице Имитированные поля. Это действие влияет на порядок полей в имитированных данных.
	**Переместить вниз**	Доступно, только если выбранная строка еще не является последней строкой таблицы Имитированные поля. Перемещает выбранную строку в конец таблицы Имитированные поля. Это действие влияет на порядок полей в имитированных данных.

Не очищать Мин и Макс при повторной подгонке. Когда выбрана эта опция, минимальное и максимальное значения не перезаписываются при обновлении распределений путем выполнения присоединенного узла Подгонка имитации.

Представление Корреляции

Часто заранее известно, что входные поля предиктивных моделей коррелируют - например, высота и ширина. Корреляции между полями, которые будут имитированы, необходимо учитывать, чтобы сохранить эти корреляции в имитированных значениях.

Если узел Генерирование имитации сгенерирован или изменен из узла Подгонка имитации с использованием хронологических данных, в представлении Корреляции можно просматривать и редактировать вычисленные корреляции между парами полей. Если у вас нет хронологических данных, можно задать корреляции вручную на основе ваших знаний о корреляции полей.

Примечание: До генерирования каких-либо данных выполняется автоматическая проверка, является ли матрица корреляций неотрицательно определенной и, таким образом, инвертируемой. Матрицу можно инвертировать, если ее столбцы линейно независимы. Если матрицу корреляций нельзя инвертировать, она будет автоматически подправлена так, чтобы сделать ее инвертируемой.

Можно выбрать вывод корреляции в формате матрицы или списка.

Матрица корреляций. Выводит корреляции между парами полей в виде матрицы. Имена полей выводятся в алфавитном порядке по вертикали с левой стороны матрицы и по горизонтали наверху матрицы. Можно редактировать только ячейки ниже диагонали; допустим ввод значений от -1,000 до 1,000 включительно. Ячейка выше диагонали обновляется, когда фокус ввода перемещается с ячейки, являющейся ее отражением выше диагонали; после этого в обеих ячейках выводится одно и то же значение. Ячейки на диагонали всегда недоступны и всегда содержат корреляцию 1,000. Значение по умолчанию для всех остальных ячеек - 0,000. Значение 0,000 указывает, что между двумя соответствующими полями нет никакой корреляции. В матрицу включены только непрерывные и порядковые поля. Номинальные, категорические и флаговые поля, а также поля, которым назначено Фиксированное распределение, не выводятся в таблице.

Список Корреляции. Выводит корреляции между парами полей в виде таблицы. Каждая строка таблицы показывает корреляцию между парой полей. Строки нельзя добавлять или удалять. Столбцы с заголовками Поле 1 и Поле 2 содержат имена полей, и их нельзя отредактировать. Столбец Корреляция содержит корреляции, и их можно редактировать; допустим ввод значений от -1,000 до 1,000 включительно. Значение

по умолчанию для всех ячеек - 0,000. В список включены только непрерывные и порядковые поля. Номинальные, категорические и флаговые поля, а также поля, которым назначено Фиксированное распределение, не выводятся в списке.

Сброс корреляций. Открывает диалоговое окно Сброс корреляций. Если доступны хронологические данные, можно выбрать одну из трех опций:

- **Подогнанные.** Заменяет текущие корреляции на корреляции, вычисленные с использованием хронологических данных.
- **Нули.** Заменяет текущие корреляции на нули.
- **Отмена.** Закрывает диалоговое окно. Корреляции остаются неизмененными.

Если исторические данные недоступны, но вы изменили корреляции, можно выбрать замену текущих корреляций на нули или отменить операцию.

Показать как. Выберите **Таблица**, чтобы показать корреляции в виде матрицы. Выберите **Список**, чтобы показать корреляции в виде списка.

Не пересчитывать корреляции при повторной подгонке. Выберите этот параметр, если необходимо вручную указать корреляции и не допустить их перезаписи при автоматической подгонке распределений с использованием узла Подгонка имитации и хронологических данных.

Использовать подогнанную многостороннюю таблицу сопряженности для вводов с категориальным распределением. По умолчанию все поля с категорическим распределением включаются в таблицу сопряженности (или в многостороннюю таблицу сопряженности, в зависимости от числа полей с категорическим распределением). Таблица сопряженности строится, как и корреляции, при выполнении узла Подгонка имитации. Таблицу сопряженности невозможно просмотреть. Когда выбрана эта опция, поля с категорическим распределением имитируются с использованием фактических процентов из таблицы сопряженности. То есть любые ассоциации между номинальными полями создаются заново в новых имитированных данных. Когда эта опция выключена, поля с категорическими распределениями имитируются с использованием ожидаемых процентов из таблицы сопряженности. Если изменить поле, это поле удаляется из таблицы сопряженности.

Представление Дополнительные опции

Число наблюдений для имитации. Выводит опции для указания числа наблюдений для имитации и названий итераций.

- **Максимальное число наблюдений.** Указывает максимальное количество наблюдений имитированных данных, а также связанных целевых значений для создания. Значение по умолчанию - 10000, минимальнон значение - 1000, максимальное значение - 2147483647.
- **Итерации.** Это число вычисляется автоматически, и его нельзя редактировать. Каждый раз, когда для параметра распределения задано несколько значений, создается новая итерация.
- **Всего строк.** Включено, только когда число итераций больше 1. Это число вычисляется автоматически по показанной формуле, и его нельзя отредактировать.
- **Создать поле итерации.** Включено, только когда число итераций больше 1. Когда выбрана эта опция, становится доступным поле **Имя**. Дополнительную информацию смотрите в разделе “Итерации” на стр. 59.
- **Имя.** Доступно, только когда включен переключатель **Создать поле итерации** и число итераций больше 1. Отредактируйте имя поля итерации, введя значение в это текстовое поле. Дополнительную информацию смотрите в разделе “Итерации” на стр. 59.

Стартовое число генератора псевдослучайных чисел. Задание стартового числа генератора псевдослучайных чисел позволяет воспроизвести имитацию.

- **Воспроизвести результаты.** Когда выбрана эта опция, становятся доступны кнопка **Генерировать** и поле **Стартовое число генератора псевдослучайных чисел**.

- **Стартовое число генератора псевдослучайных чисел.** Опция доступна, только когда включен переключатель **Воспроизвести результаты**. В этом поле можно указать число, используемое в качестве стартового числа генератора псевдослучайных чисел. Значение по умолчанию - 629111597.
- **Генерировать.** Опция доступна, только когда включен переключатель **Воспроизвести результаты**. Создает псевдослучайное целое число от 1 до 999999999 включительно в поле **Стартовое число генератора псевдослучайных чисел**.

Клонировать поле

В диалоговом окне Клонировать поле можно указать, сколько копий выбранного поля нужно создать, а также задать имя для каждой копии. Несколько копий полей полезно иметь при исследовании составных эффектов, например, процентной ставки или темпов роста за несколько последовательных периодов времени.

В строке заголовка диалогового окна выводится имя выбранного поля.

Число создаваемых копий. Содержит число копий поля, которые нужно создать. Щелкайте мышью по стрелкам, чтобы выбрать число копий, которые нужно создать. Минимальное число копий - 1, а максимальное - 512. Начальное значение числа копий - 10.

Символы суффикса копий. Содержит символы, добавляемые в конец имени поля для каждой копии. Эти символы отделяют имя поля от номера копии. Чтобы отредактировать символы суффикса, введите новые символы в этом поле. Это поле можно оставить пустым; тогда между именем поля и номером копии не будет никаких символов. Символ по умолчанию - подчеркивание.

Начальный номер копии. Содержит номер суффикса для первой копии. Чтобы выбрать начальный номер копии, щелкайте мышью по кнопкам со стрелками. Минимальный начальный номер копии - 1, а максимальный - 1000. Начальный номер копии по умолчанию - 1.

Шаг номеров копий. Содержит инкремент для номеров суффиксов. Чтобы выбрать инкремент, щелкайте мышью по кнопкам со стрелками. Минимальный инкремент - 1, а максимальный - 255. Начальное значение инкремента - 1.

Поля. Показывает предварительный просмотр имен полей для копий, который обновляется, если отредактировать какое-нибудь из полей в диалоговом окне Клонировать поле. Этот текст генерируется автоматически и его невозможно отредактировать.

OK. Генерирует все копии в соответствии с параметрами, заданными в диалоговом окне. Копии добавляются в таблицу Имитированные поля в диалоговом окне узла Имитация Генерировать непосредственно под строкой, содержащей скопированное поле.

Отмена. Закрывает диалоговое окно. Любые сделанные изменения отменяются.

Детали подгонки

Диалоговое окно Детали подгонки доступно, только если узел Имитация Генерировать был создан или изменен путем выполнения узла Подгонка имитации. В нем выводятся результаты автоматической подгонки распределения для выбранного поля. Распределения упорядочиваются по степени согласия; наиболее точно соответствующее распределение указывается первым. В этом диалоговом окне можно выполнить следующие задачи:

- Изучить распределения, подогнанные к хронологическим данным.
- Выбрать одно из подогнанных распределений.

Поле. Содержит имя выбранного поля. Этот текст нельзя изменить.

Рассматривать как (мера). Показывает тип измерения выбранного поля. Это значение берется из таблицы Имитированные поля в диалоговом окне узла Имитация Генерировать. Тип измерения можно изменить, щелкнув по стрелке и выбрав нужный тип измерения в выпадающем списке. Есть три опции: **Непрерывное**, **Номинальное** и **Порядковое**.

Распределения. В таблице Распределения выводятся все распределения, подходящие для данного типа измерения. Распределения, подогнанные к хронологическим данным, упорядочены по степени согласия, от наилучшего согласия до наихудшего. Согласие определяется статистикой согласия, выбранной в узле Подгонка имитации. Распределения, не подогнанные к хронологическим данным, выводятся в таблице в алфавитном порядке ниже подогнанных распределений.

Таблица Распределение содержит следующие столбцы:

- **Использование.** Выбранная радиокнопка указывает, какое распределение сейчас выбрано для поля. Вы можете заменить наиболее точно соответствующее распределение, выбрав радиокнопку нужного распределения в столбце Использование. Кроме того, выбор радиокнопки в столбце Использовании выводит график распределения поверх гистограммы (или столбчатой диаграммы) хронологических данных для выбранного поля. Одновременно можно выбрать только одно распределение.
- **Распределение.** Содержит имя распределения. Содержимое этого столбца нельзя отредактировать.
- **Статистика подгонки.** Содержит вычисленную статистику согласия для данного распределения. Содержимое этого столбца нельзя отредактировать. Содержимое ячейки зависит от типа измерения поля:
  - **Непрерывные.** Содержит результаты критерия Андерсона-Дарлинга или Колмогорова-Смирнова. Также показаны p-значения, связанные с тестами. Статистика согласия, выбранная в качестве критериев согласия в узле Подгонка имитации, выводится первой и используется для упорядочивания распределений. Статистики Андерсона-Дарлинга выводятся в формате: A=*a_значение* P=*p_значение*. Статистики Колмогорова-Смирнова выводятся в формате: K=*k_значение* P=*p_значение*. Если статистику невозможно вычислить, вместо числа выводится точка.
  - **Номинальное и Порядковое.** Содержит результаты теста хи-квадрат. Кроме того, выводится p-значение, связанное с этим тестом. Статистики выводятся в формате: Chi-Sq=*значение* P=*p_значение*. Если распределение не подогнано, выводится надпись `Не подогнано`. Если распределение невозможно математически подогнать, выводится надпись `Невозможно подогнать`.

  Примечание: Для Эмпирического распределения эта ячейка всегда пуста.
- **Анализ важности независимых переменных.** Содержит параметры распределения, связанные с каждым подгоняемым распределением. Эти параметры выводятся в формате *имя_параметра* = *значение_параметра*, и отдельные параметры выводятся через пробел. Для категориального распределения имена параметров являются категориями, а значения параметров являются связанными с ними вероятностями. Если распределение не подогнано к хронологическим данным, эта ячейка пуста. Содержимое этого столбца нельзя отредактировать.

Миниизображение гистограммы. Содержит график выбранного распределения поверх гистограммы хронологических данных выбранного поля.

Миниизображение распределения. Содержит объяснение и иллюстрацию к выбранному распределению.

OK. Закрывает диалоговое окно и обновляет значения в столбцах Измерение, Распределение, Параметры и Мин, Макс таблицы Имитированные поля для выбранного поля, используя информацию из выбранного распределения. Кроме того, значок в столбце Статус изменяется, отражая, является ли выбранное распределение распределением, наиболее точно соответствующим данным.

Отмена. Закрывает диалоговое окно. Любые сделанные изменения отменяются.

Указать параметры

В диалоговом окне Указать параметры можно вручную указать значения параметров для распределения выбранного поля. Кроме того, можно выбрать другое распределение для выбранного поля.

Диалоговое окно Указать параметры можно открыть тремя способами:

- Дважды щелкните по имени поля в таблице Имитированные поля в диалоговом окне узла Генерирование имитации.
- Щелкните по столбцу Параметры или Мин, Макс в таблице Имитированные поля и выберите из списка **Указать**.
- В таблице Имитированные поля выберите строку, а затем щелкните по значку **Редактировать параметры распределения**.

Поле. Содержит имя выбранного поля. Этот текст нельзя изменить.

Распределение. Содержит распределение выбранного поля. Это значение берется из таблицы Имитированные поля. Распределение можно изменить, щелкнув по стрелке и выбрав нужное распределение в выпадающем списке. Доступные распределения зависят от типа хранения выбранного поля.

Грани. Эта опция доступна, только когда в поле **Распределение** выбрано распределение игральной кости. Щелкайте мышью по стрелкам, чтобы выбрать число граней, или категорий, на которые нужно разделить поле. Минимальное число граней - 2, а максимальное - 20. Начальное значение числа граней - 6.

Параметры распределения. Таблица Параметры распределения содержит по одной строке для каждого параметра выбранного распределения.

Примечание: Распределение использует параметр интенсивности с параметром формы $\alpha = k$ и с обратным параметром масштаба $\beta = 1/\theta$.
Таблица содержит два столбца:

- **Параметр.** Содержит имена параметров. Содержимое этого столбца нельзя отредактировать.
- **Значения.** Содержит значения параметров. Если узел Генерирование имитации создан или изменен из узла Подгонка имитации, клетки в этом столбце содержат значения параметров, определенные путем подгонки распределения под хронологические данные. Если узел Генерирование имитации добавлен на холст потока из палитры Узлы источников, клетки в этом столбце будут пустыми. Значения можно редактировать, вводя их в ячейках. Дополнительную информацию о параметрах, требуемых каждым распределением, и о допустимых значениях параметров смотрите в разделе “Распределения” на стр. 59.

  Несколько значений для одного параметра надо указывать через запятую. Указание нескольких значений для параметра определяет несколько итераций имитации. Указать несколько значений можно только для одного параметра.

  Примечание: Для полей с типами хранения даты/времени следует указывать параметры распределения в виде целых чисел. Например, чтобы задать в качестве средней даты 1 января 1970 года, укажите целое число 0.

Примечание: Когда выбрано распределение броском игральной кости, таблица Параметры распределения выглядит немного иначе. Таблица содержит по одной строке для каждой грани (или категории). В таблице есть столбец Значение и столбец Вероятность. Столбец Значение содержит метку для каждой категории. Значения по умолчанию для меток - целые числа от 1 до N, где N - число граней. Метки можно редактировать, вводя значения в ячейках. В клетках можно вводить любые значения. Если вы хотите использовать значение, не являющееся числом, тип хранения поля данных должен быть изменен на строковый, если он не является строковым. Столбец Вероятность содержит вероятность для каждой категории. Вероятности нельзя изменить, они вычисляются как 1/N.

Просмотр. Показывает пример графика распределения на основе указанных параметров. Если для одного параметра задано несколько значений, выводятся примеры графиков для каждого значения параметра. Если для выбранного поля доступны хронологические данные, график распределения накладывается на гистограмму хронологических данных.

Необязательные параметры. Используйте эти опции для указания минимального значения, максимального значения или обоих этих значений для имитированных данных. Имитированные данные меньше минимального значения или больше максимального значения будут отвергнуты, даже если они будут допустимы для заданного распределения.

- **Задайте минимум.** Выберите эту опцию, чтобы сделать доступным поле **Отвергнуть значения ниже**. Этот переключатель недоступен, если выбрано Эмпирическое распределение.
- **Отвергнуть значения ниже.** Доступно, только если включен переключатель **Задайте минимум**. Введите минимальное значение для имитированных данных. Любые имитированные значения меньше этого значения будут отвергнуты.
- **Задайте максимум.** Выберите эту опцию, чтобы сделать доступным поле **Отвергнуть значения выше**. Этот переключатель недоступен, если выбрано Эмпирическое распределение.
- **Отвергнуть значения выше.** Доступно, только если включен переключатель **Задайте максимум**. Введите максимальное значение для имитированных данных. Любые имитированные значения больше этого значения будут отвергнуты.

OK. Закрывает диалоговое окно и обновляет значения в столбцах Распределение, Параметры и Мин, Макс таблицы Имитированные поля для выбранного поля. Кроме того, значок в столбце Статус обновляется в соответствии с выбранным распределением.

Отмена. Закрывает диалоговое окно. Любые сделанные изменения отменяются.

Итерации

Если вы указали несколько значений для фиксированного поля или для параметра распределения, для каждого указанного значения создается независимый набор имитированных наблюдений - а следовательно, и отдельная имитация. Это позволяет исследовать эффект от варьирования значения поля или параметра. Каждый набор имитированных наблюдений называется *итерацией*. В имитированных данных итерации изображаются в виде стопки.

Если включен переключатель **Создать поле итерации** в представлении Дополнительные опции диалогового окна узла Генерирование имитации, в имитированные данные добавляется поле итерации в виде номинального поля с числовым хранением. Имя этого поля можно отредактировать, введя значение в поле **Имя** в представлении Дополнительные опции. Это поле содержит метку, которая указывает, к какой итерации относится каждое имитированное наблюдение. Форма меток зависит от типа итерации:

- **Итерация по фиксированному полю.** Метка состоит из имени поля, знака равенства и значения поля для этой итерации, то есть:

  имя_поля = *значение_поля*
- **Итерация по параметру распределения.** Метка состоит из имени поля, двоеточия, имени параметра, по которому выполняется итерация, знака равенства и значения поля для этой итерации, то есть:

  имя_поля:*имя_параметра* = *значение_параметра*
- **Итерация по параметру распределения для категорического распределения или распределения диапазона.** Метка состоит из имени поля, слова "Iteration" и номера итерации, то есть:

  имя_поля: Iteration *номер_итерации*

Распределения

Чтобы вручную указать распределение вероятностей для любого поля, можно открыть диалоговое окно Указать параметры, выбрать нужное распределение из списка **Распределение** и ввести параметры распределения в таблице **Параметры распределения**. Далее изложены некоторые примечания по некоторым распределениям.

- **Категориальное.** Категориальные распределения описывают входное поле с фиксированным количеством числовых значений, называемых категориями. Каждая категория имеет связанную с ней вероятность. Сумма вероятностей всех категорий равняется единице.

Примечание: Если задать вероятности для категорий, сумма которых не равна 1, вы получите предупреждение.

- **Негативное биноминальное - ошибки.** Описывает распределение количества ошибок в последовательности испытаний перед обзором количества успешных исходов. Параметр *Пороговое значение* - заданное число успешных исходов, а параметр *Вероятность* - вероятность успешного исхода в каждом отдельном испытании.
- **Негативное биноминальное - испытания.** Описывает распределение количества испытаний, требуемых перед обзором количества успешных исходов. Параметр *Пороговое значение* - заданное число успешных исходов, а параметр *Вероятность* - вероятность успешного исхода в каждом отдельном испытании.
- **Диапазон.** Это распределение состоит из набора интервалов с вероятностью, назначенной каждому интервалу. Сумма вероятностей всех интервалов равна 1. Значения с заданным интервалом извлекаются из равномерного распределения, определенного на этом интервале. Интервалы указываются вводом минимального значения, максимального значения и связанной с ними вероятности.

  Например, допустим, что вы считаете, что стоимость сырого материала с вероятностью 40% будет в диапазоне от $10 до $15 за единицу, и с вероятностью 60% - в диапазоне от $15 до $20 за единицу. Вы смоделируете стоимость при помощи распределения Диапазон, которое состоит из двух интервалов - [10 - 15] и [15 - 20]. Для первого интервала вероятность составляет 0,4, для второго - 0,6. Интервалы не обязательно должны быть количественными; они могут даже пересекаться. Например, можно указать интервалы $10 - $15 и $20 - $25 или $10 - $15 и $13 - $16.
- **Распределение Вейбулла.** Параметр *Положение* - необязательный параметр положения, указывающий, где находится начальная точка распределения.

В следующей таблице показаны распределения, доступные для пользовательской подгонки распределения, и допустимые значения параметров. Некоторые из этих распределений доступны для пользовательской подгонки к отдельным типам хранения, даже если они не подгоняются к этим типам хранения автоматически узлом Подгонка имитации.

Таблица 12. Распределения, доступные для пользовательской подгонки

Распределение	**Тип хранения, поддерживаемый для пользовательской подгонки**	**Анализ важности независимых переменных**	**Предельные значения параметров**	**Примечания**
Бернулли	Целое число, действительное число, дата/время	Вероятность	0 ≤ *Вероятность* ≤ 1	
Бета	Целое число, действительное число, дата/время	Форма 1 Форма 2 Минимум Максимум	≥ 0 ≥ 0 < *Максимум* > *Минимум*	Минимум и максимум необязательны.
Биномиальное	Целое число, действительное число, дата/время	Число испытаний (n) Вероятность Минимум Максимум	> 0, целое 0 ≤ *Вероятность* ≤ 1 < *Максимум* > *Минимум*	Число испытаний должно быть целым. Минимум и максимум необязательны.
Категориальное	Целое число, действительное число, дата/время, строка	Имя (или метка) категории	0 ≤ *Значение* ≤ 1	Значение является вероятностью категории Сумма значений должна быть равна 1, иначе выводится предупреждение.

Таблица 12. Распределения, доступные для пользовательской подгонки (продолжение)

Распределение	**Тип хранения, поддерживаемый для пользовательской подгонки**	**Анализ важности независимых переменных**	**Предельные значения параметров**	**Примечания**
Игральной кости	Целое число, строка	Грани	2 ≤ *Грани* ≤ 20	Вероятность каждой категории (грани) вычисляется как 1/N, где N - число граней. Вероятности нельзя отредактировать.
Эмпирическое	Целое число, действительное число, дата/время			Эмпирическое распределение невозможно отредактировать или выбрать в качестве типа. Эмпирическое распределение доступно только при наличии хронологических данных.
Экспоненциальное	Целое число, действительное число, дата/время	Масштаб Минимум Максимум	> 0 < *Максимум* > *Минимум*	Минимум и максимум необязательны.
Фиксированное	Целое число, действительное число, дата/время, строка	Значение		Нельзя указать Фиксированное распределение для всех полей. Чтобы все поля в сгенерированных данных были фиксированными, можно использовать узел Пользовательский ввод, за которым следует узел Баланс.
Гамма	Целое число, действительное число, дата/время	Форма Масштаб Минимум Максимум	≥ 0 ≥ 0 < *Максимум* > *Минимум*	Минимум и максимум необязательны. Распределение использует параметр интенсивности с параметром формы α = k и с обратным параметром масштаба β = 1/θ.
Логнормальное	Целое число, действительное число, дата/время	Форма 1 Форма 2 Минимум Максимум	≥ 0 ≥ 0 < *Максимум* > *Минимум*	Минимум и максимум необязательны.

Таблица 12. Распределения, доступные для пользовательской подгонки (продолжение)

Распределение	Тип хранения, поддерживаемый для пользовательской подгонки	Анализ важности независимых переменных	Предельные значения параметров	Примечания
Негативное биноминальное - ошибки	Целое число, действительное число, дата/время	Пороговое значение Вероятность Минимум Максимум	≥ 0 0 ≤ *Вероятность* ≤ 1 < *Максимум* > *Минимум*	Минимум и максимум необязательны.
Негативное биноминальное - испытания	Целое число, действительное число, дата/время	Пороговое значение Вероятность Минимум Максимум	≥ 0 0 ≤ *Вероятность* ≤ 1 < *Максимум* > *Минимум*	Минимум и максимум необязательны.
Нормальный	Целое число, действительное число, дата/время	Среднее значение Среднеквадратичное отклонение Минимум Максимум	≥ 0 ≥ 0 < *Максимум* > *Минимум*	Минимум и максимум необязательны.
Пуассона	Целое число, действительное число, дата/время	Среднее значение Минимум Максимум	≥ 0 < *Максимум* > *Минимум*	Минимум и максимум необязательны.
Диапазон	Целое число, действительное число, дата/время	Begin(X) End(X) Probability(X)	0 ≤ *Значение* ≤ 1	X - индекс каждого контейнера. Сумма значений вероятности должна быть равна 1.
Треугольник	Целое число, действительное число, дата/время	Режим Минимум Максимум	*Минимум* ≤ *Значение* ≤ *Максимум* < *Максимум* > *Минимум*	
Равные	Целое число, действительное число, дата/время	Минимум Максимум	< *Максимум* > *Минимум*	
Вейбулла	Целое число, действительное число, дата/время	Курс Масштаб Положение Минимум Максимум	> 0 > 0 ≥ 0 < *Максимум* > *Минимум*	Положение, минимум и максимум необязательны.

Узел Представление данных

Узел представления данных используется для включения в поток данных, определяемых в представлении аналитических данных IBM SPSS Collaboration and Deployment Services. Представление аналитических данных определяет структуру для обращения к данным, которая описывает объекты, используемые в предсказательных моделях и бизнес-правилах. Это представление связывает структуру данных с физическими источниками данных для анализа.

Предсказательная аналитика требует, чтобы данные были организованы в таблицах, где каждая строка соответствует объекту, для которого делаются предсказания. Каждый столбец в таблице представляет измеримый атрибут объекта. Некоторые атрибуты можно получить операцией агрегирования по значениям для другого атрибута. Например, строки таблицы могут представлять покупателей со столбцами, соответствующими имени покупателя, его полу, почтовому индексу и количеству покупок этого покупателя в прошлом году, превышающих 500 долларов. Последний столбец получается из хронологии заказов покупателя, обычно хранящейся в одной или нескольких связанных таблицах.

Процесс предсказательной аналитики подразумевает использование разнообразных наборов данных на всем протяжении жизненного цикла модели. На начальном этапе разработки предсказательной модели применяются хронологические данные, результаты которых часто известны для предсказываемого события. С целью оценки эффективности и точности модели проверяется модель-кандидат на различных данных. После проверки модели вы внедряете ее в промышленное использование, чтобы сгенерировать оценки для множества объектов в процессе пакетной обработки или для одиночных объектов в процессе реального времени. В случае объединения модели с бизнес-правилами в процессе управления решениями вы при помощи имитированных данных проверяете результаты этого объединения. Однако хотя используемые данные отличаются на разных этапах процесса разработки модели, каждый набор данных должен предоставлять для модели один и тот же набор атрибутов. Набор атрибутов остается постоянным; записи данных подлежат анализируемому изменению.

В состав представления аналитических данных входят следующие компоненты, разрешающие специализированные потребности предсказательной аналитики:

- Схема представления данных или модель данных, определяющая логический интерфейс для обращения к данным как к набору атрибутов, организованных в связанные таблицы. Атрибуты в этой модели могут быть получены из других атрибутов.
- Один или несколько планов доступа к данным, предоставляющих атрибуты модели данных с физическими значениями. Вы управляете данными, доступными для модели данных, указывая, какой план доступа к данным будет активен для конкретной прикладной программы.

Важное замечание: Для использования узла представления данных сначала нужно установить и сконфигурировать IBM SPSS Collaboration and Deployment Services Repository на вашем сайте. Представление аналитических данных, на которое ссылается этот узел, обычно создается и сохранятся в репозитории при помощи IBM SPSS Collaboration and Deployment Services Deployment Manager.

Задание опций для узла представления данных

При помощи опций на вкладке **Данные** диалогового окна узла Представление данных задаются параметры данных для представления аналитических данных, выбираемого в IBM SPSS Collaboration and Deployment Services Repository.

Представление аналитических данных. Нажмите кнопку с многоточием (**...**), чтобы выбрать представление аналитических данных. Если соединение с сервером репозитория в текущий момент отсутствует, задайте URL для сервера в диалоговом окне Репозиторий: Сервер, нажмите кнопку **OK** и задайте ваши параметры аутентификации для соединения в диалоговом окне Репозиторий: Авторизация. Дополнительную информацию о входе в систему и получении объектов смотрите в руководстве пользователя IBM SPSS Modeler.

Имя таблицы. В представлении аналитических данных выберите таблицу в модели данных. Каждая таблица в этой модели данных представляет понятие или объект, участвующий в процессе предсказательной аналитики. Поля для таблиц соответствуют атрибутам представляемых этими таблицами объектов. Например, если вы анализируете заказы заказчиков, ваша модель данных может содержать таблицу для заказчиков и таблицу для заказов. У таблицы заказчиков могут быть поля для идентификатора заказчика, возраста заказчика, его пола, семейного положения и страны местожительства. У таблицы заказов могут быть поля для идентификатора заказа, числа позиций в заказе, его общей стоимости и идентификатора заказчика, разместившего заказ. Поле идентификатора заказчика может использоваться для связывания заказчиков в таблице заказчиков с их заказами в таблице заказов.

План доступа к данным. Выберите в представлении аналитических данных план доступа к данным. План доступа к данным связывает таблицы модели в представлении аналитических данных с физическими источниками данных. Представление аналитических данных обычно содержит несколько планов доступа к данным. Изменяя находящийся в использовании план доступа к данным, вы изменяете данные, используемые потоком. Например, если представление аналитических данных содержит план доступа к данным для обучения модели и план доступа к данным для ее испытания, изменив используемый план доступа к данным, можно переключиться с обучающих данных на проверочные данные.

Необязательные атрибуты. Если прикладной программе, использующей представление аналитических данных, какой-либо отдельный атрибут не требуется, его можно пометить как необязательный. Необязательные атрибуты, в отличие от обязательных, могут содержать пустые значения. Может потребоваться скорректировать прикладную программу, включив в нее обработку пустых значений для необязательных атрибутов. Например, при вызове бизнес-правила, созданного в IBM Operational Decision Manager, IBM Analytical Decision Management запрашивает службу правил, чтобы определить, какие входные поля требуются. Если подлежащая оценке запись содержит пустое значение для каких-либо требуемых службой правил полей пустое значение, правило не вызывается, а выходные поля правила заполняются значениями по умолчанию. Если пустое значение содержится в необязательном поле, правило вызывается. Правило может выполнять проверку на пустые значения для управления обработкой.

Чтобы указать атрибуты как необязательные, нажмите кнопку **Необязательные атрибуты** и выберите атрибуты, которые являются необязательными.

Включить в поле данные XML. Выберите эту опцию, чтобы создать поле, содержащее данные XML модели объекта выполняемого кода для каждой строки данных. Эта информация требуется, если данные будут использоваться с IBM Operational Decision Manager. Задайте для этого нового поля имя.

Геопространственный узел источника

Узел источника Геопространственный позволяет перенести карту или геопространственные данные в сеанс исследования данных. Импортировать данные можно одним из двух способов:

- В файл формы (`.shp`)
- Соединившись с сервером ESRI, содержащим иерархическую файловую систему, в том числе файлы карты.

  Примечание: Соединяться можно только с общедоступными службами карт.

Модели пространственно-временных предсказаний (Spatio-Temporal Prediction, STP) может включать карты или пространственные элементы в свои предсказания. Дополнительную информацию об этих моделях смотрите в теме "Узел моделирования пространственно-временных предсказаний" раздела Модели временных рядов в руководстве по узам моделирования Modeler (ModelerModelingNodes.pdf).

Задание опций для узла геопространственного источника

Тип источника данных Можно или импортировать данные из **Файла формы** (`.shp`), или соединиться со **Службой карт**.

Если вы используете **Файл формы**, или введите имя файла и путь к нему, или используйте обзор для выбора файла. Этот файл должен или находиться в локальном каталоге, или быть доступен с отображенного диска; невозможно получить доступ к этому файлу с использованием пути универсального соглашения об именовании (universal naming convention, UNC).

Примечание: Для данных формы нужны оба файла, `.shp` и `.dbf`. У этих двух файлов должны быть одинаковые имена, и они должны быть в одной папке. Файл `.dbf` автоматически импортируется при выборе файла `.shp`. Кроме того, может существовать файл `.prj`, который задает систему координат для данных форм.

Если вы используете **Службу карт**, введите URL для сервера и нажмите кнопку **Соединиться**. После соединения со службой слои из этой службы выводятся внизу диалогового окна в виде структуры дерева на панели **Доступные карты**; раскройте это дерево и выберите нужный вам слой.

Примечание: Соединяться можно только с общедоступными службами карт.

Автоматическое определение геопространственных данных

По умолчанию SPSS Modeler автоматически определяет, когда возможно, все геопространственные поля на узле источника с правильными метаданными. Метаданные могут включать в себя шкалу измерений геопространственного поля (такую как Точка или Многоугольник) и систему координат, используемую этими полями, в том числе такие сведения как начало отсчета (например, широта 0, долгота 0) и единицы измерения. Дополнительную информацию об уровнях измерения смотрите в разделе “Подуровни геопространственных измерений” на стр. 131.

Файлы `.shp` и `.dbf`, составляющие файл формы, содержат общее поле идентификатора, используемое как ключ. Например, файл `.shp` может содержать страны и использовать поле названия страны в качестве идентификатора, а файл `.dbf` может содержать сведения об этих странах с названием страны, также используемым как идентификатор.

Примечание: Если система координат не совпадает с системой координат SPSS Modeler по умолчанию, может потребоваться перепроектировать данные для использования нужной системы координат. Дополнительную информацию смотрите в разделе “Узел перепроектирования” на стр. 176.

Общие вкладки узлов источников

Для всех узлов источников, щелкнув по соответствующей вкладке, можно задать следующие опции:

- **Вкладка Данные.** Используется для изменения типа хранения по умолчанию.
- **Вкладка Фильтр.** Используется для устранения или переименования полей данных. Эта вкладка предлагает те же функции, что узел фильтра. Дополнительную информацию смотрите в разделе “Задание опций фильтрации” на стр. 142.
- **Вкладка =Типы.** Используется для задания шкал измерений. Эта вкладка предлагает те же функции, что узел типа.
- **Вкладка Аннотации.** Используемая для всех узлов эта вкладка предлагает опции для их переименования, задания пользовательских подсказок и хранения очень длинных аннотаций.

Задание шкал измерений на узле источника

Свойства полей можно задать на узле источника или отдельном узле типа. Обработка аналогична на обоих узлах. Доступны следующие свойства:

- **Поле** Щелкните дважды кнопкой мыши по имени поля, чтобы задать метки значений и полей для данных в IBM SPSS Modeler. Например, здесь можно просмотреть поля метаданных, импортированные из IBM SPSS Statistics. Таким же образом можно создать новые метки для полей и их значений. Задаваемые здесь метки выводятся повсеместно в IBM SPSS Modeler в зависимости от вариантов, выбранных вами в диалоговом окне Свойства потока.
- **Измерение** Эта шкала измерений используется для описания характеристик данных в указанном поле. Если все подробности поля известны, оно называется **полностью инстанциированным**. Дополнительную информацию смотрите в разделе “Уровни измерения” на стр. 129.

  Примечание: Шкала измерений поля отличается от типа хранения, который указывает, хранятся ли данные как строки, целые, действительные числа, даты, время, отметки времени или списки.
- **Значения** Этот столбец позволяет задать опции для чтения значений данных из набора данных; можно также использовать опцию **Задать**, чтобы задать шкалы и значения измерений в отдельном диалоговом окне. Можно также выбрать вариант передачи полей полей без чтения их значений. Дополнительную информацию смотрите в разделе “Значения данных” на стр. 133.

  Примечание: Нельзя изменить ячейку в этом столбце, если соответствующая запись **Поле** содержит список.

- **Отсутствующие** Используется для задания способа, которым будут обрабатываться пропущенные значения для поля. Дополнительную информацию смотрите в разделе “Определение пропущенных значений” на стр. 138.

  Примечание: Нельзя изменить ячейку в этом столбце, если соответствующая запись **Поле** содержит список.
- **Проверка** В этом столбце можно задать опции, гарантирующие, что значения полей будут соответствовать заданным значениям или диапазонам. Дополнительную информацию смотрите в разделе “Проверка значений типа” на стр. 138.

  Примечание: Нельзя изменить ячейку в этом столбце, если соответствующая запись **Поле** содержит список.
- **Роль** Используется для указания узлам моделирования, будут ли поля представлять собой **Входные поля** (поля предикторов) or **Поля назначения** (предсказанные поля) для процесса машинного обучения. Доступны также роли **Двойного назначения** и **Нет**, наряду с ролью **Разделение**, указывающей поле, используемое для разделения записей на отдельные выборки для обучения, испытания и проверки. Значение **Разбиение** задает, что будет выполняться построение отдельных моделей для каждого возможного значения поля. Дополнительную информацию смотрите в разделе “Задание роли поля” на стр. 139.

Дополнительную информацию смотрите в разделе “Узел Тип” на стр. 128.

Когда выполнять инстанциацию на узле источника

Существует два способа получения информации о хранении данных и значениях ваших полей. Эта **инстанциация** может происходить на любом из узлов источников при первом переносе данных в IBM SPSS Modeler или посредством вставки узла типа в поток данных.

Инстанциация на узле источника полезна, если:

- Набор данных невелик.
- Вы собираетесь получать новые поля при помощи Построителя выражений (инстанциация делает значения полей доступными из Построителя выражений).

В общем случае, если набор данных не слишком велик и вы не собираетесь добавлять поля в поток позднее, инстанциация на узле источника будет более удобным способом.

Фильтрация полей с узла источника

Вкладка Фильтр в диалоговом окне узла источника позволяет исключить поля из операций нисходящего потока на основе начального следования данных. Это полезно, например, если в данных есть дубликаты полей или если вы уже достаточно хорошо знаете данные, чтобы исключить ненужные поля. Можно также добавить в поток отдельный узел фильтра позднее. Обработка одинакова в обоих случаях. Дополнительную информацию смотрите в разделе “Задание опций фильтрации” на стр. 142.

Глава 3. Узлы операций с записями

Обзор операций с записями

Узлы операций с записями используются для внесения изменений в данные на уровне записей. Эти операции важны на фазах **изучения данных** и **подготовки данных** процесса исследования данных, поскольку позволяют адаптировать данные к конкретным потребностям бизнеса.

Например, на основе результатов аудита данных, проведенного при помощи узла аудита данных (палитра Вывод) можно решить, что следовало бы слить записи приобретений заказчиков за прошедшие три месяца. При помощи узла слияния можно выполнять слияние записей на основе значений поля ключа, такого как *ID заказчика*. Возможно также, вы обнаружите, что база данных, содержащая информацию о посещениях Web-сайтов, неуправляема по отношению к более чем миллиону записей. При помощи узла выборки можно выбрать поднабор данных для использования при моделировании.

Палитра Операции с записями содержит следующие узлы:

Узел Выбор отбирает или отбрасывает подмножество записей из потока данных на основе конкретного условия. Например, вы можете выбрать записи, принадлежащие определенному району продаж.

Узел Выборка отбирает подмножество записей. Поддерживается несколько типов выборки, в том числе стратифицированные, кластеризованные и неслучайные (структурированные) выборки. Выборки могут быть полезны для повышения производительности и для выбора групп связанных записей или транзакций для анализа.

Узел Баланс исправляет дисбаланс в наборе данных, чтобы он соответствовал заданному условию. Директива балансировки корректирует часть записей, где выполнено условие по заданному фактору.

Узел Агрегат замещает последовательность входных записей на итоговые, агрегированные выходные записи.

Узел агрегата Новизна, частота, деньги (Recency, Frequency, Monetary - RFM) позволяет рассмотреть хронологические данные транзакций клиента, исключить любые неиспользуемые данные и объединить все оставшиеся данные транзакций в одну строку, где будет представлено, когда в последний раз обращался клиент, сколько транзакций он произвел и какова общая денежная сумма этих транзакций.

Узел Сортировка сортирует записи в восходящем или убывающем порядке на основании значений одного или нескольких полей.

Узел слияния берет несколько входных записей и создает одну выходную запись, содержащую некоторые или все из входных полей. Он полезен для слияния данных из разных источников, например, из внутренних данных о клиентах и приобретенных демографических данных.

Узел присоединения проводит конкатенацию наборов записей. Это полезно для объединения наборов данных с похожими структурами, но различными данными.

Отдельный узел удаляет дублированные записи, или передавая первую отдельную запись в поток данных, или отбрасывая первую запись и вместо этого передавая в поток данных любые дубликаты.

Узел потоковых временных рядов выполняет построение и скоринг моделей временных рядов за один шаг. Этот узел можно использовать с данными как в локальной, так и в распределенной среде; в распределенной среде можно использовать мощность IBM SPSS Analytic Server.

Для многих узлов на палитре Операции с записями требуется использовать выражение CLEM. Если вы знакомы с CLEM, выражение можно ввести в нужном поле. Однако все поля выражений содержат кнопку, открывающую Построитель выражений CLEM, помогающий создавать такие выражения автоматически.

Рисунок 1. Кнопка Построитель выражений

Узел выбора

Узлы выбора можно использовать для выбора или отбрасывания я поднабора записей из потока данных на основе конкретного условия, такого как АД (артериальное давление) = "HIGH" (высокое).

Режим. Указывает, будут ли записи, соответствующие применяемому условию, включаться в поток данных или исключаться из него.

- **Включение.** Выберите эту опцию для включения записей, соответствующих условию выбора.
- **Отклонить.** Выберите эту опцию для исключения записей, соответствующих условию выбора.

Условие. Выводит условие выбора, которое будет использоваться для проверки каждой задаваемой вами записи с применением выражения CLEM. Либо введите выражение в этом окне, либо примените Построитель выражений, нажав кнопку калькулятора (Построителя выражений) справа от окна.

Если вы выбрали отбрасывание записей на основе выражения, такого как следующее:

```
(переменная1='значение1' and переменная2='значение')
```

узел выбора будет также отбрасывать по умолчанию записи, содержащие для всех полей выбора пустые значения. Во избежание этого добавьте к исходному следующее условие:

```
and not(@NULL(переменная1) and @NULL(переменная2))
```

Узлы выбора позволяют также выбрать долю записей. Обычно вы будете использовать для этой операции другой узел - узел выборки. Однако если условие, которое вы хотите задать, сложнее задаваемого указанными параметрами, вы можете создать свое собственное условие при помощи узла выбора. Например, можно создать такое условие:

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

Оно выбирает приблизительно 40% записей, показывающих высокое артериальное давление, и передает эти записи в нисходящий поток для дальнейшего анализа.

Узел выборки

При помощи узлов выборки можно выбрать поднабор записей для анализа или указать долю записей, которые следует отбросить. Поддерживаются разнообразные типы выборок, включая стратифицированную, кластеризованную и неслучайную (структурированную) выборки. Выборки могут использоваться по нескольким причинам:

- Для повышения производительности, благодаря оценке моделей для поднабора данных. Модели, оцененные по выборке, часто также же точны, как и полученные по всему набору данных, и могут быть даже точнее, если повышенная производительность позволит вам опробовать различные методы, которые иначе не удалось бы испытать.
- Для выбора групп связанных записей или транзакций для анализа (такого как выбор всех позиций в покупательской корзине интернет-магазина, или потребительской корзине) или всех свойств в конкретном ближайшем окружении.
- Для идентификации единиц или наблюдений для выборочной проверки с целью исследования гарантии качества, предотвращения мошенничества или с целью защиты.

Примечание: Если вы просто хотите разделить данные на обучающую и контрольную выборки в целях проверки, вместо этого узла можно использовать узел раздела. Дополнительную информацию смотрите в разделе “Узел раздела” на стр. 167.

Типы выборок

Кластеризованные выборки. Выборки групп или кластеров вместо отдельных единиц. Предположим, у вас есть файл данных, содержащий по одной записи для каждого студента. Если вы выполняете кластеризацию по учебным заведениям и размер выборки составляет 50%, будет выбрано 50% учебных заведений и будут выбраны все студенты в каждом выбранном учебном заведении. Студенты в невыбранных учебных заведениях будут отклонены. В среднем можно ожидать, что будет вобрано 50% студентов, но поскольку размер учебных заведений отличается, этот процент может быть неточен. Аналогично, можно кластеризовать позиции покупательских корзин по ID транзакции для гарантированной поддержки всех позиций из выбранных транзакций. Пример кластеризации свойств по городам смотрите в потоке примера *complexsample_property.str*.

Стратифицированные выборки. Выборки независимых неперекрывающихся подгрупп совокупности (или страт). Например, можно гарантировать, что мужчины и женщины будут представлены в выборке в равных долях или что будет представлен каждый район или каждая социально-экономическая группа городского населения. Можно также задать отличающийся размер выборки для каждой страты (например, если вы считаете, что одна группа была представлена в меньшем количестве в исходных данных). Пример стратификации свойств по округам смотрите в потоке примера *complexsample_property.str*.

Систематическая выборка или выборка 1-в-N. Когда трудно обеспечить случайность выборки, элементы могут отбираться систематически (через фиксированные интервалы) или последовательно.

Веса выборок. Веса выборок автоматически вычисляются при извлечении сложной выборки и приблизительно соответствуют "частоте", с которой каждый элемент выборки представлен в исходных данных. Поэтому сумма весов по всей выборке может служить оценкой размера исходных данных.

Выборочная совокупность

Выборочная совокупность определяет потенциальный источник наблюдений для включения в выборку или изучение. В некоторых случаях может оказаться выполнимым идентифицировать каждый отдельный элемент совокупности и включать его в выборку (например, при выборке позиций, сходящих с поточной линии). Но чаще всего у вас не будет доступа к каждому возможному наблюдению. Например, нельзя быть уверенным, кто проголосует на выборах, пока они не пройдут. В этом случае в качестве выборочной совокупности можно использовать список избирателей даже при том, что некоторые зарегистрированные избиратели не проголосуют, а некоторые проголосуют, несмотря на то, что не были внесены в список избирателей на момент вашей проверки списка. У тех, кого нет в выборочной совокупности, нет шансов попасть в выборку. Вопрос, достаточно ли близка ваша выборочная совокупность по характеру к совокупности, которую вы пытаетесь оценить, должен решаться для каждого реального случая.

Опции узла выборки

В соответствии с поставленными требованиями можно выбрать **Простой** или **Сложный** метод.

Опции простого метода выборки

Простой метод позволяет выбрать случайный процент записей, выбрать последовательные записи либо выбрать каждую *n*-ю запись.

Режим. Выберите, следует ли передавать (включать) или отбрасывать (исключать) записи для следующих режимов:

- **Включить выборку.** Включает выбранные записи в поток данных и отбрасывает все остальные. Например, если задать режим **Включить выборку** и опцию **1-в-N** со значением 5, будет включена каждая пятая запись и получен набор данных с размером, приблизительно равным исходному размеру. Это -режим по умолчанию при выборке данных и единственный режим при использовании сложного метода.
- **Отклонить выборку.** Исключает выбранные записи и включает все остальные. Например, если задать режим **Отбросить выборку** и опцию **1-в-N** со значением 5, каждая пятая запись будет отброшена. Этот режим доступен столько с простым методом.

Пример. Выберите метод выборки на основе следующих опций:

- **Первые.** Выберите эту опцию, чтобы использовать выборку последовательных данных. Например, если задать максимальный размер выборки 10000, будут выбраны первые 10000 записей.
- **1-в-N.** Отбирать для выборки данные, передавая или отбрасывая каждую *n*-ю запись. Например, если в качестве *n* задать 5, будет выбрана каждая пятая запись.
- **Случайный %.** Выберите эту опцию для выборки случайного процента данных. Например, если выбрать процент 20, то 20% данных будут либо переданы в поток данных, либо отброшены, в зависимости от выбранного режима. Используйте это поле для задания процента выборки. Можно также указать начальное значение рандомизации при помощи элемента управления **Задать начальное значение рандомизации**.

  Использовать выборку уровня блоков (только для In-Database). Эта опция включается, только если был выбран случайный процент выборки при выполнении исследования In-Database в базе данных Oracle или IBM DB2. При этих условиях выборка уровня данных может оказаться более эффективной.

  Примечание: При различных запусках случайной выборки с одинаковыми параметрами количества возвращаемых строк не будут точно одинаковыми. Это связано с тем, что у каждой входной записи есть вероятность включения в выборку `N/100` (где `N` - это **случайный %**, заданный для узла) и вероятности независимы, поэтому число записей может быть не равно точно `N%`.

Максимальный размер выборки. Задает максимальное количество записей для включения в выборку. Эта опция избыточна, и поэтому отключается, если выбраны опции **Первая** и **Включить**. Кроме того, учтите, что выбранная в сочетании с опцией **Случайный процент** эта опция может помешать выбору определенных записей. Например, если из набора данных с 10 миллионами записей выбрать 50% записей при

максимальном размере выборки три миллиона записей, будут выбраны 50% от первых шести миллионов записей, и не останется никакой возможности выбора для остальных четырех миллионов записей. Чтобы избежать этого ограничения, выберите **Сложный** метод выборки и затребуйте случайную выборку трех миллионов записей без задания переменной кластеризации и стратификации.

Опции сложного метода выборки

Опции сложного метода выборки обеспечивают более точное управление выборкой, включая кластерную, стратифицированную и взвешенную выборки наряду с другими опциями.

Кластеризация и стратификация. Позволяет задать поля кластеризации, стратификации и поля входных весов, если они требуются. Дополнительную информацию смотрите в разделе “Параметры кластеризации и стратификации” на стр. 72.

Тип выборки.

- **Переменный.** Выбирает кластеры или записи в каждой страте случайным образом.
- **Систематическая.** Выбирает записи через фиксированные интервалы. Эта опция работает подобно методу *1-в-N* за исключением того, что позиция первой записи изменяется в зависимости от начального значения рандомизации. Значение *n* определяется автоматически на основе размера или доли выборки.

Единицы для выборки. Можно выбрать долю или количество в качестве базовых единиц для выборки.

Объем выборки. Можно указать размер выборки несколькими способами:

- **Фиксированная.** Позволяет указать общий размер выборки в виде числа или доли.
- **Пользовательский.** Позволяет задать размер выборки для каждой подгруппы или страт. Эта опция доступна, только если во вспомогательном диалоговом окне Кластеризовать и стратифицировать было выбрано поле стратификации.
- **Переменный.** Позволяет пользователю выбрать поле, определяющее размер выборки для каждой подгруппы или страт. Значение этого поля должно быть одинаково для каждой записи в конкретной страте; например, если выборка стратифицируется по округам (county), у всех записей, где *county* = *Surrey*, должно быть одно и то же значение. Это поле должно быть числовым, а его значения должны соответствовать выбранным единицам выборки. Для долей значения должны быть больше 0 и меньше 1; для количеств минимальное значение равно 1.

Минимальный размер выборки для одной страты. Задает минимальное число записей (или минимальное число кластеров, если задано поле кластеризации).

Максимальный размер выборки для одной страты. Задает максимальное число записей или кластеров. Если выбрать эту опцию, не задав поле кластеризации или стратификации, будет выбрана случайная или систематическая выборка.

Задать начальное значение рандомизации При выборке или разделении записей на основе случайного процента эта опция позволяет продублировать одни и те же результаты в другом сеансе. Задав начальное значение, используемое генератором случайных чисел, можно обеспечить назначение одних и тех же записей при каждом вызове узла. Введите нужное начальное значение рандомизации или нажмите кнопку **Сгенерировать**, чтобы автоматически сгенерировать случайное значение. Если эта опция не выбрана, каждый раз при вызове узла будет генерироваться отличающаяся выборка.

Примечание: При использовании опции **Задать начальное значение рандомизации** с записями, читаемыми из базы данных, перед выборкой может потребоваться узел сортировки для обеспечения одинаковых результатов при каждом вызове узла. Это связано с зависимостью начального значения рандомизации от порядка записей, который не гарантированно будет оставаться одним и тем же в реляционной базе данных. Дополнительную информацию смотрите в разделе “Узел сортировки” на стр. 79.

Параметры кластеризации и стратификации

В диалоговом окне Кластеризовать и стратифицировать можно выбрать поля кластеризации, стратификации и взвешивания при построении сложной выборки.

Кластеры. Задает категориальное поле, используемое для кластеризации записей. Записи отбираются на основе принадлежности к кластерам, причем некоторые кластеры включаются, а другие - нет. Но если из данного кластера включается какая-либо запись, то включаются все записи. Например при анализе связей продуктов в покупательских корзинах можно кластеризовать позиции по ID транзакции, чтобы обеспечить поддержку всех позиций из выбранных транзакций. Вместо выборки записей (которая свела бы к нулю информацию о том, какие позиции продаются вместе) можно подготовить выборку транзакций, чтобы обеспечить сохранность всех записей для выбранных транзакций.

Стратифицировать по. Задает категориальное поле, используемое для стратификации записей, чтобы стратифицированная выборка включала в себя отбор элементов из независимых неперекрывающихся подгрупп совокупности (страт). В случае 50% выборки, стратифицированной, например, по полу, должны быть отобраны две 50% выборки: одна для мужчин и одна для женщин. Например, страты могут быть социоэкономическими группами, рабочими категориями, возрастными или этническими группами, что позволяет обеспечить адекватные размеры выборок для исследуемых подгрупп. Если в исходном наборе данных женщин в три раза больше, чем мужчин, это соотношение будет сохранено, благодаря подготовке выборки отдельно для каждой группы. Можно также задать несколько полей стратификации (например, подготовив выборку линеек продуктов в регионах или наоборот).

Примечание: Если выполняется стратификация по полю, в котором есть пропущенные значения (пустые или системные пропущенные значения, пробелы и пробельные или пользовательские пропущенные значения), то вы не сможете задать пользовательские размеры выборок для страт. Если вы хотите использовать пользовательские размеры выборок при стратификации по полю с пропущенными или пробельными значениями, их нужно будет заполнить в восходящем потоке.

Использовать входные веса. Задает полек, используемое для взвешивания перед выборкой. Например, если поле веса содержит значения, ранжируемые от 1 до 5, вероятность выбора записей с весом 5 будет в пять раз выше. Значения этого поля будут перезаписаны окончательными выходными весами, сгенерированными узлом (смотрите следующую тему).

Новый выходной вес. Задает имя поля, где записываются веса, если не задано поле входного веса. (Если поле входного веса задано, его значения заменяются окончательными весами, как замечено выше, и никакого отдельного поля выходного веса не создается.) Значения выходного веса указывают число записей, представляемых каждой записью выборки в исходных данных. Сумма значений весов дает оценку размера выборки. Например, при 10% выборке выходной вес составит 10 для всех записей, указывая, что каждая запись выборки представляет примерно десять записей в исходных данных. В стратифицированной или взвешенной выборке значения выходных весов могут отличаться, в зависимости от доли выборки для каждой страты.

Комментарии

- Кластеризованная выборка полезна, если невозможно получить полный список совокупности, выборку которой вы хотите подготовить, но можно получить полные списки для определенных групп кластеров. Она также используется, если случайная выборка сгенерировала бы список объектов исследования, обращение к которому оказалось бы практически невозможным. Например, проще было бы посетить фермеров в одном округе, чем в случае выбора фермеров, рассеянных по всем округам в стране.
- Можно задать и поля кластеризации, и поля стратификации для независимой выборки кластеров в каждой из страт. Например, можно подготовить выборку значений свойств, стратифицированных по округам, и кластеризацию по городам в каждом округе. Это гарантирует построение независимой выборки городов из каждого округа. Некоторые города будут включены, а другие - нет, но для каждого включенного города будут включены все его свойства.
- Для случайной выборки единиц из каждого кластера можно связать между собой два узла выборки. Например, сначала можно подготовить выборку по районам, стратифицированным по округам, как

описано выше. Затем присоединить второй узел и выбрать *город* как поле стратификации, позволяющее подготовить выборку доли записей из каждого района.

- В случаях, где для уникальной идентификации кластеров требуется сочетание полей, можно сгенерировать новое поле при помощи узла извлечения. Например, если несколько магазинов используют одинаковую систему нумерации для транзакций, можно получить новое поле, соединяющее последовательно ID магазинов и транзакций.

Размеры выборки для страт

При построении стратифицированной выборки в качестве опции по умолчанию используется выборка одинаковых долей записей или кластеров для каждой страты. Если одна группа превосходит численностью другую, например, в 3 раза, это соотношение, как правило, будет предпочтительней сохранить. Однако в противном случае можно задать размер выборки отдельно для каждой страты.

В диалоговом окне Размеры выборок для страт выводится список с указанием каждого значения поля стратификации, позволяющего переопределить размер по умолчанию для данной страты. Если выбрано несколько полей стратификации, в списке указывается каждое возможное сочетание этих значений, что позволяет, например, задать размер каждой этнической группы в каждом городе в каждой стране. Размеры задаются как доли или количества, определяемые текущим значением параметра на узле выборки.

Чтобы задать размеры выборки для страт

1. На узле выборки выберите опцию **Сложный** и выберите одно или несколько полей стратификации. Дополнительную информацию смотрите в разделе “Параметры кластеризации и стратификации” на стр. 72.
2. Выберите **Пользовательский** и выберите **Задать размеры**.
3. В нижнем левом углу диалоговом окне Размеры выборок для страт нажмите кнопку **Читать значения**, чтобы заполнить вывод на экран. В некоторых случаях может потребоваться инстанциировать значения в источнике восходящего потока или на узле Тип. Дополнительную информацию смотрите в разделе “Что такое инстанциация?” на стр. 133.
4. Щелкните по любой строке, чтобы переопределить размер по умолчанию для данной страты.

Замечания о размере выборки

Пользовательские размеры выборки могут оказаться полезны в случае отличающейся дисперсии различных страт, например, чтобы сделать размеры выборки пропорциональными среднеквадратичному отклонению. (Если наблюдения в страте отличаются больше, для получения представительной выборки потребуется выборка большего их числа.) Если же страта мала, можно использовать более высокую долю выборки для гарантированного включения в нее минимального числа наблюдений.

Примечание: Если выполняется стратификация по полю, в котором есть пропущенные значения (пустые или системные пропущенные значения, пробелы и пробельные или пользовательские пропущенные значения), то вы не сможете задать пользовательские размеры выборок для страт. Если вы хотите использовать пользовательские размеры выборок при стратификации по полю с пропущенными или пробельными значениями, их нужно будет заполнить в восходящем потоке.

Узел Баланс

Узлы Баланс можно использовать для исправления условий дисбаланса в наборах данных, чтобы они соответствовали заданным критериям тестирования. Допустим, в наборе данных есть только два значения, *low* (низкое) и *high* (высокое), и что 90% наблюдений - *low*, тогда как всего 10% наблюдений - *high*. Множество методов моделирования испытывают трудности с такими смещенными данными, так как стремятся изучить только исход *low* и игнорировать исход *high*, поскольку он более редок. Если данные хорошо сбалансированы, с примерно одинаковым количеством исходов *low* и *high*, у моделей будет выше шанс найти шаблоны, различающие эти две группы. В этом случае узел Баланс полезен для создания директивы балансировки, уменьшающей число наблюдений с исходом *low*.

Балансировка выполняется посредством дублирования и последующего отбрасывания записей на основе задаваемого вами условия. Записи, для которых никакое условие не удерживается, всегда ее проходят. Поскольку этот процесс выполняется посредством дублирования и/или отбрасывания записей, исходная последовательность ваших данных теряется в операциях нисходящего потока. Перед добавлением узла Баланс в поток данных обязательно получите все значения, связанные с последовательностью.

Примечание: Узлы балансировки могут быть обобщены автоматически из диаграмм и гистограмм распределения. Например, вы можете сбалансировать данные, чтобы показать равные соотношения по всем категориям категориального поля, как показано на графике распределения.

Пример. При построении потока RFM для выявления недавних заказчиков, положительно ответивших на предыдущие маркетинговые кампании, маркетинговый отдел компании продаж использует узел Баланс для балансировки разницы между ответами true и false в данных.

Задание опций для узла балансировки

Директивы балансировки записей. Возвращает список текущих директив балансировки. Каждая директива содержит коэффициент и условие, указывающие программам "увеличить соотношение записей пропорционально заданному коэффициенту, где соблюдается указанное условие". Коэффициент меньше 1,0 означает, что соотношение указанных записей будет понижено. Например, если вы хотите уменьшить число записей, где препарат Y является терапевтическим, можно создать директиву балансировки с коэффициентом 0,7 и условием `Drug = "drugY"`. Эта директива означает, что число записей, где препарат Y является терапевтическим, будет уменьшено до 70% для всех операций нисходящего потока.

Примечание: Коэффициенты балансировки для операции уменьшения можно задавать с числом знаков после запятой не более четырех. Коэффициенты меньше 0,0001 приведут к ошибке, поскольку результаты будут вычислены неверно.

- **Создайте условия**, нажав кнопку справа от текстового поля. Будет вставлена пустая строка для ввода новых условий. Чтобы создать для условия выражение CLEM, нажмите кнопку Построитель выражений.
- **Удалите директивы** при помощи красной кнопки удаления.
- **Отсортируйте директивы** при помощи кнопок со стрелками вверх и вниз.

Сбалансировать только данные обучения. Если в потоке представлено поле раздела, эта опция балансирует данные только в разделе обучения. В частности, это может оказаться полезным при генерировании скорректированных оценок склонности, где требуется несбалансированный раздел тестирования или проверки. Если поле раздела в потоке не представлено (либо задано несколько полей раздела), эта опция игнорируется, и балансируются все данные.

Узел агрегации

Агрегация - это задача по подготовке данных, часто используемая для сокращения размера набора данных. Перед агрегацией надо уделить время на очистку данных, сосредоточив внимание, главным образом, на пропущенных значениях. После агрегирования потенциально полезная информация, относящаяся к пропущенным значениям, может быть потеряна.

При помощи узла агрегации можно заменить последовательность входных записей сводными, агрегированными выходными записями. Например, у вас может быть набор входных записей о продажах, таких как показанные в следующей таблице.

Таблица 13. Входной пример записей о продажах

Возраст	Пол	Регион	Отрасль	Число продаж
23	Н	Ю	8	4
45	Н	Ю	16	4
37	Н	Ю	8	5

Таблица 13. Входной пример записей о продажах (продолжение)

Возраст	Пол	Регион	Отрасль	Число продаж
30	Н	Ю	5	7
44	Н	С	4	9
25	Н	С	2	11
29	Ж	Ю	16	6
41	Ж	С	4	8
23	Ж	С	6	2
45	Ж	С	4	5
33	Ж	С	6	10

Эти записи можно агрегировать с полями *Пол* и *Регион* в качестве ключевых. Затем выбрать опцию агрегирования *Возраст* с режимом **Среднее** и *Продажи* с режимом **Сумма**. В диалоговом окне узла агрегации выберите опцию **Включить число записей в поле**, и агрегированный вывод станет таким, как в следующей таблице.

Таблица 14. Пример агрегированной записи

Возраст (средний)	Пол	Регион	Число продаж (сумма)	Число записей
35,5	Ж	С	25	4
29	Ж	Ю	6	1
34,5	Н	С	20	2
33,75	Н	Ю	20	4

Отсюда можно узнать, например, что средний возраст четырех сотрудниц отдела сбыта в северном регионе - 35,5 лет, а суммарный итог их продаж - 25 единиц.

Примечание: Такие поля, как *Отрасль*, отбрасываются автоматически, если никакой режим агрегирования не задан.

Задание опций для узла агрегации

На узла агрегации задаются следующие опции.

- Одно или несколько полей ключей для использования в качестве категорий для агрегации.
- Одно или несколько полей агрегации, для которых вычисляются значения агрегации.
- Одна или несколько мод агрегации (типов агрегации) для вывода для каждого поля агрегации.

Можно также задать моды агрегации по умолчанию, используемые для вновь добавляемых полей, и использовать выражения (подобные формулам) для категоризации агрегаций.

Заметим, что в отношении производительности операции агрегации могут выиграть от включения поддержки параллельной обработки.

Поля ключей. Возвращает список полей, которые можно использовать в качестве категорий для агрегации. В качестве ключей можно использовать и непрерывные (числовые), и категориальные поля. Если выбрать нескольких полей ключей, будут образованы сочетания значений с целью генерирования значения ключа для агрегирования записей. Для каждого поля ключа уникальности будет сгенерировано по одной агрегированной записи. Например, если полями ключей являются *Пол* и *Регион*, у каждого уникального сочетания *М* и *Ж* с регионами *С* и *Ю* (четыре уникальных сочетания) будет агрегированная запись. Для добавления поля ключа используйте кнопку Средство выбора полей в правой части окна.

Оставшаяся часть этого диалогового окна разделена на две основные области - **Основные агрегации** и **Выражения агрегации**.

Базовые агрегации

Агрегированные поля. Возвращает список полей, для которых будут агрегированы значения, наряду с выбранными режимами агрегации. Для добавления полей в этот список используйте кнопку Средство выбора полей в правой части окна. Доступны следующие режимы агрегации.

Примечание: Некоторые режимы неприменимы к нечисловым полям (например, **Сумма** для поля даты/времени). Режимы, которые нельзя использовать с выбранным полем агрегации, отключаются.

- **Сумма.** Выберите эту опцию для возврата суммированных значений для каждого сочетания полей. Эта сумма представляет собой итог для всех значений по всем наблюдениям без пропущенных значений.
- **Среднее значение.** Выберите эту опцию для возврата средних значений для каждого сочетания полей. Среднее - это мера центральной тенденции; арифметическое среднее (сумма, деленная на число наблюдений).
- **Минимум.** Выберите эту опцию для возврата минимальных значений для каждого сочетания полей.
- **Максимум.** Выберите эту опцию для возврата максимальных значений для каждого сочетания полей.
- **Ср. отклонение.** Выберите эту опцию для возврата среднеквадратичного отклонения для каждого сочетания полей. Среднеквадратичное отклонение - это мера разброса относительно среднего значения; оно равно квадратному корню из дисперсии.
- **Медиана.** Выберите эту опцию для возврата значений медианы для каждого сочетания полей. Медиана представляет собой меру центральной тенденции, которая нечувствительна к выбросам, в отличие от среднего значения, которое могут исказить несколько экстремально больших или малых значений. Другое название - 50-й процентиль или 2-й квартиль.
- **Количество.** Выберите эту опцию для возврата непустых значений для каждого сочетания полей.
- **Дисперсия.** Выберите эту опцию для возврата значений дисперсии для каждого сочетания полей. Дисперсия - это мера разброса относительно среднего значения; равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньшее числа наблюдений.
- **1-я квартиль.** Выберите эту опцию для возврата значений 1-й квартили (25-й процентили) для каждого сочетания полей.
- **3-я квартиль.** Выберите эту опцию для возврата значений 3-й квартили (75-й процентили) для каждого сочетания полей.

Примечание: Когда выполняется поток, содержащий узел агрегации, значения, возвращаемые для 1-й и 3-й квартилей при обратном переносе SQL в базу данных Oracle, могут отличаться от соответствующих значений, возвращаемых в собственном режиме.

Режим по умолчанию. Задайте моду агрегации по умолчанию, используемую для вновь добавляемых полей. Если вы часто используете одну и ту же агрегацию, выберите здесь одну или несколько мод и нажмите справа кнопку Применить ко всем, чтобы применить выбранные моды ко всем полям, указанным ниже.

Расширение имени нового поля. Выберите эту опцию, чтобы добавить суффикс или префикс (такой как "1" или "new") к дубликатам агрегированных полей. Например, если вы выбрали опцию Суффикс и задали в качестве расширения "1", в результате агрегации минимальных значений для поля *Возраст* будет сгенерировано имя поля *Age_Min_1*. *Примечание*: В новое поле автоматически добавляются расширения агрегации, такие как *_Min* или *Max_*, указывая тип выполненной агрегации. Выберите **Суффикс** или **Префикс**, чтобы указать предпочитаемый вами тип расширения.

Включить количество записей в поле. Выберите эту опцию, чтобы включить в каждую выходную запись избыточное поле, называемое *Record_Count* (по умолчанию). Это поле указывает, сколько записей было агрегировано для образования каждой записи агрегации. Создайте для этого поля пользовательское имя, введя его в поле редактирования.

Примечание: При вычислении агрегатов системные пустые значения исключаются, но они включаются в число записей. С другой стороны, пробельные значения включаются и в агрегацию, и в число записей. Чтобы исключить эти значения, можно при помощи узла заполнения заменить пробелы на пустые значения. Пробелы можно также удалить при помощи узла выбора.

Выражения агрегации

Выражения подобны формулам, создаваемым из значений, имен полей, операций и функций. В отличие от функций, которые работают в каждый момент с одной записью, выражения агрегации работают с группой, набором или собранием записей.

Примечание: Создавать выражения агрегации можно, только если поток содержит соединение с базой данных (при помощи узла Источник баз данных).

Новые выражения создаются как производные поля; для создания выражения вы используете функции *Агрегации баз данных*, доступные в построителе выражений.

Дополнительную информацию о построителе выражений смотрите в Руководстве пользователя IBM SPSS Modeler (ModelerUsersGuide.pdf).

Заметим, что существует связь между **Ключевыми полями** и любыми выражениями агрегации, которые вы создаете, поскольку выражения агрегации группируются по ключевому полю.

Допустимые выражения агрегации - это те, которые оценивают агрегированный выход; пара примеров допустимых выражений агрегации и правила, которые ими управляют, приведены ниже.

- Вы можете использовать скалярные функции, чтобы объединить вместе несколько функций агрегации и получить единый результат агрегации. Например:

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

- Функция агрегации может работать с результатами нескольких скалярных функций. Например:

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

Параметры оптимизации агрегации

На вкладке Оптимизация задаются следующие опции.

Ключи непрерывны. Выберите эту опцию, если знаете, что все записи с одними и теми же значениями ключей сгруппированы вместе на входе (например, если ко входу применена сортировка по полям ключей). Выбрав эту опцию, можно увеличить производительность.

Разрешить приближения для медианы и квартилей. Порядковые статистики (медиана, первая квартиль и третья квартиль) на данный момент не поддерживаются при обработке данных в Analytic Server. Если используется Analytic Server, можно включить этот переключатель для использования приблизительного значения для этой статистики вместо полученного разбиением данных по интервалам с последующим вычислением оценки для статистики на основании распределения по интервалам. По умолчанию эта опция не включена.

Число интервалов. Доступно только при включении переключателя **Разрешить приближения для медианы и квартилей**. Выберите количество интервалов, которые будут использоваться для оценки статистики; это количество интервалов влияет на **Максимальный % ошибки**. По умолчанию число интервалов - 1000, что соответствует максимальной ошибке в 0,1% диапазона.

Узел агрегации RFM

Узел агрегации RFM (Recency, Frequency, Monetary - недавность, частота, деньги) позволяет взять хронологические данные сделок клиентов, убрать из них все неиспользуемые данные и объединить все оставшиеся данные сделок в одну строку, используя в качестве ключа уникальный ID клиента, указывающий, когда последний раз у них с вами было дело (недавность), сколько сделок они заключили (частота) и какова итоговая ценность этих сделок (деньги).

Перед любой агрегацией надо уделить время на очистку данных, сосредоточив внимание, главным образом, на всех пропущенных значениях.

После идентификации и преобразовании данных при помощи узла агрегации RFM можно применить узел анализа RFM для выполнения дальнейшего анализа. Дополнительную информацию смотрите в разделе “Узел анализа RFM” на стр. 163.

Имейте в виду, что после прохождения файла данных через узел агрегации RFM у него не будет никаких значений назначения, поэтому перед тем, как его можно будет использовать в качестве входных данных для дальнейшего предсказательного анализа с какими-либо узлами моделирования (такими как C5.0 или CHAID), этот файл нужно будет слить с другими данными о клиентах (например, путем сопоставления ID клиентов). Дополнительную информацию смотрите в разделе “Узел слияния” на стр. 80.

Узлы агрегации RFM и анализа RFM в IBM SPSS Modeler конфигурируются под использование независимого разделения на интервалы; то есть, они ранжируют данные и разделяют на интервалы для каждого измерения значения недавности, частоты или денег безотносительно к остальным двум их значениям.

Задание опций для узла агрегации RFM

Вкладка Параметры на узле агрегации RFM содержит следующие поля.

Вычислить недавность относительно. Задайте дату, относительно которой будет вычислена недавность сделок (транзакций). Это может быть либо **Фиксированная дата**, вводимая вами, либо **Сегодняшняя дата**, задаваемая системой. **Сегодняшняя дата** вводится по умолчанию и автоматически обновляется при вызове узла.

Последовательные ID. Если данные предварительно были отсортированы так, что все записи с одним и тем же ID появляются вместе в потоке данных, выберите эту опцию, чтобы повысить скорость обработки. Если данные предварительно не были отсортированы (или вы в этом не уверены), оставьте эту опцию невыбранной, и узел отсортирует данные автоматически.

ID. Выберите поле, которое будет использоваться для идентификации заказчиков и их сделок. Для вывода списка доступных для выбора полей используйте кнопку Средство выбора полей в правой части окна.

Дата. Выберите поле даты, которое будет использоваться для вычисления недавности относительно даты. Для вывода списка доступных для выбора полей используйте кнопку Средство выбора полей в правой части окна.

Имейте в виду, что для использования в качестве входного требуется поле с хранением даты или отметки времени в подходящем формате. Например, при наличии строкового поля со значениями, такими как *Jan 2007*, *Feb 2007* и так далее его можно преобразовать в поле дат при помощи поля фильтра и функции `to_date()`. Дополнительную информацию смотрите в разделе “Преобразование хранения при помощи узла заполнения” на стр. 152.

Значение. Выберите поле, которое будет использоваться для вычисления итогового денежного значения сделок заказчика. Для вывода списка доступных для выбора полей используйте кнопку Средство выбора полей в правой части окна. *Примечание*: Это должно быть числовое значение.

Расширение имени нового поля. Выберите эту опцию, чтобы добавлять к вновь генерируемым полям недавности, частоты и денег суффикс или префикс, например: "12_месяц". Выберите **Суффикс** или **Префикс**, чтобы указать предпочитаемый вами тип расширения. Например, это может быть полезно при проверке нескольких периодов времени.

Отбрасывать записи со значение ниже. При необходимости можно задать минимальное значение, ниже которого никакие подробности сделок при вычислении итогов RFM использоваться не будут. Единицы значения связываются с выбранным полем **Значение**.

Включить только недавние транзакции. При анализе большой базы данных можно указать использовать только самые последние записи. Можно выбрать использование данных, записанных либо после определенной даты, либо за недавний промежуток времени.

- **Транзакции с датой после.** Задайте дату транзакций, после которой записи будут включаться в анализ.
- **Транзакции за последние.** Задайте число периодов и их тип (дни, недели, месяцы или годы) за датой **Вычислять недавность относительно**, после которой записи будут включаться анализ.

Сохранить дату второй наиболее свежей транзакции. Если вы хотите знать дату второй наиболее свежей транзакции для каждого заказчика, включите этот переключатель. После этого можно также включить и переключатель **Сохранить дату третей наиболее свежей транзакции**. Например, это поможет выявить заказчиков, которые могли заключить множество сделок значительное время тому назад, но только одну сделку недавно.

Узел сортировки

При помощи узлов сортировки можно отсортировать записи по возрастанию или по убыванию на основе значений одного или нескольких полей. Например, узлы сортировки часто используются для просмотра и выбора записей с чаще всего встречающимися значениями данных. Как правило, данные сначала можно агрегировать при помощи узла агрегации, а затем, применив узел сортировки, отсортировать агрегированные данные по убыванию количества записей. Вывод этих результатов в виде таблицы позволит вам изучить данные и принять решения, такие как выбор записей лучших 10 покупателей.

Вкладка Параметры на узле сортировки содержит следующие поля.

Сортировать по. Все поля, выбираемые для использования в качестве ключей сортировки, выводятся в таблице. Поле ключа лучше всего работает для сортировки, если оно числовое.

- **Добавьте поля** в этот список при помощи кнопки выбора полей в правой части окна.
- **Выберите порядок**, нажав кнопку со стрелкой **По возрастанию** или **По убыванию** в столбце *Порядок* этой таблицы.
- **Удалите поля** при помощи красной кнопки удаления.
- **Отсортируйте директивы** при помощи кнопок со стрелками вверх и вниз.

Порядок сортировки по умолчанию. Выберите либо **По возрастанию**, либо **По убыванию**, чтобы использовать этот порядок в качестве порядка сортировки по умолчанию при добавлении новых полей выше.

Примечание: Узел сортировки не применяется, если ниже по потоку модели есть отличительный узел. Информацию об узле Разделять смотрите в разделе “Отличительный узел” на стр. 89.

Параметры оптимизации сортировки

При работе с данными, которые, как вы знаете, уже были отсортированы по некоторым полям ключей, можно указать, какие поля уже отсортированы, что позволит системе отсортировать остальные данные более эффективно. Например, вы хотите отсортировать данные по полям *Возраст* (по убыванию) и *Препарат* (по возрастанию), но знаете, что данные уже отсортированы по полю *Возраст* (по убыванию).

Данные предварительно отсортированы. Задает, отсортированы ли уже данные по одному или нескольким полям.

Задать существующий порядок сортировки. Укажите поля, которые уже были отсортированы. В диалоговом окне Выбрать поля добавьте поля в список. В столбце *Упорядочить* укажите, сортируется ли каждое поле по возрастанию или убыванию. Если задается несколько полей, убедитесь, что вы указываете их в списке в правильном порядке сортировки. При помощи кнопок со стрелками справа от списка расставьте поля в правильном порядке. Если при указании верного существующего порядка сортировки вы сделаете ошибку, при вызове потока появится сообщение об ошибке с номером записи, где сортировка не соответствует заданной вами.

Примечание: Скорость сортировки может выиграть от включения поддержки параллельной обработки.

Узел слияния

Узел слияния предназначен для создания из нескольких входных записей одной, содержащей все или некоторые входные поля. Эта операция полезна, если вы хотите слить данные из различных источников, например, данные о покупателях и приобретенные демографические данные. Слияние данных можно выполнить несколькими способами.

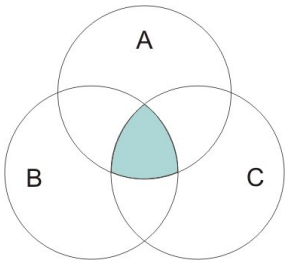
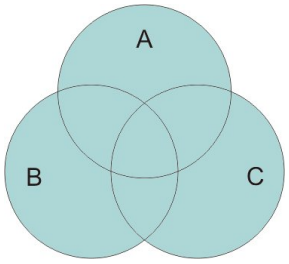
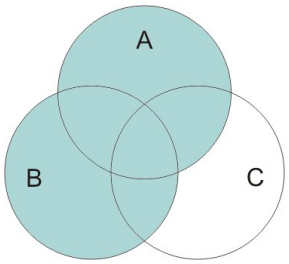
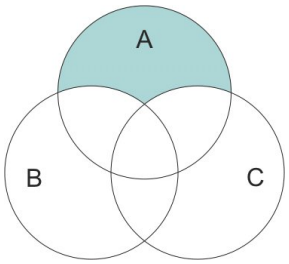
- Слияние по **Порядку** конкатенирует соответствующие записи изо всех источников по порядку их ввода, пока не будет исчерпан наименьший источник данных. При использовании этой опции важно, чтобы данные были отсортированы при помощи узла сортировки.
- Слияние при помощи поля **Ключ**, такого как *ID покупателя*, позволяет указать способ сопоставления записей из одного источника данных с записями из одного или нескольких других источников. Возможны объединения нескольких типов, включая внутреннее объединение, полное внешнее объединение, частичное внешнее объединение и антиобъединение. Дополнительную информацию смотрите в разделе “Типы объединений”.
- Слияние по **Условию** означает, что можно задать условие, которое должно выполняться для возможности слияния. Условие можно задать непосредственно на узле или построить при помощи Построителя выражений.
- Слияние по **Ранжированному условию** - это левое внешнее объединение с задаваемым условием, которое должно быть выполнено, чтобы произошло слияние, и с выражением ранжирования, по которому производится сортировка совпадений от низкого до высокого приоритета. Наиболее часто используемое для слияния геопространственных данных условие можно задать непосредственно на узле или построить с помощью Построителя выражений.

Типы объединений

При использовании поля ключа для слияния данных полезно потратить немного времени, чтобы продумать, какие записи исключать и какие включать. Существуют разнообразные объединения, которые подробно описаны ниже.

Два основных типа объединений называют внутренними и внешними объединениями. Эти методы часто используются для слияния таблиц из связанных баз данных на основе общепринятых значений поля ключа, такого как *ID клиента*. Внутренние объединения допускают чистое слияние и выходной набор данных, включающий только полные записи. Внешние объединения тоже включают полные записи из объединенных слиянием данных, но допускают также и включение уникальных данных из одной или нескольких входных таблиц.

Эти разрешенные типы объединений описаны гораздо подробнее ниже.

	**Внутреннее объединение** включает только записи, с которых значение для поля ключа является общим для всех входных таблиц. То есть, несовпадающие записи во внешнее соединение включены не будут.
	**Полное внешнее объединение** включает все записи (и несовпадающие, и совпадающие) из входных таблиц. Левое и правое внешние объединения, называемые частичными внешними объединениями, описаны ниже.
	**Частичное внешнее объединение** включает все записи, сопоставленные при помощи поля ключа, а также несопоставленные записи из заданных таблиц. (Или, иными словами, все записи из некоторых таблиц и только совпадающие записи из остальных таблиц.) Таблицы (такие как показанные здесь A и B) можно выбрать для включение во внешнее объединение при помощи кнопки Выбрать на вкладке Слить. Частичные объединения называют также левыми или правыми внешними объединениями, если выполняется слияние всего двух таблиц. Поскольку IBM SPSS Modeler разрешает выполнять слияние более двух таблиц, мы называем это частичным внешним объединением.
	**Антиобъединение** включает только несопоставленные записи для первой входной таблицы (показанной здесь таблицы A). Этот тип объединения противоположен внутреннему объединению и не включает полные записи в выходном наборе данных.

Например, при наличии у вас информации о фермах в одном наборе данных и связанных с фермами страховых исках в другом, вы можете при помощи опций слияния сопоставить записи из первого источника с записями во втором.

Чтобы определить, заполнил ли клиент в вашем примере с фермами страховой иск, воспользуйтесь опцией объединения для возврата списка, показывающего где совпадают все ID из этих двух примеров.

	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...

Рисунок 2. Вывод примера для слияния внутреннего объединения

Опция полного внешнего объединения при ее использовании возвращает из входных таблиц и совпадающие, и несовпадающие записи. Для всех недостающих значений будет использоваться системное пропущенное значение (`$null$`).

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

Рисунок 3. Вывод примера для слияния полного внешнего объединения

Частичное внешнее объединение включает все записи, сопоставленные при помощи поля ключа, а также несопоставленные записи из заданных таблиц. В таблице выводятся все сопоставленные записи из поля ID, а также все сопоставленные записи из первого набора данных.

	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

Рисунок 4. Вывод примера для слияния частичного внешнего объединения

При использовании антиобъединения в таблице возвращаются только несопоставленные записи для первой входной таблицы.

	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

Рисунок 5. Вывод примера для слияния антиобъединения

Задание метода и ключей слияния

Вкладка Слить на узле слияния содержит следующие поля.

Способ слияния Выберите способ, который будет использоваться для слияния записей. Выбор одной из опций, **Ключи** или **Условие**, активирует нижнюю половину диалогового окна.

- **По порядку** Эта опция выполняет слияние записей, упорядочивая их так, что что *n*-ная запись из каждого входного поля объединяется слиянием для генерирования *n*-го выходного поля. Когда какая-либо запись остается без соответствующей входной записи, выходные записи более не генерируются. Это означает, что число созданных записей будет равно числу записей в наименьшем наборе данных.
- **Ключи** Использует поле ключа, такое как *ID транзакции*, для слияния записей с одним и тем же значением в поле ключа. Это эквивалентно EQUI-объединению базы данных. Если значение ключа встречается несколько раз, возвращаются все возможные сочетания. Например, если записи с одинаковым значением ключа *A* содержат в других полях отличающиеся значения *B*, *C* и *D*, слитые поля генерируют отдельную запись для каждого сочетания: *A* со значением *B*, *A* со значением *C* и *A* со значением *D*.

  Примечание: Пустые значения (NULL) в методе слияния по ключу не считаются идентичными и объединены не будут.
- **Условие** Используйте эту опцию для задания условия для слияния. Дополнительную информацию смотрите в разделе “Задание условий для слияния” на стр. 83.
- **Ранжированное условие** Используйте эту опцию для указания, будут ли сливаться все парные сравнения строк в первичном и во вторичном наборе данных; используйте выражение ранжирования, чтобы

упорядочить несколько совпадений от низкого до высокого приоритета. Дополнительную информацию смотрите в разделе “Задание условий ранжирования для слияния” на стр. 84.

Возможные ключи Возвращает в списке только поля с точно совпадающими именами полей во всех входных источниках данных. Выберите в этом списке поле и при помощи кнопки со стрелкой добавьте его в качестве поля, используемого для слияния записей. Может использоваться несколько полей ключа. Несовпадающие входные поля можно переименовать при помощи узла фильтра или вкладки Фильтр узла источника.

Ключи для слияния Возвращает список всех полей, используемых для слияния записей из входных источников данных на основе полей ключей. Для удаления ключа из списка выберите его и при помощи кнопки со стрелкой возвратите этот ключ в список Возможные ключи. При выборе нескольких полей ключей включается опция ниже.

Комбинировать повторения ключевых полей Если выше выбрано несколько полей ключа, эта опция гарантирует, что будет существовать всего одно входное поле с данным именем. Эта опция включается по умолчанию за исключением тех случаев, когда потоки были импортированы из прежних версий IBM SPSS Modeler. Если эта опция отключена, дубликаты полей ключей должны быть переименованы или исключены при помощи вкладки Фильтр в диалоговом окне узла слияния.

Включать только соответствующие записи (внутреннее объединение) Выберите эту опцию для слияния только полных записей.

Включать и соответствующие, и несоответствующие записи (полное внешнее объединение) Выберите эту опцию для выполнения "полного внешнего объединения". Это означает, что если значения для поля ключа представлены не во всех входных таблицах, неполные записи все еще будут сохраняются. Неопределенное значение (`$null$`) добавляется в поле ключа и включается в выходную запись.

Включать соответствующие и выбранные несоответствующие записи (частичное внешнее объединение) Выберите эту опцию для выполнения "частичного внешнего объединения" таблиц, выбираемых во вспомогательном диалоговом окне. Нажмите кнопку **Выбрать**, чтобы задать таблицы, для которых будут сохраняться неполные записи при слиянии.

Включать в первый набор данных записи, не совпадающие ни с какими другими (антиобъединение) Выберите эту опцию для выполнения типа "антиобъединения", при котором только несовпадающие записи из первого набора передаются вниз по потоку. Порядок входных наборов данных можно задать при помощи кнопок со стрелками на в кладке Входные поля. Этот тип объединения не включает полные записи в выходной набор данных. Дополнительную информацию смотрите в разделе “Типы объединений” на стр. 80.

Задание данных для частичных объединений

Для частичного внешнего объединения нужно выбрать одну или несколько таблиц, для которых будут сохраняться неполные записи. Например, вы можете сохранить все записи из таблицы Клиент, при сохранении из таблицы Ссуда под недвижимость только сопоставленных записей.

Столбец Внешнее объединение. В столбце *Внешнее объединение* выберите наборы данных для их полного включения. Для частичного объединения будут сохранены перекрывающиеся записи, а также неполные записи для выбранных здесь наборов данных. Дополнительную информацию смотрите в разделе “Типы объединений” на стр. 80.

Задание условий для слияния

Задав метод слияния **Условие**, можно задать одно или несколько условий, которые должны выполняться для возможности слияния.

Условия можно либо ввести непосредственно в поле Условие, либо построить их при помощи Построителя выражений, щелкнув по значку калькулятора справа от поля.

Добавить теги дублированным именам полей для исключения конфликтов слияния Если в двух или более наборах данных для слияния есть одинаковые имена полей, включите этот переключатель, чтобы добавить различные теги префикса в начало заголовков столбцов полей. Например, если есть два поля с именем *Имя*, в слиянии будут поля *1_Имя* и *2_Имя*. Если тег был переименован в источнике данных, это новое имя будет использоваться вместо числового тега префикса. Если не включить этот переключатель, а среди данных окажутся дублированные имена, появится предупреждение справа от переключателя.

Задание условий ранжирования для слияния

Слияние по ранжированному условию можно рассматривать как слияние левого внешнего объединения по условию; левая часть слияния - это первичный набор данных, где каждая запись представляет собой событие. Например, в модели, используемой для обнаружения паттернов в данных о преступлении, первичный набор данных - это данные о самом преступлении и связанная с ним информация (положение, тип и так далее). В этом примере правая сторона может содержать соответствующие наборы геопространственных данных.

Слияние использует условие слияния и выражение ранжирования. Условие слияния может использовать геопространственную функцию, такую как *within* или *close_to*. При слиянии все поля наборов данных правой части добавляются к набору данных левой части, но при нескольких совпадениях возникает поле списка. Например:

- Левая часть: данные преступления
- Правая часть: набор данных Районы и набор данных Дороги
- Условия слияния: данные преступления *within* (внутри) районов и *close_to* (близко к) дорогам, а также дано определение, что считается близким.

В данном примере, если преступление произошло на необходимом расстоянии *close_to* трем дорогам (и для числа возвращаемых совпадений задано по крайней мере значение 3), все три дороги будут возвращены как элемент списка.

Задав метод слияния **Ранжированное условие**, можно задать одно или несколько условий, которые должны выполняться для возможности слияния.

Первичный набор данных Выберите первичный набор данных для слияния; поля из всех других наборов данных будут добавляться к выбранному набору. Он может рассматриваться как левая часть выражения слияния внешнего объединения.

При выборе первичного набора данных все другие входные наборы данных, которые соединены с узлом Слияние, автоматически перечисляются в таблице **Слияния**.

Добавить теги дублированным именам полей для исключения конфликтов слияния Если в двух или более наборах данных для слияния есть одинаковые имена полей, включите этот переключатель, чтобы добавить различные теги префикса в начало заголовков столбцов полей. Например, если есть два поля с именем *Имя*, в слиянии будут поля *1_Имя* и *2_Имя*. Если тег был переименован в источнике данных, это новое имя будет использоваться вместо числового тега префикса. Если не включить этот переключатель, а среди данных окажутся дублированные имена, появится предупреждение справа от переключателя.

Слияния

Набор данных

Показывает имена вторичных наборов данных, подключенных как входные наборы на узел Слияние. По умолчанию при наличии нескольких вторичных наборов данных они перечисляются в том порядке, в котором они подсоединились к узлу Слияние.

Условие слияния

Введите уникальные условия для слияния каждого из наборов данных в таблице с первичным набором данных. Условия можно либо ввести непосредственно в ячейку, либо построить их при помощи Построителя выражений, щелкнув по значку калькулятора справа от ячейки. Например,

можно использовать геопространственные предикаты для создания условия слияния, по которому данные о преступлениях из одного набора данных помещаются в данные о районах другого набора данных. Условие слияния по умолчанию зависит от геопространственной шкалы измерения, как показано в списке ниже.

- Точка, Ломаная, Несколько точек, Мультиломаная - для них условие по умолчанию *close_to*.
- Многоугольник, Мультиполигон - для них условие по умолчанию *within*.

Дополнительную информацию об этих уровнях смотрите в разделе “Подуровни геопространственных измерений” на стр. 131.

Если набор данных содержит несколько геопространственных полей разного типа, используемое условие по умолчанию зависит от первой шкалы измерений, обнаруженной в данных, в следующем порядке по убыванию.

- Точки
- Ломаная
- Многоугольник

Примечание: Значения по умолчанию доступны только в том случае, когда во вторичной базе данных есть поле геопространственных данных.

Выражение для ранжирования

Укажите выражение для ранжирования слияния наборов данных; это выражение используется для сортировки нескольких совпадений в порядке, основанном на критерии ранжирования. Условия можно либо ввести непосредственно в ячейку, либо построить их при помощи Построителя выражений, щелкнув по значку калькулятора справа от ячейки.

Условия ранжирования по умолчанию для расстояний и областей предоставляются в Построителе выражений вместе с возможностью определения приоритета от низшего к высшему, например, лучшее совпадение для расстояния может соответствовать наименьшему значению. Примером ранжирования по расстоянию может быть наличие первичного набора данных, содержащего сведения о преступлениях и связанных с ними положениях, и всех других наборов данных, содержащих объекты с положениями; в этом случае расстояние от места преступления до объектов может использоваться как критерий ранжирования. Выражение ранжирования по умолчанию зависит от геопространственной шкалы измерений, как показано в списке ниже.

- Точка, Ломаная, Несколько точек, Мультиломаная - для них выражение по умолчанию *distance* (расстояние).
- Многоугольник, Мультиполигон - для них выражение по умолчанию *area* (площадь).

Примечание: Значения по умолчанию доступны только в том случае, когда во вторичной базе данных есть поле геопространственных данных.

Количество совпадений

Задайте количество возвращаемых совпадений на основе выражений для условий и ранжирования. Количество совпадений по умолчанию зависит от геопространственной шкалы измерений во вторичном наборе данных, как показано в списке ниже; однако можно дважды щелкнуть по ячейке и ввести собственное значение не более 100.

- Точка, Ломаная, Несколько точек, Мультиломаная - значение по умолчанию 3.
- Многоугольник, Мультиполигон - значение по умолчанию 1.
- Набор данных, не содержащий геопространственных полей, - значение по умолчанию 1.

Например, если вы конфигурируете слияние, основанное на **Условии слияния** *close_to* и **Выражении ранжирования** *distance*, первые три (ближайшие) совпадения из вторичных наборов данных для каждой записи в первичном наборе данных будут возвращаться как значения в получающемся поле списка.

Фильтрация полей с узла слияния

Узлы слияния обеспечивают удобный способ фильтрации или переименования дубликатов полей после слияния нескольких источников данных. Щелкните в диалоговом окне по вкладке **Фильтр**, чтобы выбрать опции фильтрации.

Представленные здесь опции почти идентичны опциям узла Фильтр. Однако в меню Фильтр доступны дополнительные опции, которые здесь не рассматриваются. Дополнительную информацию смотрите в разделе “Фильтрация или переименование полей” на стр. 141.

Поле. Выводит входные поля из подключенных в текущий момент источников данных.

Тег. Выводит имя (или номер) тега, связанное со ссылкой к источнику данных. Щелкните по вкладке **Входные поля**, чтобы изменить активные ссылки на узел слияния.

Узел источника. Выводит узел источника, данные которого подлежат слиянию.

Подсоединенный узел. Выводит имя для узла, подсоединенного к узлу слияния. Часто при сложном исследовании данных требуется несколько операций слияния или добавления, в которых может участвовать один и тот же узел источника. Имя подсоединенного узла обеспечивает способ, отличающий эти поля.

Фильтр. Выводит текущие соединения между входным и выходным полем. Для активных соединений выводится сплошная стрелка. Соединения с красным крестиком (X) указывают на отфильтрованные поля.

Поле. Список выходных полей после операции слияния или добавления. Дубликаты полей выводятся красным. Щелкните по полю Фильтр выше, чтобы отключить поля-дубликаты.

Просмотр текущих полей. Выберите эту опцию, чтобы просмотреть информацию о полях, выбранных для использования в качестве полей ключа.

Просмотр неиспользуемых параметров полей. Выберите эту опцию, чтобы просмотреть информацию о полях, не используемых в текущий момент.

Задание входного порядка и тегов

При помощи вкладки Входные поля в диалоговых окнах узлов слияния и добавления можно задать порядок входных данных и внести любые изменения в имя тега для каждого источника.

Теги и порядок входных наборов данных. Выберите эту опцию для слияния или добавления только полных записей.

- **Тег.** Возвращает список имен тегов для каждого источника входных данных. Имена тегов или **теги** представляют собой способ уникальной идентификации связей данных для операция слияния или добавления. Например, представьте себе воду из различных труб, объединяемую в одной точке и текущую по одной трубе. Поток данных в IBM SPSS Modeler аналогичен. Теги обеспечивают способ управления входами ("трубами") на узлы слияния или добавления, чтобы в случае резервирования или отсоединения узла связи остались бы и легко идентифицировались.

При подключении к узлам слияния или добавления дополнительных источников данных теги создаются автоматически при помощи номеров, представляющих порядок подсоединения вами узлов. Этот порядок не связан с порядком полей ни во входных, ни в выходных наборах данных. Тег по умолчанию можно изменить, введя новое имя в столбце *Тег*.

- **Узел источника.** Выводит узел источника, данные которого подлежат объединению.
- **Подсоединенный узел.** Выводит имя для узла, подсоединенного к узлу слияния или добавления. Часто при сложном исследовании данных требуется несколько операций слияния, в которых может участвовать один и тот же узел источника. Имя подсоединенного узла обеспечивает способ, отличающий эти поля.
- **Поля.** Возвращает число полей в каждом источнике данных.

Просмотреть текущие теги. Выберите эту опцию для просмотра тегов, активно используемых узлом слияния или добавления. Другими словами, текущие теги идентифицируют связи с узлом, через которые проходят данные. Текущие теги аналогичны метафорическим трубам с реально существующим потоком воды.

Просмотреть неиспользуемые параметры тегов. Выберите эту опцию для просмотра тегов или связей, ранее использовавшихся для узла слияния или добавления, на не подсоединенных в текущий момент к источнику данных. Они аналогичны пустым, но все еще исправным трубам в системе водоснабжения. Вы можете подсоединить эти "трубы" к новому источнику или удалить их. Чтобы удалить неиспользуемые теги с узла, нажмите кнопку **Очистить**. Все неиспользуемые теги будут сразу удалены.

Параметры оптимизации слияния

Система предоставляет две опции, которые могут помочь выполнить слияние данных в некоторых ситуациях более эффективно. Эти опции позволяют оптимизировать слияние, если один набор данных значительно больше других наборов данных или если данные уже отсортированы по всем или некоторым полям ключей, используемым вами для слияния.

Примечание: Выполняемая на этой вкладке оптимизация применяется только к вызову собственного узла IBM SPSS Modeler, то есть когда узел Слияние не выполняет преобразование pushback в SQL. Параметры оптимизации не влияют на генерирование SQL.

Один входной набор данных - сравнительно большой. Выберите эту опцию, чтобы указать, что один из входных наборов данных намного больше остальных. Система кэширует более мелкие наборы данных в памяти, а затем выполняет слияние, обрабатывая указанный большой набор данных без его кэширования и сортировки. Обычно вы будете использовать этот тип объединения с данными, полученными при помощи схемы звезда или аналогичного проекта, где используется большая центральная таблица совместно используемых данных (например, в транзакционных данных). Если вы выбрали эту опцию, нажмите кнопку **Выбрать**, чтобы указать большой набор данных. Имейте в виду, что можно выбрать всего *один* большой набор данных. Следующая таблица содержит сводку объединений, которые можно оптимизировать при помощи этого метода.

Таблица 15. Сводка оптимизаций объединений.

Тип объединения	Можно оптимизировать для большого входного набора данных?
Внутренняя	Да
Частично	Да, если в этом большом наборе данных нет неполных записей.
Заполнено	Нет
Антиобъединение	Да, если указанный большой набор данных - первый входной набор.

Все входные данные уже отсортированы по полям ключей. Выберите эту опцию, чтобы указать, что входные данные уже отсортированы по одному или нескольким полям ключей, используемым вами для слияния. Убедитесь, что отсортированы *все* входные наборы данных.

Задать существующий порядок сортировки. Укажите поля, которые уже были отсортированы. В диалоговом окне Выбрать поля добавьте поля в список. Можно выбрать только поля, используемые для слияния (заданные на вкладке Слить). В столбце *Упорядочить* укажите, сортируется ли каждое поле по возрастанию или убыванию. Если задается несколько полей, убедитесь, что вы указываете их в списке в правильном порядке сортировки. При помощи кнопок со стрелками справа от списка расставьте поля в правильном порядке. Если при указании верного существующего порядка сортировки вы сделаете ошибку, при вызове потока появится сообщение об ошибке с номером записи, где сортировка не соответствует заданной вами.

В зависимости от учета регистра символов в методе упорядочения, используемом базой данных, оптимизация может работать неверно там, где одно или несколько входных полей будут отсортированы базой данных. Например, при наличии двух входных полей, в одном из которых регистр символов учитывается, а в другом не учитывается, результаты сортировки могут отличаться. Оптимизация слияния

приводит к обработке записей, при которой используется их порядок сортировки. Вследствие этого, если для сортировки входных полей используются разные методы упорядочения, узел слияния сообщает об ошибке и выводит номер записи, где сортировка несогласована. Если все входные поля - из одного источника либо отсортированы при помощи взаимонепротиворечивых методов упорядочения, записи могут быть успешно слиты.

Примечание: Скорость слияния может выиграть от включения поддержки параллельной обработки.

Узел добавления

Узлы добавления можно использовать для выполнения конкатенации наборов записей. В отличие от узлов слияния, выполняющих объединение записей из различных источников, узлы добавления читают и передают вниз по потоку все записи из одного источника, пока не останется ни одной записи. Затем читаются записи из следующего источника, причем используется та же структура данных (число записей, число полей и так далее), что и у первого или первичного входного источника. Если у первичного источника больше полей, чем у другого входного источника, для всех недостающих значений используется пустая строка (`$null$`).

Узлы добавления полезны для создания сочетаний наборов данных со схожими структурами, но различными данными. Например, у вас могут быть данные транзакций, хранимые в разных файлах за разные периоды времени, допустим, в файле данных продаж за март и отдельном файле за апрель. Предположив, что у них одна и та же структура (одни и те же поля в одном и том же порядке), узел добавления объединит их вместе в один большой файл, который затем вы сможете проанализировать.

Примечание: Чтобы можно было добавить файлы, шкалы измерений полей должны быть однотипны. Например, *номинальное* поле нельзя добавлять к полю, у шкалы измерений которого *непрерывный* тип.

Задание опций узла добавления

Сопоставлять поля по. Выберите метод, используемый при добавлении сопоставляемых полей.

- **Положению.** Выберите эту опцию для добавления наборов данных на основе положения полей в главном источнике данных. При использовании этого метода данные должны быть отсортированы для гарантии правильного добавления.
- **Имя.** Выберите эту опцию для добавления наборов данных на основе имен полей во входных наборах данных. Кроме того, выберите **Учитывать регистр**, чтобы включить учет регистра при сопоставлении имен полей.

Поле вывода. Возвращает список узлов источников, соединенных с узлом добавления. Первый узел в списке соответствует первичному входному источнику. Поля в выводе на экран можно отсортировать, щелкнув по соответствующему заголовку столбца. Эта сортировка не изменяет порядок полей в наборе данных в действительности.

Включить поля из. Выберите **Только главный набор данных**, чтобы сгенерировать выходные поля на основе полей в главном наборе данных. Главный набор данных - это первое входное поле, задаваемое на вкладке Входные поля. Выберите **Все наборы данных**, чтобы сгенерировать выходные поля для всех полей во всех наборах данных независимо от того, существует ли сопоставляемое поле во всех входных наборах данных.

Записи тегов путем включения в поле исходного набора данных. Выберите эту опцию, чтобы добавить дополнительное поля в выходной файл, значения которого указывают исходный набор данных для каждой записи. Введите имя в текстовом поле. Имя поля по умолчанию - *Входное поле*.

Отличительный узел

Чтобы можно было начать исследование данных, сначала нужно удалить повторения записей из набора данных. Например, в базе данных маркетинговых исследований отдельные лица могут быть представлены несколькими позициями с различной информацией об адресе или компании. Отличительный узел позволяет найти повторения записей в наборе данных или удалить их оттуда либо создать из группы дублированных записей одну составную запись.

Чтобы использовать отличительный узел, сначала нужно определить набор ключевых полей, по которым распознается, когда две записи следует считать дубликатами.

Если вы не выбираете все ваши поля как ключевые, две "дублированные" записи могут оказаться на самом деле не идентичными, так как они все еще будут отличаться по значениям в остальных полях. В этом случае можно определить также порядок сортировки, применимый в каждой группе дублированных записей. Этот порядок сортировки предоставляет вам возможность точного управления, какая из записей будет считаться первой в группе. В противном случае все дубликаты будут рассматриваться как взаимозаменяемые, и может быть выбрана любая из записей. Входящий порядок записей не учитывается, поэтому использование расположенного выше узла сортировки не поможет (смотрите ниже тему "Сортировка записей на отличительном узле").

Режим. Укажите, следует ли создать составную запись либо включить или исключить (отбросить) первую запись.

- **Создать для каждой группы составную запись.** Предоставляет способ агрегирования нечисловых полей. Выбор этой опции делает доступной вкладку Составные, на которой вы указываете, как создавать составные записи. Дополнительную информацию смотрите в разделе “Особые составные параметры” на стр. 91.
- **Включить только первую запись в каждой группе.** Выбирает первую запись из каждой группы дубликатов и отбрасывает остальные. *Первая* запись определяется в соответствии с определенным ниже порядком сортировки, а не по порядку входящих записей.
- **Исключить только первую запись в каждой группе.** Отбрасывает первую запись из каждой группы дублированных записей и выбирает вместо нее оставшуюся. *Первая* запись определяется в соответствии с определенным ниже порядком сортировки, а не по порядку входящих записей. Эта опция полезна для *обнаружения* повторений в данных для возможности последующего их исследования в потоке.

Ключевые поля для группировки. Возвращает поле или список полей, используемых для установления идентичности записей. Вы можете:

- Добавить в этот список поля при помощи кнопки инструмента выбора полей в правой части окна.
- Удалить поля из списка при помощи кнопки удаления с красным крестиком (X).

В группах сортировать записи по. Перечисляет поля, используемые для определения, как сортируются записи в каждой группе дубликатов и в каком порядке они сортируются, в возрастающем или в убывающем. Вы можете:

- Добавить в этот список поля при помощи кнопки инструмента выбора полей в правой части окна.
- Удалить поля из списка при помощи кнопки удаления с красным крестиком (X).
- Переместить поля при помощи кнопок со стрелками вверх или вниз в случае их сортировки по нескольким полям.

Если вы выбрали включение или исключение первой записи из каждой группы, необходимо указать порядок сортировки, так как вам нужно знать, какая из записей рассматривается как первая.

Для некоторых опций на вкладке Составные вам может также потребоваться задать порядок сортировки в случае выбора создания составной записи. Дополнительную информацию смотрите в разделе “Особые составные параметры” на стр. 91.

Порядок сортировки по умолчанию. Укажите, как следует по умолчанию сортировать записи, по **По возрастанию** или **По убыванию** значений ключей сортировки.

Сортировка записей на отличительном узле

Если для вас важен порядок записей в группе дубликатов, необходимо задать этот порядок с помощью опции **Сортировать записи в группах по** на отличительном узле. Не используйте данные расположенного выше узла сортировки. Помните, что входящий порядок записей не учитывается, берется в расчет только заданный на узле порядок.

Если вы не указываете поля сортировки (или указываете не все из них), записи в каждой группе дубликатов будут неупорядоченными (или не полностью упорядоченными), и результат может быть непредсказуемым.

Допустим, например, что у вас есть очень большой набор записей журнала, относящихся к нескольким компьютерам. Журнал может содержать примерно такие данные:

Таблица 16. Данные журнала компьютера

Метка даты/времени	Компьютер	Температура
17:00:22	Компьютер A	31
13:11:30	Компьютер B	26
16:49:59	Компьютер A	30
18:06:30	Компьютер X	32
16:17:33	Компьютер A	29
19:59:04	Компьютер C	35
19:20:55	Компьютер Y	34
15:36:14	Компьютер X	28
12:30:41	Компьютер Y	25
14:45:49	Компьютер C	27
19:42:00	Компьютер B	34
20:51:09	Компьютер Y	36
19:07:23	Компьютер X	33

Чтобы уменьшить количество записей и оставить самую последнюю запись для каждого компьютера, используйте в качестве ключевого поле `Компьютер`, а как поле сортировки - поле `Отметка времени` (в порядке убывания). Порядок ввода на результате не отразится, поскольку выбор сортировки задает, какие из множества строк для данного компьютера должны быть возвращены, и окончательный вывод данных будет следующим.

Таблица 17. Сортированные данные журнала компьютера

Метка даты/времени	Компьютер	Температура
17:00:22	Компьютер A	31
19:42:00	Компьютер B	34
19:59:04	Компьютер C	35
19:07:23	Компьютер X	33
20:51:09	Компьютер Y	36

Отличительные параметры оптимизации

Если данные, с которыми вы работаете, содержат лишь малое число записей или если они уже были отсортированы, вы можете оптимизировать способ их обработки, включив для IBM SPSS Modeler поддержку более эффективной обработки данных.

Примечание: Если выбрана опция **У входного набора данных мало отличительных ключей** или используется генерирование SQL для узла, в значении отличительного ключа может быть возвращена любая строка; для управления возвращением строк в отличительном ключе нужно задать порядок сортировки при помощи полей **В группах сортировать записи по** на вкладке Параметры. Опции оптимизации не влияют на вывод результатов отличительным узлом, пока вы не зададите порядок сортировки на вкладке Параметры.

У входного набора данных мало отличительных ключей. Выберите эту опцию при наличии небольшого числа записей или/и небольшого числа уникальных значений полей ключей. Выбрав эту опцию, можно увеличить производительность.

Входной набор данных уже упорядочен по полям группировки и сортировки на вкладке Параметры. Выбирайте эту опцию, только если данные уже отсортированы по всем полям, указанным в списке **В группах сортировать записи по** на вкладке Параметры, и если порядок сортировки данных по возрастанию или по убыванию один и тот же. Выбрав эту опцию, можно увеличить производительность.

Отключить построение SQL. Выберите эту опцию, чтобы отключить генерирование SQL для узла.

Особые составные параметры

Если данные, с которыми вы работаете, содержат множество записей (например, для одного и того же сотрудника), вы можете оптимизировать способ их обработки, создав для процесса одну составную запись (или запись агрегации). Если установлена возможность IBM SPSS Modeler Entity Analytics, с ее помощью можно также объединять или сглаживать дублированные записи в выводе SPSS Entity Analytics.

Примечание: Эта вкладка доступна, только если на вкладке Параметры выбрано действие **Создать для каждой группы составную запись**.

Допустим, SPSS Entity Analytics помечает три записи как представляющие собой один и тот же объект, как показано в следующей таблице.

Таблица 18. Пример нескольких записей, относящихся к одному и тому же объекту.

$EA-ID	Имя	Возраст	Банк	Высшее образование	Общая сумма задолженности
0003	Bob Jones	27	K	Институт	27000
0003	Robert Jones	35	C	Степень	42000
0003	Robbie Jones	27	D	PhD	7000

Наша цель- агрегировать эти три записи в одну запись, которую мы затем будем использовать вниз по потоку. Мы могли бы при помощи узла агрегации сложить общую сумму задолженности и вычислить средний возраст; однако мы не сможем усреднить такие подробности, как имена, банки и тому подобное. Если мы укажем, какие подробности использовать для создания составной записи, то сможем получить одну запись.

Из нашей таблицы мы можем создать составную запись, выбрав следующие подробности.

- Для поля **Имя** использовать первую запись
- Для поля **Возраст** взять наибольшую
- Для поля **Банк** конкатенировать все значения без разделителя
- Для поля **Высшее образование** взять первую найденную в списке (PhD, Степень, Институт)

- Для **задолженности** взять общую сумму

Объединением (или агрегированием) этих значений мы завершаем работу, получая одну составную запись, содержащую следующие подробности.

- Имя: Bob Jones
- Возраст: 35
- Банк: KND
- Высшее образование: PhD
- Задолженность: 76000

Эта запись дает нам лучшее представление о Бобе Джонсе, сотруднике с докторской степенью, возрастом как минимум 35 лет, с тремя известными банковскими счетами и большой общей задолженностью.

Задание опций для вкладки Составные

Поле. В этом столбце выводятся все поля (кроме полей ключей в модели данных) в их естественном порядке сортировки; если узел не соединен, никакие поля не выводятся. Чтобы отсортировать строки по именам полей в алфавитном порядке, щелкните по соответствующему заголовку столбца. Несколько строк можно выбрать щелчком мыши с нажатой клавишей Shift или Ctrl. Кроме того, при щелчке по полю правой кнопкой мыши появляется меню, в котором можно выбрать следующие действия: выбрать все строки, отсортировать строки по возрастанию или убыванию имен или значений полей, выбрать поля по типу показателей или значений или выбрать значение для автоматического добавления одной и той же записи **Заполнить значениями на основе** в каждую выбираемую строку.

Заполнить значениями на основе. Выберите тип значений, который должен использоваться для составной записи для столбца **Поле**. Доступные опции зависят от типа поля.

- Для полей числового диапазона можно выбрать:
  - Первая запись в группе
  - Последняя запись в группе
  - Всего
  - Среднее значение
  - Минимум
  - Максимум
  - Пользовательское:
- Для полей времени или даты можно выбрать:
  - Первая запись в группе
  - Последняя запись в группе
  - Самые ранние
  - Самые последние
  - Пользовательское:
- Для строковых полей или полей без типа можно выбрать:
  - Первая запись в группе
  - Последняя запись в группе
  - Первую по алфавиту
  - Последнюю по алфавиту
  - Пользовательское:

В любом случае опция **Пользовательское** позволяет реализовать дополнительное управление значением, используемым для заполнения составной записи. Дополнительную информацию смотрите в разделе “Составные особого типа - Вкладка Пользовательские” на стр. 93.

Включить количество записей в поле. Выберите эту опцию, чтобы включить в каждую выходную запись избыточное поле, по умолчанию называемое Record_Count. Это поле указывает, сколько записей было агрегировано для образования каждой записи агрегации. Чтобы создать для этого поля пользовательское имя, введите это имя в поле редактирования.

Составные особого типа - Вкладка Пользовательские

Диалоговое окно Пользовательское заполнение предоставляет дополнительный элемент управления, позволяющий использовать значение для заполнения новой составной записи. Имейте в виду, что перед использованием этой опции ваши данные нужно инстанцировать, если на вкладке Составные выполняется настройка только одной строки данных.

Примечание: Это диалоговое окно становится доступно только при выборе пользовательского значения в столбце **Заполнить значениями на основе** на вкладке Составные.

В зависимости от типа поля, можно выбрать одну из следующих опций.

- **Выбрать по частоте.** Выберите значение на основе частоты его встречаемости в записи данных.

  Примечание: Для полей непрерывного типа, полей без типа и полей даты/времени эта опция недоступна.
  - **Использование.** Выберите наименее или наиболее часто встречающееся.
  - **Совпадающие наблюдения.** При наличии нескольких записей с одинаковой частотой встречаемости укажите, как выбирать требуемую запись. Можно выбрать одну из четырех опций: Использовать первую, Использовать последнюю, Использовать наименьшую или Использовать наибольшую.
- **Включает значение (T/F).** Выберите эту опцию для преобразования поля в флаг, указывающий, есть ли у какой-либо из записей в группе заданное значение. Затем можно выбрать значение в списке **Значение** для выбранного поля.

  Примечание: В случае выбора нескольких строк полей на вкладке Составные эта опция недоступна.
- **Первое совпадение в списке.** Выберите эту опцию, чтобы определить, какое значение в первую очередь присваивать составной записи. Затем можно выбрать один элемент в списке **Элементы** для выбранного поля.

  Примечание: Если вы выберете несколько строк полей на вкладке Составные, эта опция будет недоступна.
- **Конкатенировать значения.** Выберите эту опцию, чтобы сохранять все значения в группе, конкатенируя их в строку. Нужно указать, какой использовать разделитель между значениями.

  Примечание: Это единственная опция, доступная, если выбрана одна или несколько строк полей непрерывного типа, полей без типа или полей даты/времени.
- **Использовать разделитель.** Для использования в конкатенированной строке в качестве значения разделителя можно выбрать **пробел** или **Запятую**. Другой вариант - ввести в поле **Другой** свой собственный символ для значения разделителя.

  Примечание: Доступно, только если выбрана опция**Конкатенировать значения**.

узел потоковых временных рядов

Узел потоковых временных рядов можно использовать для построения и скоринга моделей временных рядов за один шаг. Отдельная модель временного ряда строится для каждого поля назначения, однако слепки моделей не добавляются на палитру сгенерированных моделей, и информацию о моделях невозможно просмотреть.

Методам для моделирования данных временных рядов требуется универсальный интервал между измерениями, со всеми пропущенными значениями, указанными пустыми строками. Если данные еще не отвечают этому требованию, вам надо преобразовать значения должным образом.

Другие моменты, которые следует отметить в связи с узлами временных рядов:

- Поля должны быть числовыми.
- Поля дат нельзя использовать в качестве входных.
- Разделы игнорируются.

Узел потоковых временных рядов оценивает модели экспоненциального сглаживания, одномерные и многомерные модели авторегрессии и проинтегрированного скользящего среднего, модели АРПСС (Autoregressive Integrated Moving Average, ARIMA) (или передаточных функций) для временных рядов и составляет прогнозы на основе данных временных рядов. Доступен также эксперт построения моделей, который пытается автоматически идентифицировать и оценить наиболее подходящую модель АРПСС или экспоненциального сглаживания для одного или нескольких полей назначения.

Дополнительную информацию о моделировании временных рядов смотрите в разделе Модели временных рядов в руководстве Узлы моделирования SPSS Modeler.

Узел потоковых временных рядов поддерживается для использования в среде внедрения потоков через IBM SPSS Modeler Solution Publisher с использованием службы скоринга IBM SPSS Collaboration and Deployment Services или IBM InfoSphere Warehouse.

Узел потоковых временных рядов - опции полей

На вкладке Поля указывается, будут ли использоваться значения ролей полей, уже определенные на расположенных выше узлах, или назначение полей будет выполнено вручную.

Использовать предопределенные роли Эта опция использует значения ролей (назначения, предикторы и так далее) с расположенного выше узла Тип (или с вкладки Типы лежащего выше узла источника).

Использовать пользовательские назначения полей Выберите эту опцию, чтобы назначить вручную объекты назначения, предикторы и другие роли.

Поля При помощи кнопок со стрелками назначьте элементы из этого списка вручную для различных полей ролей в правой части экрана. Значки обозначают допустимые уровни измерения для каждого поля роли.

Чтобы выбрать все поля в списке, нажмите кнопку **Все**, или же, чтобы выбрать все поля с этим уровнем измерения, нажмите кнопку для отдельного уровня измерения.

Назначение Выберите одно поле в качестве назначения для предсказания.

Входные поля - кандидаты Выберите одно или несколько полей как входные поля для предсказания.

События и вмешательства Используйте эту область, чтобы обозначить определенные входные поля как поля событий или вмешательств. Это указание определяет поле как содержащее данные временных рядов, которые могут затрагивать события (предсказуемые повторяющиеся ситуации, такие как маркетинговые акции) или вмешательствами (одноразовыми инцидентами, такими как отключение электроэнергии или забастовка сотрудников).

Узел потоковых временных рядов - опции спецификации данных

На вкладке Спецификации данных задаются все опции для данных, подлежащих включению в вашу модель. Можно, конечно, просто нажать кнопку **Выполнить**, чтобы построить модель со всеми опциями по умолчанию, но скорее всего вы захотите настроить конструкцию для своих собственных целей.

Вкладка содержит несколько различных панелей, где можно задать настройки, относящиеся конкретно к вашей модели.

Узел потоковых временных рядов - наблюдения

Используйте эти параметры, чтобы указать поля, определяющие наблюдения.

Наблюдения, которые задаются полем даты/времени

Можно задать, что наблюдения определяются полем даты, времени или отметки времени. В дополнение к полю, определяющему наблюдения, выберите подходящий интервал времени, описывающий это наблюдение. В зависимости от указанного интервала времени можно задать также другие параметры, такие как интервал между наблюдениями (инкремент) или число дней в неделю. К интервалу времени применяются следующие возможности:

- Используйте значение **Нерегулярный**, когда наблюдения нерегулярно распределены по времени, например, задаются в моменты времени обработки заказов на продажу. Когда выбрана опция **Нерегулярный**, нужно указать интервал времени анализа, используя параметры **Интервал времени** на вкладке Спецификации данных.
- Когда наблюдения представляют дату и время, а интервал времени - это Часы, Минуты или Секунды, используйте параметры **Часов в день**, **Минут в день** или **Секунд в день**. Когда наблюдение представляет время (длительность)без указания на дату и интервал времени в часах, минутах или секундах, используйте параметры **Часы (не периодически)**, **Минуты (не периодически)** или **Секунды (не периодически)**.
- На основании выбранного интервала времени процедура может обнаружить пропущенные наблюдения. Обнаружение пропущенных наблюдений необходимо, так как процедура предполагает равномерное распределение всех наблюдений во времени через равные промежутки и отсутствие пропущенных наблюдений. Например, если интервал времени - это Дни, а после даты 2015-10-27 следует 2015-10-29, это указывает на пропущенное наблюдение для 2015-10-28. Для всех пропущенных наблюдений импутируются значения; с помощью области **Обработка пропущенных значений** вкладки Спецификации данных задайте параметры для обработки пропущенных значений.
- Заданный интервал времени позволяет процедуре обнаружить несколько наблюдений в одном интервале времени, которые нужно объединить вместе, и выровнять наблюдения по границе интервала, например по первому числу месяца, для обеспечения эквидистантности наблюдений. Например, если интервал времени - это Месяцы, несколько дат одного месяца объединяются вместе. Такой тип объединения называется *группировка*. По умолчанию наблюдения при группировке суммируются. Для группировки можно задать и другой способ обработки значений, такой как вычисление среднего, используя параметры **Агрегирование и распределение** на вкладке Спецификации данных.
- Для некоторых интервалов времени дополнительные параметры могут определить разрывы в обычно равномерно расположенных интервалах. Например, если интервал времени - это Дни, но допускаются только рабочие дни, можно указать, что в неделе есть только пять дней и она начинается в понедельник.

Наблюдения определяются как периоды или циклические периоды

Наблюдения можно определять одним или несколькими целочисленными полями, представляющими периоды или повторяющиеся циклы периодов вплоть до произвольного числа уровней циклов. При помощи такой структуры можно описать ряд наблюдений, который не подходит ни под один из обычных интервалов времени. Например, финансовый год всего с 10 месяцами можно описать полем цикла, представляющим годы, и полем периода, представляющим месяцы, где длина одного цикла равна 10.

Поля, задающие циклические периоды, определяют иерархию периодических уровней, где низший уровень определен полем **Период**. Следующий высший уровень задается циклическим полем с уровнем 1, после него - циклическим полем с уровнем 2, и т.д. Значения полей для каждого уровня, кроме наивысшего, должны быть периодическими по отношению к следующему высшему уровню. Значения для наивысшего уровня не могут быть периодическими. Например, в случае 10-месячного финансового года месяцы периодически повторяются по годам, но годы не периодические.

- Длина цикла на конкретном уровне - это период следующего низшего уровня. Для примера с финансовым годом есть только один циклический уровень и длина его цикла равна 10, так как следующий низший уровень представляет месяцы, а в каждом заданном финансовом году 10 месяцев.
- Укажите начальное значение для любого периодического поля, не начинающегося с 1. Этот параметр необходим для обнаружения пропущенных значений. Например, если периодическое поле начинается с 2, но начальное значение задано как 1, процедура предположит, что существует пропущенное значение для первого периода в каждом цикле этого поля.

Узел потоковых временных рядов - интервал времени для анализа

Интервал времени, используемый вами для анализа, может отличаться от интервала времени наблюдения. Например, если интервал времени наблюдений - это Дни, для анализа вы можете выбрать интервал времени Месяцы. Затем данные агрегируются из ежедневных в ежемесячные, прежде чем строится модель. Вы можете распределить данные также из более длительного в короткий интервал времени. Например, если наблюдения квартальные, данные можно распределить из ежеквартальных в ежемесячные.

Используйте эти параметры, чтобы указать интервал времени для анализа. Способ, которым объединяются или распределяются данные, задается параметрами **Агрегирование и распределение** на вкладке Спецификации данных.

Возможные варианты выбора интервала времени анализа зависят от того, как определены наблюдения, и от интервала времени этих наблюдений. В частности, когда наблюдения определены по циклическим периодам, поддерживается только агрегирование. В этом случае интервал времени анализа должен быть не меньше, чем интервал времени наблюдений.

Узел потоковых временных рядов - опции агрегации и распределения

С помощью параметров на этой панели задаются значения параметров для агрегирования или распределения входных данных в соответствии с интервалами времени наблюдений.

Функции агрегации

Когда используемый для анализа интервал времени превышает временной интервал наблюдений, производится агрегирование входных данных. Например, агрегирование применяется, если интервал наблюдений - это Дни, а интервал анализа - Месяцы. Доступны следующие функции агрегирования: mean, sum, mode, min или max.

Функции распределения

Когда используемый для анализа интервал короче интервала наблюдений, входные данные распределяются. Например, распределение производится, если интервал наблюдений - это Кварталы, а интервал анализа - Месяцы. Доступны следующие функции распределений: mean или sum.

Функции группировки

Группировки применяются, когда наблюдения определяются значениями дата/время и в одном интервале времени производится несколько наблюдений. Например, если временной интервал наблюдений - это Месяцы, несколько дат одного месяца можно сгруппировать и связывать с месяцем, которому они принадлежат. Доступны следующие функции группировки: mean, sum, mode, min или max. Группировка производится всегда, когда наблюдения определены значениями даты/времени, а временной интервал наблюдений задан как Нерегулярный.

Примечание: Хотя группировка - это вариант агрегирования, она производится до любой обработки пропущенных значений, а формальное агрегирование выполняется после такой обработки. Когда временной интервал наблюдений задан как нерегулярный, агрегирование производится только через функцию группировки.

Объединить межсуточные наблюдения на вчерашний день

Указывает, агрегировать ли наблюдения, время которых перешло за границу дня, со значениями за предыдущий день. Например, для почасовых измерений в течение суток, разделенных на интервалы по 8 часов, начинающихся в 20:00, этот параметр указывает, включать ли наблюдения от 00:00 до

04:00 в агрегированные результаты предыдущего дня. Этот параметр применим только в том случае, когда интервал времени наблюдений - это Часов в день, Минут в день или Секунд в день, а временной интервал анализа - Дни.

Пользовательские параметры для заданных полей

Функции агрегирования, распределения и группировки можно задать, используя поле за полем. Эти параметры переопределят параметры по умолчанию для функций агрегирования, распределения и группировки.

Узел потоковых временных рядов - опции отсутствующих значений

С помощью параметров на этой панели задается способ замены пропущенных значений во входных данных на импутированное значение. Возможны следующие способы замены:

Линейная интерполяция

Заменяет пропущенные значения с помощью линейной интерполяции. Для интерполяции используются последнее валидное (непропущенное) значение перед пропущенным и первое валидное значение после пропущенного. Если в первом или последнем наблюдении ряда есть пропущенное значение, используются два ближайших непропущенных значения в начале или в конце ряда.

Среднее значение ряда

Заменяет пропущенные значения средним для всего ряда.

Среднее близких точек

Заменяет пропущенные значения средним из валидных окружающих значений. Интервал ближайших точек здесь - количество точек до и после текущей, которые используются при вычислении среднего.

Медиана близких точек

Заменяет пропущенные значения медианой из валидных окружающих значений. Интервал ближайших точек здесь - количество точек до и после текущей, которые используются при вычислении медианы.

Линейный тренд

Эта опция использует все непропущенные наблюдения в ряду для подгонки по простой модели линейной регрессии, которая затем используется для импутации пропущенных значений.

Другие параметры:

Максимальная процентная доля пропущенных значений (%)

Задает максимальную процентную долю пропущенных значений, разрешенных для любого ряда. Ряды с большим количеством пропущенных значений, чем задано этим пределом, исключаются из анализа.

Узел потоковых временных рядов - интервал оценки

На вкладке Период оценивания можно задать диапазон записей для использования при оценивании модели. По умолчанию период оценивания начинается с самого раннего наблюдения и завершается самым поздним наблюдением в ряду.

По начальному и конечному времени

Можно задать и начальное, и конечное время периода оценивания, а также указать только начало или только конец. Если не задается начальное или конечное время периода оценивания, используется значение по умолчанию.

- Если наблюдения определяются полем дата/время, введите значения начального и конечного времени в том же формате, который используется для поля даты/времени.
- Для наблюдений, определенных циклическими интервалами, задайте значение для каждого из полей циклических интервалов. Каждое поле выводится в отдельном столбце.

По самым поздним и самым ранним интервалам времени

Определяет период оценивания как заданное количество интервалов времени, начинающихся самым ранним интервалом или заканчивающихся самым поздним интервалом в данных с учетом возможного сдвига. В этом контексте интервал времени относится к интервалу времени анализа.

Допустим, например, что наблюдения ежемесячные, но интервал анализа квартальный. Если задать опцию **Последний** и значение 24 для параметра **Число интервалов**, это будет означать последние 24 квартала.

Если хотите, можно исключить указанное число интервалов времени. Например, указание последних 24 интервалов времени и значения 1 для исключаемого количества означает, что период оценивания состоит из 24 интервалов, предшествующих последнему интервалу.

Узел потоковых временных рядов - опции построения

На вкладке Параметры конструкции задаются все опции для построения вашей модели. Можно, конечно, просто нажать кнопку **Выполнить**, чтобы построить модель со всеми опциями по умолчанию, но скорее всего вы захотите настроить конструкцию для своих собственных целей.

Вкладка содержит две различных панели, на которых можно задать настройки, относящиеся конкретно к вашей модели.

Узел потоковых временных рядов - общие опции построения

Опции, доступные на этой панели, зависят от того, какой из следующих трех параметров вы выберите в списке **Метод**:

- **Эксперт построения моделей** Выберите эту опцию, чтобы использовать эксперт построения моделей, который автоматически находит наиболее подходящую модель для каждого зависимого ряда.
- **Экспоненциальное сглаживание** Используйте эту опцию для задания пользовательской модели экспоненциального сглаживания.
- **ARIMA** Используйте эту опцию для задания пользовательской модели АРПСС.

Эксперт построения моделей

Тип моделей Выберите тип моделей, которые вы хотите построить.

- **Все модели** Эксперт создания моделей рассматривает и модели АРПСС, и модели экспоненциального сглаживания.
- **Только модели экспоненциального сглаживания** Эксперт создания моделей рассматривает только модели экспоненциального сглаживания.
- **Только модели АРПСС** Эксперт создания моделей рассматривает только модели АРПСС.

Эксперт создания моделей рассматривает сезонные модели Эта опция включается только в том случае, если для активного набора данных включена периодичность. Когда выбрана эта опция, эксперт построения моделей рассматривает и сезонные, и несезонные модели. Если эта опция не выбрана, эксперт создания моделей рассматривает только несезонные модели.

Автоматически обнаруживать выбросы По умолчанию автоматическое обнаружение выбросов не производится. Выберите эту опцию, чтобы выполнять автоматическое обнаружение выбросов, после чего выберите нужные типы выбросов.

У входных полей должна быть шкала измерений *Флаговая*, *Номинальная* или *Порядковая*, и они должны быть числовыми (например, для флагового поля должно быть указано, 1/0, а не True/False); только тогда они включаются в этот список.

Эксперт построения моделей рассматривает только простую регрессию и не рассматривает произвольные передаточные функции для входных полей, определенных как поля событий или вмешательств на вкладке **Поля**.

Экспоненциальное сглаживание

Тип моделей Модели экспоненциального сглаживания классифицируются как сезонные или несезонные.[1] Сезонные модели доступны, только если периодичность, определенная при помощи панели интервалов времени на вкладке Спецификации данных, сезонная. Типы сезонной периодичности: периоды циклов, года, кварталы, месяцы, дни в неделю, часы в день, минуты в день и секунды в день. Доступны следующие типы моделей:

- **Простая** Эта модель подходит для рядов, в которых отсутствует тренд и сезонность. Единственный релевантный параметр такой модели предназначен для сглаживания уровня ряда. Простая модель экспоненциального сглаживания в наибольшей степени напоминает модель АРПСС с нулевым порядком авторегрессии, единичными порядками дифференцирования и скользящего среднего, и не имеющую константы.
- **Линейный тренд Хольта** Эта модель подходит для рядов, в которых имеется линейный тренд и отсутствует сезонность. Относящиеся к ней параметры предназначены для сглаживания уровня и тренда, независимого в этой модели. Модель экспоненциального сглаживания Хольта является более общей, чем модель Брауна, но вычисление оценок для нее может занять больше времени в случае длинных рядов. Модель экспоненциального сглаживания Хольта в наибольшей степени напоминает модель АРПСС с нулевым порядком авторегрессии и двумя порядками дифференцирования и скользящего среднего.
- **Линейный тренд Брауна** Эта модель подходит для рядов, в которых имеется линейный тренд и отсутствует сезонность. Релевантными сглаживающими параметрами для нее являются уровень и тренд, но в данной модели они предполагаются равными. Поэтому модель Брауна представляет собой частный случай модели Холта. Модель экспоненциального сглаживания Брауна в наибольшей степени напоминает модель АРПСС с нулевым порядком авторегрессии, порядками дифференцирования и скользящего среднего, равными двум, для которой коэффициент скользящего среднего для второго лага равен квадрату половины коэффициента для первого лага в квадрате.
- **Демпфированный тренд** Эта модель подходит для рядов, в которых линейный тренд затухает, а сезонность отсутствует. Ее релевантные параметры предназначены для сглаживания уровня, тренда и скорости затухания тренда. Затухающая модель экспоненциального сглаживания в наибольшей степени напоминает модель АРПСС с единичными порядками авторегрессии и дифференцирования, имеющую порядок скользящего среднего, равный двум.
- **Простая сезонная** Эта модель подходит для ряда, в котором нет никакого тренда, а сезонная вариация постоянна во времени. Ее релевантные параметры сглаживания - уровень и сезонная составляющая. Сезонная модель экспоненциального сглаживания в наибольшей степени напоминает модель АРПСС с нулевым порядком авторегрессии (единичными порядками несезонного и сезонного дифференцирования) и ненулевые коэффициенты скользящего среднего для лагов 1, p и p+1, где p - число периодов сезонности. Для ежемесячных данных p = 12.
- **Аддитивная Уинтерса** Эта модель подходит для ряда с линейным трендом и сезонной вариацией, не меняющейся с течением времени. Ее релевантные параметры сглаживания - уровень, тренд и сезонная составляющая. Сезонная аддитивная модель Винтера в наибольшей степени напоминает модель АРПСС с нулевым порядком авторегрессии (единичными порядками несезонного и сезонного дифференцирования) и ненулевые коэффициенты скользящего среднего для лагов p, где p - число периодов сезонности. Для ежемесячных данных p = 12.
- **Мультипликативная Уинтерса** Эта модель подходит для ряда с линейным трендом и сезонной вариацией, изменяющейся с величиной ряда. Ее релевантные параметры сглаживания - уровень, тренд и сезонная составляющая. Мультипликативная модель экспоненциального сглаживания Винтерса не похожа ни на одну из моделей АРПСС.

Преобразование назначения Можно задать преобразование, выполняемое для каждой зависимой переменной перед ее моделированием.

- **Нет** Преобразование не выполняется.
- **Квадратный корень** Выполняется преобразование Квадратный корень.

---

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

- **Натуральный логарифм** Выполняется преобразование Натуральный логарифм.

АРПСС

Задайте структуру пользовательской модели АРПСС.

Порядок АРПСС В соответствующих ячейках сетки введите значения для различных компонентов АРПСС вашей модели. Все значения должны быть неотрицательными целыми числами. Для авторегрессии и компонентов скользящего среднего такое значение представляет максимальный порядок. В модель включаются все положительные меньшие порядки. Например, если задать значение 2, в модель будут включены порядки 2 и 1. Ячейки в столбце Сезонность включаются только в том случае, если для активного набора данных определена периодичность.

- **Авторегрессия (p)** Количество порядков авторегрессии в модели. Порядки авторегрессии задают, какие предыдущие значения из ряда использовались для предсказания текущих значений. Например, порядок авторегрессии 2 означает, что для предсказания текущего значения использовалось на два периода более раннее значение из ряда.
- **Разностная (d)** Задает порядок исчисления разностей, применимый к ряду до оценки моделей. Вычисление разностей необходимо при наличии трендов (ряды с трендами обычно нестационарные, а моделирование АРПСС предполагает стационарность) и используется для удаления этих эффектов. Порядок исчисления разностей соответствует степени тренда ряда - разности первого порядка учитывают линейные тренды, разности второго порядка - квадратичные, и так далее.
- **Скользящее среднее (q)** Количество порядков скользящего среднего в модели. Порядки скользящего среднего задают, как отклонения от среднего значения ряда предыдущих значений используются для предсказания текущих значений. Например, порядки скользящего среднего 1 и 2 указывают, что отклонения от среднего значения ряда для каждого значения за прошлые два периода будут рассматриваться для предсказания текущих значений ряда.

Сезонная Сезонные компоненты авторегрессии, скользящего среднего и исчисления разностей играют ту же роль, что и их несезонные аналоги. Однако для сезонных порядков на текущие значения ряда влияют предыдущие значения, отделенные одним или несколькими сезонными периодами. Например, для ежемесячных данных (сезонный период 12) сезонный порядок 1 означает, что на текущее значение ряда влияет значение ряда за 12 периодов до текущего. Тем самым для ежемесячных данных сезонный порядок 1 - это то же самое, что несезонный порядок 12.

Определять выбросы автоматически Выберите эту опцию, чтобы выполнять автоматическое обнаружение выбросов, и выберите один или несколько из доступных типов выбросов.

Обнаруживаемые типы выбросов Выберите типы выбросов, обнаружение которых вы хотите выполнять. Поддерживаются следующие типы:

- Аддитивные (по умолчанию)
- Сдвиг уровня (по умолчанию)
- Инновационные
- Переходные
- Сезонные аддитивные
- Локальный тренд
- Аддитивная вставка

Порядки и преобразования передаточных функций Чтобы задать преобразования и определить передаточные функции для каких-либо или всех входных полей в модели АРПСС, нажмите кнопку **Задать**; появится отдельное диалоговое окно, в котором надо ввести подробности передачи и преобразования.

Включить константу в модель Включение константы - это стандартный прием, если вы не уверены, что общее среднее значение ряда равно 0. Рекомендуется исключить константу, если применяется исчисление разностей.

Функции передачи и преобразования: С помощью диалогового окна Порядки и преобразования передаточных функций задаются преобразования и определяются передаточные функции для каких-либо или всех полей в модели АРПСС.

Преобразования назначения На этой панели можно задать преобразование, выполняемое для каждой переменной назначения перед ее моделированием.

- **Нет** Преобразование не выполняется.
- **Квадратный корень** Выполняется преобразование Квадратный корень.
- **Натуральный логарифм** Выполняется преобразование Натуральный логарифм.

Передаточные функции входных рядов-кандидатов и преобразование С помощью передаточных функций задается способ использования прошлых значений входных полей для прогнозирования будущих значений ряда назначения. В списке в левой части панели выводятся все входные поля. Остальная информация на этой панели относится конкретно к выбираемому вами входному полю.

Порядок передаточных функций В соответствующих ячейках сетки **Структура** введите значения для различных компонентов передаточной функции. Все значения должны быть неотрицательными целыми числами. Для компонентов числителя и знаменателя это значение представляет максимальный порядок. В модель включаются все положительные меньшие порядки. Кроме этого, порядок 0 всегда включен для компонентов числителя. Например, если задать значение 2 для числителя, в модель будут включены порядки 2, 1 и 0. Если задать 3 для знаменателя, в модель будут включены порядки 3, 2 и 1. Ячейки в столбце Сезонность включаются только в том случае, если для активного набора данных определена периодичность.

Числитель Порядок числителя передаточной функции задает, какие предыдущие значения из выбранного независимого ряда (предикторы) используются для предсказания значений зависимого ряда. Например, порядок числителя 1 означает, что для предсказания текущего значения каждого зависимого ряда используется значение независимого ряда на один период раньше, а также текущее значение независимого ряда.

Знаменатель Порядок знаменателя передаточной функции задает, как отклонения от среднего значения ряда для предыдущих значений из выбранного независимого ряда (предикторов) используются для предсказания значений зависимого ряда. Например, порядок знаменателя 1 означает, что отклонения от среднего значения независимого ряда на один период в прошлом рассматриваются для предсказания текущего значения каждого зависимого ряда.

Разность Задает порядок исчисления разностей, применимый к выбранному независимому ряду (предиктору) до оценки моделей. Вычисление разностей необходимо при наличии трендов и используется для удаления их влияния.

Сезонность Сезонные числитель, знаменатель и компоненты разностей играют ту же роль, что и их несезонные аналоги. Однако для сезонных порядков на текущие значения ряда влияют предыдущие значения, отделенные одним или несколькими сезонными периодами. Например, для ежемесячных данных (сезонный период 12) сезонный порядок 1 означает, что на текущее значение ряда влияет значение ряда на 12 периодов ранее текущего. Тем самым для ежемесячных данных сезонный порядок 1 - это то же самое, что несезонный порядок 12.

Задержка Определение задержки приводит к тому, что влияние входного поля задерживается на заданное число интервалов. Например, если задана задержка 5, значение входного поля в момент времени t не влияет на прогноз, пока не пройдет пять периодов ($t + 5$).

Преобразование Спецификация передаточной функции для набора независимых переменных включает в себя также необязательное преобразование, которое будет выполнено с этими переменными.

- **Нет** Преобразование не выполняется.
- **Квадратный корень** Выполняется преобразование Квадратный корень.

- **Натуральный логарифм** Выполняется преобразование Натуральный логарифм.

Узел потоковых временных рядов - опции модели

Ширина границ доверительного интервала (%) Доверительные интервалы вычисляются для предсказаний моделей и автокорреляций остатков. Вы можете указать любое положительное число, меньшее 100. По умолчанию используется 95%-ный доверительный интервал.

Прогноз

- Опция **Распространить записи на будущее** задает число интервалов оценивания для прогнозов на время после окончания периода оценивания. Интервал времени в этом случае - это интервал времени анализа, заданный вами на вкладке Спецификации данных. Если затребованы прогнозы, выполняется автоматическое построение моделей авторегрессии для всех входных рядов, не являющихся также и назначениями. Затем эти модели используются для генерирования значения данных входных рядов на период прогноза. Максимального ограничения для этого параметра нет. for this setting.
- **Вычислить будущие значения входных данных** Если выбрать эту опцию, будут вычисляться прогнозные значения для показателей предсказания, предсказания шума, оценка дисперсии и будущие значения времени.

Сделать доступной для скоринга Здесь можно задать значения по умолчанию для опций скоринга, которые появятся в диалоговом окне для слепка модели.

- **Вычислять верхнюю и нижнюю границы доверительных интервалов** Эта опция, если она выбрана, создает новые поля (с префиксами по умолчанию $TSLCI- и $TSUCI-) для нижних и верхних доверительных интервалов, для каждого поля назначения, а также итоги этих значений.
- **Рассчитать шумовые остатки** Эта опция, если она включена, создает новое поле (с префиксом по умолчанию $TSNR-) для остатков модели, для каждого поля назначения, а также итог этих значений.

Узел потока TCM

Узел потока TCM можно использовать для построения и скоринга причинных моделей времени за один шаг.

Дополнительную информацию об этих причинных моделях времени смотрите в теме Причинные модели времени в разделе Модели временных рядов Руководства по узлам моделирования SPSS Modeler.

Узел потока TCM - Опции временных рядов

На вкладке Поля используйте параметры **Временные ряды**, чтобы задать ряды, которые будут включены в модельную систему.

Выберите опцию структуры данных, применимой к вашим данным. Для многомерных данных щелкните по **Выбрать измерения**, чтобы указать поля измерений. Порядок указанных полей измерений определяет тот порядок, в котором они появятся во всех последующих диалоговых окнах и в выходных данных. Используйте стрелки вниз и вверх в дополнительном диалоговом окне Выбрать измерения, чтобы изменить порядок полей измерений.

Для данных на основе столбцов у термина *ряд* такое же значение, как и у термина *поле*. Для многомерных данных поля, содержащие временной ряд, называются полями *показателей*. Временной ряд для многомерных данных определен полем показателей и значениями для каждого из полей измерений. Следующие возможности применимы и к данным на основе столбцов, и к многомерным данным.

- Ряды, указанные как ряды - кандидаты входных значений или как входные ряды и ряды назначения, рассматриваются для включения в модель для каждого назначения. Модель для каждого элемента назначения всегда содержит сдвиг значений относительно друг друга (лаг).
- Ряды, указанные как принудительные входные данные, всегда включаются в модель для каждого элемента назначения.

- По крайней мере один ряд должен быть задан или как ряд назначения, или как одновременно ряд назначения и входных данных.
- Когда выбрана опция **Использовать предопределенные роли**, поля с ролью Вход задаются как кандидаты входных рядов. Никакая предопределенная роль не отображается на Принудительные входные данные.

Многомерные данные

Для многомерных данных задаются поля показателей и связанные роли в решетке, где каждая строка в решетке задает один показатель и роль. По умолчанию система моделей включает в себя ряды для всех комбинаций полей измерений для каждой строки в решетке. Например, если есть измерения для переменных *регион* и *торговая_марка*, по умолчанию указание показателя *продажи* как назначения означает, что существуют отдельные ряды назначения продаж для каждой комбинации *регион* и *торговая_марка*.

Для каждой строки в решетке можно настроить набор значений для каждого из полей измерений, нажав кнопку с многоточием для измерения. Это действие откроет дополнительное диалоговое окно Выбрать значения измерения. Вы можете также добавить, удалить или скопировать строки решетки.

Столбец **Число рядов** выводит количество наборов значений измерений, которые в настоящее время заданы для связанного показателя. Выведенное значение может быть больше, чем фактическое количество рядов (один ряд на набор). Это условие возникает, когда некоторые из заданных комбинаций значений измерений не соответствуют рядам, входящим в связанный показатель.

Узел потока TCM - Опции выбора измерений

Для многомерных данных можно настроить анализ, указывая, какие значения измерений применить к конкретному полю показателей с конкретной ролью. Например, если *продажи* - это поле показателей, а *канал* - измерение со значениями 'розница' и 'интернет', можно задать продажи 'интернет' как входные данные, а продажи через 'розницу' как назначение.

Все значения

Указывает, что включаются все значения текущего поля измерения. Эта опция выбрана по умолчанию.

Выберите значения для включения или исключения

Используйте эту опцию, чтобы задать значения для текущего поля измерения. Когда выбрано значение **Включить** для опции **Режим**, включаются только значения, указанные в списке **Выбранные значения**. При выборе значения **Исключить** для опции **Режим** включаются все значения, отличные от указанных в списке **Выбранные значения**.

Набор значений, из которых производится выбор, можно отфильтровать. Значения, удовлетворяющие условиям фильтра, появятся на вкладке **Соответствуют**, а все другие значения - на вкладке **Не соответствуют** в списке **Невыбранные значения**. На вкладке **Все** перечислены все невыбранные значения, независимо от условий фильтра.

- При определении фильтра можно использовать звездочку (*) для символов подстановки.
- Чтобы очистить текущий фильтр, задайте пустое значение для термина поиска в диалоговом окне Фильтровать выводимые значения.

Узел потока TCM - Опции наблюдений

На вкладке Поля используйте параметры **Наблюдения**, чтобы указать поля, определяющие наблюдения.

Наблюдения, определенные датой/временем

Можно задать, что наблюдения определяются полем даты, времени или отметки времени. В дополнение к полю, определяющему наблюдения, выберите подходящий интервал времени, описывающий это наблюдение. В зависимости от указанного интервала времени можно задать также другие параметры, такие как интервал между наблюдениями (инкремент) или число дней в неделю. К интервалу времени применяются следующие возможности:

- Используйте значение **Нерегулярный**, когда наблюдения нерегулярно распределены по времени, например, задаются в моменты времени обработки заказов на продажу. Когда выбрана опция **Нерегулярный**, нужно указать интервал времени анализа, используя параметры **Интервал времени** на вкладке Спецификации данных.
- Когда наблюдения представляют дату и время, а интервал времени - это Часы, Минуты или Секунды, используйте параметры **Часов в день**, **Минут в день** или **Секунд в день**. Когда наблюдение представляет время (длительность)без указания на дату и интервал времени в часах, минутах или секундах, используйте параметры **Часы (не периодически)**, **Минуты (не периодически)** или **Секунды (не периодически)**.
- На основании выбранного интервала времени процедура может обнаружить пропущенные наблюдения. Обнаружение пропущенных наблюдений необходимо, так как процедура предполагает равномерное распределение всех наблюдений во времени через равные промежутки и отсутствие пропущенных наблюдений. Например, если интервал времени - это Дни, а после даты 2014-10-27 следует 2014-10-29, это указывает на пропущенное наблюдение для 2014-10-28. Для всех пропущенных наблюдений импутируются значения. Параметры для обработки пропущенных значений можно задать на вкладке Спецификации данных.
- Заданный интервал времени позволяет процедуре обнаружить несколько наблюдений в одном интервале времени, которые нужно объединить вместе, и выровнять наблюдения по границе интервала, например по первому числу месяца, для обеспечения эквидистантности наблюдений. Например, если интервал времени - это Месяцы, несколько дат одного месяца объединяются вместе. Такой тип объединения называется *группировка*. По умолчанию наблюдения при группировке суммируются. Для группировки можно задать и другой способ обработки значений, такой как вычисление среднего, используя параметры **Агрегирование и распределение** на вкладке Спецификации данных.
- Для некоторых интервалов времени дополнительные параметры могут определить разрывы в обычно равномерно расположенных интервалах. Например, если интервал времени - это Дни, но допускаются только рабочие дни, можно указать, что в неделе есть только пять дней и она начинается в понедельник.

Наблюдения определяются периодами или циклическими периодами

Наблюдения можно определять одним или несколькими целочисленными полями, представляющими периоды или повторяющиеся циклы периодов вплоть до произвольного числа уровней циклов. При помощи такой структуры можно описать ряд наблюдений, который не подходит ни под один из обычных интервалов времени. Например, финансовый год всего с 10 месяцами можно описать полем цикла, представляющим годы, и полем периода, представляющим месяцы, где длина одного цикла равна 10.

Поля, задающие циклические периоды, определяют иерархию периодических уровней, где низший уровень определен полем **Период**. Следующий высший уровень задается циклическим полем с уровнем 1, после него - циклическим полем с уровнем 2, и т.д. Значения полей для каждого уровня, кроме наивысшего, должны быть периодическими по отношению к следующему высшему уровню. Значения для наивысшего уровня не могут быть периодическими. Например, в случае 10-месячного финансового года месяцы периодически повторяются по годам, но годы не периодические.

- Длина цикла на конкретном уровне - это период следующего низшего уровня. Для примера с финансовым годом есть только один циклический уровень и длина его цикла равна 10, так как следующий низший уровень представляет месяцы, а в каждом заданном финансовом году 10 месяцев.
- Укажите начальное значение для любого периодического поля, не начинающегося с 1. Этот параметр необходим для обнаружения пропущенных значений. Например, если периодическое поле начинается с 2, но начальное значение задано как 1, процедура предположит, что существует пропущенное значение для первого периода в каждом цикле этого поля.

Узел потока TCM - Опции интервала времени

Интервал времени, используемый для анализа, может отличаться от интервала времени наблюдения. Например, если интервал времени наблюдений - это Дни, для анализа вы можете выбрать интервал времени Месяцы. Затем данные агрегируются из ежедневных в ежемесячные, прежде чем строится модель. Вы можете распределить данные также из более длительного в короткий интервал времени. Например, если наблюдения квартальные, данные можно распределить из ежеквартальных в ежемесячные.

Возможные варианты выбора интервала времени анализа зависят от того, как определены наблюдения, и от интервала времени этих наблюдений. В частности, когда наблюдения определены по циклическим периодам, поддерживается только агрегирование. В этом случае интервал времени анализа должен быть не меньше, чем интервал времени наблюдений.

Интервал времени для анализа задается из параметров **Интервал времени** на вкладке Спецификации данных. Способ, которым объединяются или распределяются данные, задается параметрами **Агрегирование и распределение** на вкладке Спецификации данных.

Узел потока TCM - Опции Агрегация и Распределение

Функции агрегации

Когда используемый для анализа интервал времени превышает временной интервал наблюдений, производится агрегирование входных данных. Например, агрегирование применяется, если интервал наблюдений - это Дни, а интервал анализа - Месяцы. Доступны следующие функции агрегирования: mean, sum, mode, min или max.

Функции распределения

Когда используемый для анализа интервал короче интервала наблюдений, входные данные распределяются. Например, распределение производится, если интервал наблюдений - это Кварталы, а интервал анализа - Месяцы. Доступны следующие функции распределений: mean или sum.

Функции группировки

Группировки применяются, когда наблюдения определяются значениями дата/время и в одном интервале времени производится несколько наблюдений. Например, если временной интервал наблюдений - это Месяцы, несколько дат одного месяца можно сгруппировать и связывать с месяцем, которому они принадлежат. Доступны следующие функции группировки: mean, sum, mode, min или max. Группировка производится всегда, когда наблюдения определены значениями даты/времени, а временной интервал наблюдений задан как Нерегулярный.

Примечание: Хотя группировка - это вариант агрегирования, она производится до любой обработки пропущенных значений, а формальное агрегирование выполняется после такой обработки. Когда временной интервал наблюдений задан как нерегулярный, агрегирование производится только через функцию группировки.

Объединить межсуточные наблюдения на вчерашний день

Указывает, агрегировать ли наблюдения, время которых перешло за границу дня, со значениями за предыдущий день. Например, для почасовых измерений в течение суток, разделенных на интервалы по 8 часов, начинающихся в 20:00, этот параметр указывает, включать ли наблюдения от 00:00 до 04:00 в агрегированные результаты предыдущего дня. Этот параметр применим только в том случае, когда интервал времени наблюдений - это Часов в день, Минут в день или Секунд в день, а временной интервал анализа - Дни.

Пользовательские параметры для заданных полей

Функции агрегирования, распределения и группировки можно задать, используя поле за полем. Эти параметры переопределят параметры по умолчанию для функций агрегирования, распределения и группировки.

Узел потока TCM - Опции Значение отсутствия

Пропущенные значения во входных данных заменяются на импутированное значение. Возможны следующие способы замены:

Линейная интерполяция
Заменяет пропущенные значения с помощью линейной интерполяции. Для интерполяции используются последнее валидное (непропущенное) значение перед пропущенным и первое валидное значение после пропущенного. Если в первом или последнем наблюдении ряда есть пропущенное значение, используются два ближайших непропущенных значения в начале или в конце ряда.

Среднее значение ряда
Заменяет пропущенные значения средним для всего ряда.

Среднее близких точек
Заменяет пропущенные значения средним из валидных окружающих значений. Интервал ближайших точек здесь - количество точек до и после текущей, которые используются при вычислении среднего.

Медиана близких точек
Заменяет пропущенные значения медианой из валидных окружающих значений. Интервал ближайших точек здесь - количество точек до и после текущей, которые используются при вычислении медианы.

Линейный тренд
Эта опция использует все непропущенные наблюдения в ряду для подгонки по простой модели линейной регрессии, которая затем используется для импутации пропущенных значений.

Другие параметры:

Максимальная процентная доля пропущенных значений (%)
Задает максимальную процентную долю пропущенных значений, разрешенных для любого ряда. Ряды с большим количеством пропущенных значений, чем задано этим пределом, исключаются из анализа.

Узел потока TCM - Общие опции данных

Максимальное количество отдельных значений на поле измерения
Этот параметр применяется к многомерным данным и задает максимальное количество отдельных значений, которое разрешено для любого поля измерений. По умолчанию для этого предела задано значение 10000, но оно может быть увеличено до сколь угодно большого числа.

Узел потока TCM - Общие опции построения

Ширина доверительного интервала (%)
Этот параметр управляет доверительными интервалами для параметров прогнозов и модели. Вы можете указать любое положительное число меньше 100. По умолчанию используется 95%-ный доверительный интервал.

Максимальное количество входных элементов для каждого элемента назначения
Этот параметр указывает максимальное количество входных элементов, которые разрешены в модели для каждого элемента назначения. Можно указать целое число в диапазоне от 1 до 20. Модель для каждого элемента назначения всегда содержит сдвиг значений относительно друг друга (лаг), поэтому задание для этого параметра 1 указывает, что единственный входной элемент и есть элемент назначения.

Допуск модели
Этот параметр управляет итерационным процессом, используемым для определения наилучшего набора входных элементов для каждого элемента назначения. Можно указать любое положительное значение. По умолчанию используется значение 0.001.

Порог выбросов (%)

Наблюдение помечается флагом как выброс, если вычисленная в модели вероятность, что данное наблюдение - это выброс, превосходит этот порог. Можно указать значение в диапазоне от 50 до 100.

Количество задержек для каждого ввода

Этот параметр задает количество шагов задержки для каждого входного элемента в модели для элемента назначения. По умолчанию количество шагов задержки автоматически определяется из интервала времени, используемого в анализе. Например, если интервал времени - это месяцы (с инкрементом в один месяц), количество задержек равно 12. Кроме этого, количество задержек можно задать явным образом. Заданное значение должно быть целым числом от 1 до 20.

Продолжить оценивание, используя существующие модели

Если вы уже сгенерировали причинную модель времени, выберите эту опцию, чтобы повторно использовать параметры критериев, заданные для этой модели, а не строить новую модель. Таким образом можно сэкономить время на повторное оценивание и создать новый прогноз на основе той же модели, что и раньше, но используя более новые данные.

Узел потока TCM - Опции Период оценки

По умолчанию период оценивания начинается с самого раннего наблюдения и завершается самым поздним наблюдением в ряду.

По начальному и конечному времени

Можно задать и начальное, и конечное время периода оценивания, а также указать только начало или только конец. Если не задается начальное или конечное время периода оценивания, используется значение по умолчанию.

- Если наблюдения определяются полем дата/время, введите значения начального и конечного времени в том же формате, который используется для поля даты/времени.
- Для наблюдений, определенных циклическими интервалами, задайте значение для каждого из полей циклических интервалов. Каждое поле выводится в отдельном столбце.

По самым поздним и самым ранним интервалам времени

Определяет период оценивание как заданное количество интервалов времени, начинающихся самым ранним интервалом или заканчивающихся самым поздним интервалом в данных с учетом возможного сдвига. В этом контексте интервал времени относится к интервалу времени анализа. Допустим, например, что наблюдения ежемесячные, но интервал анализа квартальный. Если задать опцию **Последний** и значение 24 для параметра **Число интервалов**, это будет означать последние 24 квартала.

Если хотите, можно исключить указанное число интервалов времени. Например, указание последних 24 интервалов времени и значения 1 для исключаемого количества означает, что период оценивания состоит из 24 интервалов, предшествующих последнему интервалу.

Узел потока TCM - Опции модели

Имя модели

Для модели можно задать пользовательское имя или же принять сгенерированное автоматически имя *TCM*.

Прогноз

Опция **Распространить записи на будущее** задает число интервалов оценивания для прогнозов на время после окончания периода оценивания. Интервал времени в этом случае - это интервал времени анализа, заданный на вкладке Спецификации данных. Если затребованы прогнозы, выполняется автоматическое построение моделей авторегрессии для всех входных рядов, не являющихся также и назначениями. Затем эти модели используются для генерирования значения данных входных рядов на период прогноза. Максимального ограничения для этого параметра нет. for this setting.

Узел пространственно-временных диапазонов

Пространственно-временные диапазоны (Space-Time-Boxes, STB) - это расширение пространственных положений с географической привязкой. Более конкретно, STB - это алфавитно-цифровая строка, представляющая область в пространстве и времени правильной формы.

Например, STB вида **dr5ru7|2013-01-01 00:00:00|2013-01-01 00:15:00** составлен из трех частей:

- Геохеш **dr5ru7**
- Начальная отметка времени - **2013-01-01 00:00:00**
- Конечная отметка времени - **2013-01-01 00:15:00**

Например, можно использовать информацию о месте и времени для повышения вероятности, что два объекта одинаковы, так как они находятся фактически в одном и том же месте в одно и то же время. Другой вариант - повысить точность идентификации взаимосвязи, показав, что два объекта связаны своей близостью в пространстве и времени.

Для соответствия вашим требованиям можно выбрать режим **Индивидуальные записи** или **Сосредоточение**. В обоих режимах требуются следующие основные подробности:

Поле широты. Выберите поле, определяющее широту (в системе координат WGS84).

Поле долготы. Выберите поле, определяющее долготу (в системе координат WGS84).

Поле отметки времени. Выберите поле, определяющее время или дату.

Опции отдельных записей

Используйте эту опцию, чтобы добавить дополнительное поле к записи и определить ее положение в данное время.

Произвести от. Выберите одно или несколько значений плотности пространства и времени, из которых будет производиться значение в новом поле. Дополнительную информацию смотрите в разделе “Определение плотности в пространственно-временном диапазоне” на стр. 110.

Расширение имени поля. Введите расширение, которое вы хотели бы добавить к имени нового поля. Можно выбрать добавление этого расширения в виде **Суффикса** или **Префикса**.

Опции аттрактора

Аттрактор можно трактовать как характеристику положения или времени, где или когда некоторый объект постоянно или регулярно бывает. Например, эту характеристику можно использовать для транспортного средства, регулярно выполняющего рейсы по одному маршруту, и соответственно определять отклонения от нормы.

Детектор аттракторов отслеживает движение объектов и помечает флагами условия, при которых объект оказывается в области "захвата в аттрактор". Детектор аттракторов автоматически назначает каждый аттрактор с флагом одному или нескольким STB и использует отслеживание события и объекты в памяти для обнаружения аттракторов с оптимальной эффективностью.

Плотность STB. Выберите одно или несколько значений плотности пространства и времени, из которых будет производиться значение в новом поле. Например, значение **STB_GH4_10MINS** может соответствовать четырехсимвольному блоку географической сетки размером примерно 20 км на 20 км и временному окну 10 минут. Дополнительную информацию смотрите в разделе “Определение плотности в пространственно-временном диапазоне” на стр. 110.

Поле ID объекта. Выберите объект, который будет использоваться как идентификатор аттрактора. Это поле ID определяет событие.

Минимальное число событий. Событие - это строка в данных. Выберите минимальное число появления событий для объекта, чтобы он рассматривался как аттрактор. Аттрактор должен квалифицироваться также на основании следующего поля **Время пребывания - по крайней мере**.

Минимальное время пребывания. Задайте минимальную длительность пребывания объекта в одном положении. Например, это поможет исключить из рассмотрения автомобили, останавливающиеся на светофоре, который нельзя рассматривать как аттрактор. Аттрактор должен квалифицироваться также на основании предыдущего поля **Минимальное количество событий**.

Далее подробно описывается, что квалифицируется как аттрактор:

Пусть $e_1, ..., e_n$ обозначает все упорядоченные по времени события, полученные от данного ID объекта в течение интервала времени (t_1, t_n). Эти события квалифицируются как аттрактор, если:

- n >= *минимальное число событий*
- $t_n - t_1$ >= *минимальное время пребывания*
- Все события $e_1, ..., e_n$ происходят в одном STB

Разрешить аттракторам пересекать границы STB. При выборе этой опции определение аттрактора становится менее жестким и позволяет, например, включить аттракторы, распространяющиеся на несколько пространственно-временных диапазонов. Например, если ваши STB определены как целые часы, при выборе этой опции будут распознаваться допустимыми объекты, присутствующие где-то в течение часа, хотя этот час состоит из 30 минут до полуночи и 30 минут после. Если эта опция не выбрана, все 100% времени присутствия должны попадать в один пространственно-временной диапазон.

Минимальная доля событий в квалификационном диапазоне по времени (%). Доступно только при выборе опции **Разрешить аттракторам пересекать границы STB**. Используйте эту опцию для управления тем, в какой степени сообщаемый аттрактор для одного STB может фактически перекрываться с другим. Выберите минимальную долю событий, которые должны произойти в одном STB, чтобы определить аттрактор. Если задать значение 25%, а доля событий составит 26%, это будет рассматриваться как аттрактор.

Допустим, например, что для вашей конфигурации детектора аттракторов требуется по крайней мере два события (минимальное число событий = 2) и время непрерывного пребывания не меньше двух минут в пространственном блоке 4-байтового геохеша и в 10-минутном блоке времени (STB_NAME = STB_GH4_10MINS). Аттрактор может детектироваться, например, при пребывании объекта в одном пространственном блоке 4-байтового геохеша, когда происходят три квалифицирующих события в 10-минутном промежутке межу 16:57 и 17:07, а именно, в 16:58, 17:01 и 17:03. Значение процентной доли квалификационного интервала времени задает STB, который будет выделен аттрактору, следующим образом:

- **100%.** Есть сообщение об аттракторе во временном блоке 17:00 - 17:10, нет сообщения для временного блока 16:50 - 17:00 (события в 17:01 и в 17:03 удовлетворяют всем обязательным условиям для квалификации аттрактора, и 100% этих событий произошло во временном блоке 17:00 - 17:10).
- **50%.** Сообщено об аттракторах в обоих временных блоках (события в 17:01 и 17:03 удовлетворяют всем обязательным условиям для квалификации аттрактора и по крайней мере 50% этих событий произошло во временном блоке 16:50 - 17:00, а по крайней мере 50% этих событий произошло во временном блоке 17:00 - 17:10).
- **0%.** Сообщено об аттракторах в обоих интервалах времени.

Когда задано значение 0%, отчеты об аттракторах включают в себя STB, представляющие все временные блоки, затронутые в течение квалификационного интервала. Квалификационный интервал должен быть не

больше длительности соответствующего временного блока в STB. Другими словами, не должно возникнуть такой ситуации, что 10-минутный STB сконфигурирован вместе с 20-минутный квалификационным интервалом.

Об аттракторе сообщается, как только выполнены квалификационные условия, и несколько раз для одного STB об аттракторе не сообщается. Допустим, для аттрактора квалифицировано три события, а в течение квалификационного интервала времени внутри одного STB всего произошло 10 событий. В этом случае об аттракторе сообщается при возникновении третьего квалификационного события. Ни одно из дополнительных семи событий не инициирует отправку отчета об аттракторе.

Примечание:

- Данные события в памяти для детектора аттракторов не используются совместно несколькими процессами. Поэтому у отдельного объекта есть родственность с отдельным узлом детектора аттракторов. Таким образом, входящие данные о движении для объекта всегда должны согласованно передаваться на узел детектора аттракторов, отслеживающего этот объект, что обычно происходит на одном узле в течение всего выполнения задания.
- Данные события в памяти для детектора аттракторов непостоянные. При всяком закрытии и перезапуске детектора аттракторов все аттракторы текущей работы теряются. Это означает, что остановка и перезапуск процесса может привести к тому, что система не сообщит о действительных аттракторах. Потенциальным средством от этого может быть воспроизведение некоторых хронологических данных о движении (например, можно вернуться на 48 часов назад и воспроизвести записи движения, доступные для любого перезапущенного узла).
- Данные должны направляться в детектор аттракторов последовательно по времени.

Определение плотности в пространственно-временном диапазоне

Выберите размер (плотность) вашего пространственно-временного диапазона (Space-Time-Boxe, STB), задав физическую область и время наблюдений для включения в каждый из диапазонов.

Пространственная плотность. Выберите размер области для включения в каждый STB.

Интервал времени. Выберите длительность в часах для включения в каждый STB.

Имя поля. Присваивается префикс STB; выполняется автоматически на основании выбора в предыдущих двух полях.

Глава 4. Узлы операций с полями

Обзор операций с полями

После начального изучения данных может потребоваться выбрать, очистить или построить данные при подготовке к анализу. Палитра Операции с полями содержит множество узлов, полезных для этого преобразования и подготовки данных.

Например, при помощи узла извлечения можно создать атрибут, не представленный в текущий момент в данных. А при помощи узла разделения на интервалы можно перекодировать значения полей автоматически для намеченного анализа. Скорее всего, вы будете часто использовать узел Тип, позволяющий назначить шкалу измерений, значения и роль моделирования для каждого поля в наборе данных. Его операции полезны для обработки пропущенных значений и моделирования нисходящего потока.

Палитра Операции с полями содержит следующие узлы:

Узел автоматической подготовки данных (Automated Data Preparation, ADP) может анализировать ваши данные и находит исправления, выявляет проблемные и малополезные поля, создает при необходимости производные атрибуты и повышает производительность, применяя интеллектуальные способы анализа и выборки. Этот узел можно использовать в полностью автоматическом режиме, позволив ему выбирать и применять исправления или предварительно просматривать изменения перед тем, как они сделаны и приняты, а при желании применять, отклонять или исправлять их.

Узел Тип задает метаданные и свойства полей. Например, можно задать уровень измерений (количественный, номинальный, порядковый или флаговый) для каждого поля, задать опции для обработки отсутствующих значений и системных null, задавать роль поля для целей моделирования, задавать метки полей и значений и задавать значения для поля.

Узел Фильтр фильтрует (отбрасывает) поля, переименовывает поля и отображает поля с одного узла источника на другой.

Узел извлечения изменяет значения данных или создает новые поля из одного или нескольких существующих полей. Он создает поля формулы типа, флага, номинала, состояния, количества и условного выражения.

Узел Ансамбль объединяет два или более слепков моделей для получения более точных предсказаний, чем можно получить от любой модели.

Узел заполнителя замещает значения полей и заменяет систему хранения. Вы можете заменить значения на основе условия CLEM, такого как `@BLANK(@FIELD)`. Как вариант, вы можете выбрать замещение всех пустых значений или значений null на конкретное значение. Узел заполнителя часто используется вместе с узлом Тип для замены пропущенных значений.

Узел анонимизации преобразует способ представления имен и значений полей уровнем ниже, маскируя таким образом исходные данные. Это может быть полезно, если вы хотите разрешить другим пользователям построить модели, используя чувствительные данные, такие как имена клиентов или другие подробности.

Узел переклассификации преобразует один набор категориальных значений в другой. Переклассификация полезна для сворачивания категорий или для перегруппировки данных для анализа.

Узел разделения на интервалы автоматически создает новые номинальные поля на основе значений одного или нескольких существующих количественных полей (числового диапазона). Например, можно преобразовать количественное входное поле в новое категориальное поле, содержащее группы входных данных, как отклонения от среднего. После создания интервалов для нового поля вы можете сгенерировать узел извлечения на основе точек деления.

Узел анализа Новизна, частота, деньги (Recency, Frequency, Monetary - RFM) позволяет вам количественно определить, какие клиенты вероятно будут лучшими, исследуя, насколько недавно они сделали свои последние покупки (новизна), как часто они покупали (частота) и сколько денег потратили на все транзакции (деньги).

Узел Разделы генерирует поле раздела, которое разбивает данные на отдельные подмножества для стадий обучения, испытания и проверки при построении моделей.

Узел Задать как флаг извлекает несколько полей флагов на основании категориальных значений, определенных для одного или нескольких номинальных полей.

Узел реструктуризации преобразует номинальное или флаговое поле в группу полей, которые можно заполнить значениями еще одного поля. Например, если задано поле с именем *тип_платежа*, у которого могут быть значения *кредит*, *наличные* и *дебет*, могут быть заданы три новые поля (*кредит*, *наличные*, *дебет*), каждое из которых может содержать значение фактического выполненного платежа.

Узел Транспонирование меняет данные в строках и столбцах, чтобы записи становились полями, а поля записями.

Узел Интервалы времени используется для задания интервалов и получения нового поля времени для операций оценки или прогноза. Поддерживается весь диапазон интервалов времени от секунд до лет.

Узел Хронология создает новые поля, содержащие данные из полей в предыдущих записях. Хронологические узлы чаще всего используются для последовательных данных, таких как данные временных рядов. Перед использованием узла Хронология может потребоваться отсортировать данные с использованием узла Сортировка.

Узел переупорядочения полей определяет естественный порядок, используемый для вывода полей нижележащего уровня. Этот порядок влияет на показ полей во многих положениях, таких как таблицы, списки и средство выбора полей. Эта операция полезна при работе с обширными наборами данных, чтобы сделать нужные поля более наглядными.

В SPSS Modeler элементы, такие как пространственные функции построителя выражений, узел Пространственное предсказание (Spatio-Temporal Prediction, STP) и узел Визуализация карт используют систему координат проекции. При помощи узла Репроецирование можно изменить систему координат для любых импортируемых данных, где используется географическая система координат.

Часть этих узлов можно сгенерировать непосредственно из отчета аудита, создаваемого узлом аудита данных. Дополнительную информацию смотрите в разделе “Генерирование других узлов для подготовки данных” на стр. 304.

Автоматическая подготовка данных

В любом проекте подготовка данных для анализа - один из важнейших шагов; именно этот шаг традиционно требовал наибольших затрат времени. Инструмент Автоматическая подготовка данных (АПД) решает эту задачу, для чего анализирует данные и находит решения выявленных проблем, выявляет проблемные и малополезные поля, создает при необходимости производные атрибуты и повышает производительность, применяя интеллектуальные методы скрининга. Можно использовать полностью **автоматический** алгоритм, доверив ему выбор и принятие исправлений, а можно просмотреть предлагаемые изменения в **интерактивном** режиме и принять нужные.

Благодаря автоматической подготовке данных вы сможете легко и быстро подготовить данные для моделирования, даже если не были раньше знакомы с используемыми при этом понятиями статистики. После этого модели будут строиться и оцениваться быстрее; более того, благодаря автоматической подготовке данных достигаются более сильные результаты при автоматическом моделировании в таких процессах, как обновление модели и чемпион / претендент.

Примечание: когда автоматическая подготовка данных готовит поле для анализа, она не заменяет существующие значения и свойства существующего поля, а создает новое поле, скорректированное или преобразованное. Старое поле не используется в дальнейшем анализе, для его роли задается значение Нет.

Пример. Страховая компания с ограниченными ресурсами для исследования страховых исков домовладельца желает построить модель для того, чтобы отмечать подозрительные, потенциально мошеннические иски. Перед построением модели компания готовит данные для моделирования и при этом использует автоматическую подготовку данных. Поскольку желательно пересмотреть предложенные преобразования перед их применением, используется автоматическая подготовка данных в интерактивном режиме.

Группа предприятий автомобильной промышленности отслеживает продажи ряда легковых автомобилей. С целью выявить успешные и неудачные модели сотрудники стараются выявить взаимосвязь между продажами транспортных средств и их характеристиками. Чтобы подготовить данные к анализу, используется автоматическая подготовка данных; затем строятся и сравниваются между собой модели, использующие неподготовленные и подготовленные данные.

Какова ваша цель? Автоматическая подготовка данных рекомендует шаги по подготовке данных, влияющие на скорость работы алгоритмов при построении моделей и на точность прогноза соответствующих моделей. Эти шаги могут включать преобразования, конструирование и отбор показателей. Целевое поле также может быть преобразовано. Можно указать приоритеты построения моделей, на которых должна сосредоточиться подготовка данных.

- **Сбалансировать скорость и точность.** Эта опция при подготовке данных отдает равный приоритет как скорости обработки данных алгоритмами построения моделей, так и точности прогнозов.
- **Оптимизировать скорость.** Эта опция при подготовке данных отдает приоритет скорости обработки данных алгоритмами построения моделей. Выберите эту опцию при работе с большими наборами данных, а также когда нужен быстрый ответ.
- **Оптимизировать точность.** Эта опция при подготовке данных отдает приоритет точности прогнозов, генерируемых алгоритмами построения моделей.
- **Настроить анализ.** Выберите этот вариант, если нужно вручную изменить алгоритм на вкладке Параметры. Обратите внимание на то, что этот выбор производится автоматически, если на вкладке Параметры сделать изменения, несовместимые с одной из других целей.

Обучение узла

Узел автоматической подготовки данных реализован как узел процесса и работает аналогично узлу Тип; **обучение** узла автоматической подготовки данных соответствует созданию экземпляра узла Тип. После анализа указанные преобразования применяются к данным без дальнейшего анализа, пока не изменится использующая их модель данных. Подобно узлам Тип и Фильтр, когда узел автоматической подготовки данных отсоединяется, он запоминает модель данных и преобразования, так что при повторном подключении его не нужно обучать еще раз; таким образом, его можно обучить на поднаборе типичных данных, а затем скопировать или внедрить для использования с реальными данными, создав столько экземпляров, сколько нужно.

Использование панели инструментов

С панели инструментов можно запустить анализ данных и обновить панель анализа, а также сгенерировать узлы, которые можно использовать совместно с первоначальными данными.

- **Генерировать** Из этого меню можно сгенерировать узел фильтра или производного поля. Обратите внимание на то, что это меню доступно, только когда на вкладке Анализ показан тот или иной анализ.

  Узел Фильтр удаляет преобразованные входные поля. Если для узла автоматической подготовки данных задано оставлять в наборе данных первоначальные входные поля, эта опция восстанавливает первоначальный набор входных полей, что позволяет интерпретировать поле оценки в терминах входных полей. Например, это может быть полезно, когда нужно сгенерировать диаграмму зависимости поля оценки от различных входных полей.

  Узел извлечения может восстановить первоначальный набор данных и единицы поля назначения. Сгенерировать узел извлечения можно, только когда узел автоматической подготовки данных содержит анализ, изменяющий шкалу диапазонного поля назначения (то есть на панели Подготовить входные поля и поля назначения выбрано изменение шкалы Бокса-Кокса). Нельзя сгенерировать узел извлечения, если поле назначения - не диапазон или если не выбрано изменение шкалы Бокса-Кокса. Дополнительную информацию смотрите в разделе “Генерирование узла извлечения” на стр. 126.
- **Вид** Содержит опции, управляющие выводом информации на вкладке Анализ. Сюда относятся элементы управления для редактирования диаграмм и выбор вывода для основной панели и дополнительных панелей.
- **Предварительный просмотр** Выводит пример преобразований, которые будут применены ко входным данным.
- **Анализировать данные** Запускает анализ с использованием текущих значений параметров и выводит результаты на вкладке Анализ.
- **Очистить анализ** Удаляет существующий анализ (это доступно, только если текущий анализ существует).

Состояние узла

Состояние узла автоматической подготовки данных на холсте IBM SPSS Modeler изображается стрелкой или делением на значке; оно показывает, был ли выполнен анализ.

Вкладка Поля

Перед построением модели необходимо указать поля, которые должны служить полями назначения и входными полями. За немногими исключениями, все узлы моделирования будут использовать информацию о полях из узла Тип, расположенного выше. Если вы используете узел Тип для выбора входных полей и полей назначения, на этой вкладке можно ничего не менять.

Использовать значения узла типа. Эта опция указывает узлу, что следует использовать информацию о полях из узла Тип, расположенного выше. Это вариант по умолчанию.

Использование настраиваемых параметров. Эта опция указывает узлу, что следует использовать информацию о полях, заданную здесь, а не ту, что задана на расположенных выше узлах Тип. После выбора этого варианта задайте приведенные ниже поля, как это потребуется.

Цель. Для тех моделей, для которых требуется одна или несколько полей назначения, выберите поле назначения или несколько полей. Это аналогично заданию для поля роли *Поле назначения* на узле Тип.

Входные. Выберите одно или несколько входных полей. Это аналогично заданию для поля роли *Входное* на узле Тип.

Вкладка Параметры

Вкладка Параметры содержит несколько различных групп параметров, которые можно изменять, чтобы точно настроить то, как алгоритм будет обрабатывать имеющиеся данные. Если в настройку параметров по умолчанию внести изменения, которые несовместимы с другими целями, то выбор на вкладке Цель будет автоматически изменен на **Настроить анализ**.

Параметры полей

Использовать поле частоты. Эта опция позволяет выбрать поле частотного веса. Используйте это поле, если каждая запись в данных обучения представляет несколько объектов; например, при использовании агрегированных данных. Значениями в поле должны быть количества объектов, представляемых каждой записью.

Использовать поле веса. Эта опция позволяет выбрать поле веса наблюдений. Веса наблюдений используются для учета различий в дисперсии разных уровней поля назначения.

Как обрабатывать поля, исключаемые из моделирования. Задайте, что происходит с исключенными полями; можно выбрать их отфильтровывание из данных или просто задание для их *Роли* значения **Нет**.

Примечание: Эта опция будет применяться также к полю назначения, если оно преобразовывается. Например, если в качестве поля **Назначение** используется новая извлеченная версия назначения, исходное поле назначения или отфильтровывается, или для него задается значение **Нет**.

Если входящие поля не соответствуют существующему анализу. Укажите, что случится, если одно или несколько обязательных входных полей будут отсутствовать во входящем наборе данных при выполнении обученного узла ADP.

- **Остановить выполнение и сохранить текущий анализ.** При этом процесс выполнения останавливается, сохраняется текущая информация анализа и выводится сообщение об ошибке.
- **Очистить существующий анализ и проанализировать новые данные.** При этом существующие данные анализа очищаются, входные данные анализируются и к этим данным применяются рекомендованные преобразования.

Подготовить даты и значения времени

Многие алгоритмы моделирования не умеют обрабатывать непосредственную информацию о датах и значениях времени; эти параметры дадут вам возможность взять даты и значения времени в существующих данных и произвести новые данные о продолжительности, которые могут служить входными данными модели. Поля, содержащие даты и значения времени, нужно заранее определить как поля с типами даты и времени. Первоначальные поля даты и времени не будут рекомендованы как входные поля модели после автоматической подготовки данных.

Подготовить даты и время к построению моделей. При отмене этой опции остальные элементы управления панели Подготовить даты и значения времени отключаются, но их состояние запоминается.

Вычислить время, прошедшее до определенной даты. Результат - число лет, месяцев или дней от опорной даты для каждой переменной, содержащей даты.

- **Дата.** Укажите дату, от которой будет отсчитываться продолжительность исходя из информации о дате во входных данных. Если задать **Сегодняшняя дата**, это будет значить, что при выполнении автоматической подготовки данных всегда будет использоваться текущая дата системы. Чтобы использовать некоторую конкретную дату, выберите **Фиксированная дата** и введите нужную дату. При создании узла в поле **Фиксированная дата** автоматически вводится текущая дата.
- **Единица продолжительности времени.** Укажите, что автоматическая подготовка данных должна автоматически определять единицу продолжительности для дат, или выберите **Фиксированные единицы** - Годы, Месяцы или Дни.

Вычислить время, прошедшее до определенного момента времени. Результат - число часов, минут или секунд от опорного времени для каждого значения времени, заданного в переменной.

- **Время.** Укажите время, от которого будет отсчитываться продолжительность исходя из информации о времени во входных данных. Если задать **Текущее время**, это будет значить, что при выполнении автоматической подготовки всегда будет использоваться текущее время системы. Чтобы использовать некоторое конкретное время, выберите **Фиксированное время** и введите нужные сведения. При создании узла в поле **Фиксированное время** автоматически вводится текущее время.
- **Единица продолжительности времени.** Укажите, что автоматическая подготовка данных должна автоматически определять единицу продолжительности времени, или выберите **Фиксированные единицы** - Часы, Минуты или Секунды.

Извлечь циклические элементы времени. Эти параметры служат для извлечения из поля даты или времени одного или нескольких полей. Например, если включены все три переключателя в группе даты, то входное значение даты "1954-05-23" разделяется на три значения 1954, 5 и 23; эти значения записываются в поля с суффиксами, определенными на панели **Имена полей**, а первоначальное поле игнорируется.

- **Извлечь из дат:** Укажите, нужно ли из всех входных полей даты извлекать годы, месяцы и дни в любом сочетании.
- **Извлечь из значений времени.** Укажите, нужно ли из всех входных полей времени извлекать часы, минуты и секунды в любом сочетании.

Исключение полей

Низкое качество данных может ухудшить точность прогнозов, поэтому есть смысл задавать уровень приемлемого качества данных для входных характеристик. Все поля, которые оказываются константами, и поля, в которых 100% значений отсутствует, исключаются автоматически.

Исключить входные поля низкого качества. Отмена этой опции отключает остальные элементы управления исключением полей, запоминая их состояние.

Исключить поля со слишком большим количеством пропущенных значений. Поля, где процент отсутствующих значений превосходит указанный максимум, отстраняются от дальнейшего анализа. Задайте значение от 0

до 100; значение 0 эквивалентно отмене опции, а значение 100 не имеет особого смысла, потому что поля со 100 процентами отсутствующих значений исключаются автоматически. Значение по умолчанию - 50.

Исключить номинальные поля со слишком большим количеством уникальных категорий. Номинальные поля, где число категорий превосходит указанный максимум, отстраняются от дальнейшего анализа. Задайте целое положительное число. Значение по умолчанию - 100. Это полезно для автоматического отстранения от моделирования тех полей, которые содержат уникальную информацию о записи - ID, адрес или имя.

Исключить категориальные поля со слишком большим количеством значений в одной категории. Порядковые и номинальные поля, в которых одна из категорий содержит более указанного процента записей, отстраняются от дальнейшего анализа. Задайте значение от 0 до 100; значение 0 эквивалентно отмене опции, а значение 100 не имеет особого смысла, потому что поля, оказавшиеся константой, исключаются автоматически. Значение по умолчанию - 95.

Подготовка входных и выходных данных

Так как данные еще не в идеальном состоянии для обработки, вам может потребоваться скорректировать некоторые из параметров перед запуском процесса анализа. Например, это может включить в себя удаление выбросов, задание способов обработки пропущенных значений или настройку типа.

Примечание: Если вы изменяете значения на этой панели, автоматически изменяется и вкладка **Цели**, на которой выбирается опция **Пользовательский анализ**.

Подготовить входные поля и поля назначения для моделирования. Переключает все поля на панели в положение on или off.

Настроить тип и улучшить качество данных. Для входных и выходных полей можно отдельно задать несколько преобразований данных; это вызвано тем, что вам может не требоваться изменять значения поля назначения. Например, предсказанное значение прибыли в долларах более понятно, чем предсказание, выраженное логарифмом от долларового значения. Кроме этого, если в поле назначения есть пропущенные значения, нет никакого выигрыша предсказания при заполнении пропущенных значений, и в то же время заполнение пропущенных входных значений позволит использовать некоторые алгоритмы для обработки информации, которая в противном случае была бы потеряна.

Дополнительные параметры для этих преобразований, такие как значения отсечения для выбросов, общие для входных и выходных полей.

Для входных и/или выходных полей можно выбрать следующие настройки:

- **Скорректировать тип числовых полей.** Выберите эту возможность для определения, можно ли преобразовать числовые поля с уровнем измерения *Порядковый* на *Количественный*, или наоборот. Для управления преобразованием можно задать минимальное и максимальное значение порогов.
- **Переупорядочивание номинальных полей.** Выберите эту опцию для сортировки номинальных полей (набор) по порядку от самой редкой до самой частой категории.
- **Замените значения выбросов в количественных полях.** Задайте, заменять ли выбросы; используйте вместе с опцией **Способ замены выбросов** .
- **Количественные поля: заменить пропущенные значения средним значением.** Выберите для замены отсутствующих значений количественных показателей (диапазон).
- **Номинальные поля: замените пропущенные значения модой.** Выберите для замены отсутствующих значений номинальных показателей (набор).
- **Порядковые поля: замените пропущенные значения медианой.** Выберите для замены отсутствующих значений порядковых показателей (упорядоченный набор).

Максимальное количество значений для порядковых полей. Задайте порог для переопределения порядковых полей (упорядоченный набор) в количественные (диапазон). Значение по умолчанию 10; поэтому при наличии у порядкового поля более десяти категорий оно переопределяется в количественное (диапазон).

Минимальное количество значений для количественных полей. Задайте порог для переопределения количественный полей (диапазон) в порядковые (упорядоченный набор). Значение по умолчанию - 5; поэтому при наличии у количественного поля менее пяти значений оно переопределяется в порядковое (упорядоченный набор).

Значение отсечения выбросов. Задайте критерий отсечения выбросов, измеряемый в среднеквадратичных отклонениях; значение по умолчанию - 3.

Метод замены выбросов. Выберите, как заменять выбросы, или принудительным урезанием до значения усечения, или удалением и заданием для них значения отсутствия. Все выбросы, замещенные на значение отсутствия, обрабатываются согласно параметрам, выбираемым выше.

Разместить все количественные входные поля на общей шкале. Для нормализации количественных входных полей включите этот переключатель и выберите способ нормализации. По умолчанию используется **преобразование z-оценки**, где можно задать **Итоговое среднее** со значением по умолчанию 0 и **Итоговое среднеквадратичное отклонение** со значением по умолчанию 1. Как вариант, можно выбрать использование **Преобразования мин/макс** и задать минимальное и максимальное значения со значениями по умолчанию 0 и 100 соответственно.

Это поле особенно полезно при выборе опции **Выполнить конструирование показателя** на панели Конструировать и выбрать показатели.

Перемасштабировать количественное поле назначения при помощи преобразования Бокса-Кокса. Включите этот переключатель для нормализации количественного поля назначения (масштаб или диапазон). Значения по умолчанию для преобразования Бокса-Кокса равны 0 для **Итогового среднего значения** и 1 для **Итогового среднеквадратичного отклонения**.

Комментарий: Если вы выбираете нормализацию поля назначения, преобразуется и размерность поля назначения. В этом случае вам может потребоваться сгенерировать узел Извлечение для применения обратного преобразования, чтобы переключить преобразованные объекты обратно в распознаваемый формат для будущей обработки. Дополнительную информацию смотрите в разделе “Генерирование узла извлечения” на стр. 126.

Выбор конструкций и возможностей

Чтобы повысить точность прогноза по вашим данным, можно преобразовать входные поля или сконструировать новые на основе существующих полей.

Примечание: Если вы изменяете значения на этой панели, вкладка **Цели** автоматически обновляется для выбора опции **Пользовательский анализ**.

Преобразовать, построить и выбрать входные поля для увеличения точности прогноза. Переключает все поля на панели в положение on или off.

Объединить малочисленные категории, чтобы максимизировать связь с целевым полем. Включите, чтобы сделать модель более "экономной" путем уменьшения числа переменных, обрабатываемых в связи с полем назначения. При необходимости измените значение вероятности по умолчанию 0,05.

Обратите внимание, что если все категории сливаются в одну, исходная и полученная версии поля исключаются из модели, поскольку они не представляют ценности как предиктор.

Если переменной назначения нет, слить малочисленные категории на основе количеств. Если вы работает с данными, у которых нет поля назначения, можно слить малочисленные категории порядковых (упорядоченный набор) и/или номинальных (набор) объектов. Задайте минимальную процентную долю наблюдений или записей в данных, которые определяют категории для слияния; значение по умолчанию равно 10.

Категории сливаются с использованием следующих правил:

- Слияние не выполняется для бинарных полей.
- Если при слиянии остается только две категории, слияние останавливается.
- Если и у исходных категорий, и у создаваемых при слиянии категорий процентная доля содержащихся в них наблюдений больше заданного минимума, слияние останавливается.

Разделение количественных полей на интервалы с сохранением прогнозирующих возможностей. Если у вас есть данные, включающие категориальное поле назначения, можно количественные входные поля с сильными связями разделить на интервалы для повышения производительности обработки. При необходимости измените значение по умолчанию (0,05) вероятности для однородных подмножеств.

Если в результате категоризации некоторого конкретного поля получается всего одна категория, исходная и полученная версии поля исключаются, поскольку не представляют ценности как предиктор.

Примечание: Разделение на интервалы в ADP отличается от оптимального разделения, используемого в других частях IBM SPSS Modeler. При оптимальном разбиении для преобразования количественной переменной в категориальную используется информация об энтропии; для этого требуется сортировка данных и хранение всех данных в оперативной памяти. ADP использует однородные подмножества для категоризации количественной переменной, то есть при этом ADP не требуется сортировка данных и их хранение в оперативной памяти. Использование однородных подмножеств для категоризации количественной переменной означает, что число категорий после категоризации всегда меньше или равно числу категорий в поле назначения.

Выполнить отбор показателей. Выберите эту опцию для удаления объектов с низким коэффициентом корреляции. При необходимости измените значение вероятности по умолчанию 0,05.

Эта опция применяется к количественным входным показателям (только в том случае, если переменная назначения количественная) и к категориальных входным показателям.

Выполнить конструирование показателей. Выберите эту опцию для извлечения новых показателей из комбинации нескольких существующих показателей (которые затем отбрасываются из модели).

Эта опция применяется либо к непрерывным характеристикам, если поле назначения - непрерывное, либо когда поле назначения отсутствует.

Имена полей

Для лучшей узнаваемости новых и преобразованных характеристик автоматическая подготовка данных создает и применяет новые базовые имена, префиксы и суффиксы. Вы можете дополнительно отредактировать эти имена с учетом ваших потребностей и данных. Если нужно изменить другие метки, это нужно сделать в узле Тип, расположенном под данным узлом.

Преобразованные и сконструированные поля. Укажите расширения имен, которые будут применяться к преобразованным полям назначения и входным полям.

Обратите внимание на то, что в узле автоматической подготовки данных задание пустых строчных полей может, при некоторых вариантах обработки неиспользуемых полей, приводить к ошибке. Если на панели Параметры полей вкладки Параметры для опции **Как обрабатывать поля, исключенные из моделирования** задано значение **Отфильтровать неиспользуемые поля**, то расширения для входных полей и полей назначения можно задать пустыми. Первоначальные поля отфильтровываются, и преобразованные поля сохраняются поверх них; в этом случае у новых преобразованных полей имена совпадают с первоначальными.

Но если выбрать опцию **Задать для неиспользуемых полей направление 'Нет'**, пустые расширения имен для полей назначения и входных полей приведут к попыткам создать дубликаты имен полей, то есть к ошибкам.

Задайте, кроме того, префикс, который будет применяться к характеристикам, конструируемым согласно параметрам Отобрать и сконструировать. Новое имя создается путем добавления числового суффикса к этому сочетанию префикс - корень. Формат числа зависит от количества новых производных характеристик, например:

- если конструируется от 1 до 9 характеристик, они получат имена от характеристика1 до характеристика9.
- если конструируется от 10 до 99 характеристик, они получат имена от характеристика01 до характеристика99.
- если конструируется от 100 до 999 характеристик, они получат имена от характеристика001 до характеристика999.

Это гарантирует, что сконструированные характеристики будут осмысленно сортироваться при любой численности.

Значения продолжительности, вычисленные по датам и значениям времени. Укажите расширения имен, которые будут применяться к значениям продолжительности, вычисленной по датам и значениям времени.

Циклические элементы, извлеченные из дат и значений времени. Укажите расширения имен, которые будут применяться к циклическим элементам, извлеченным из дат и значений времени.

Вкладка Анализ

1. Когда вы будете удовлетворены параметрами автоматической подготовки данных, включая все изменения на вкладках Цель, Поля и Параметры, нажмите кнопку **Анализировать данные**; алгоритм применит параметры ко входным данным и выведет результаты на вкладке Анализ.

Вкладка Анализ содержит, в табличном и графическом формате, сводку об обработке данных и рекомендации о возможном изменении или улучшении данных для назначения оценок. Пересматривая рекомендации, вы можете принимать или отклонять их.

Вкладка Анализ состоит из двух панелей: основного представления слева и связанного, или дополнительного, справа. Есть три основных представления:

- Сводка по обработке полей (по умолчанию). Дополнительную информацию смотрите в разделе “Сводка по обработке полей” на стр. 121.
- Поля. Дополнительную информацию смотрите в разделе “Поля” на стр. 121.
- Итоги по действиям. Дополнительную информацию смотрите в разделе “Сводка действий” на стр. 122.

Есть четыре вида связанных (дополнительных) представлений:

- Мощность предсказания (по умолчанию). Дополнительную информацию смотрите в разделе “Точность прогноза” на стр. 122.
- Таблица переменных. Дополнительную информацию смотрите в разделе “Таблица полей” на стр. 123.
- Свойства поля. Дополнительную информацию смотрите в разделе “Свойства поля” на стр. 123.
- Свойства действия. Дополнительную информацию смотрите в разделе “Свойства действия” на стр. 124.

Связи между представлениями

В основном представлении подчеркнутый текст служит для открытия связанного представления. Щелчком по такому тексту выводятся подробности о соответствующем поле, наборе полей или шаге обработки. Последняя выбранная ссылка выделяется более темным цветом для наглядной связи между содержимым двух панелей представлений.

Восстановление представлений

Чтобы снова вывести первоначальные рекомендации анализа, отменив все изменения, которые вы внесли в представлениях Анализ, нажмите кнопку **Сброс** в нижней части основной панели.

Сводка по обработке полей

Таблица Сводка по обработке полей содержит снимок планируемого общего влияния обработки, включая перевод характеристик в другое состояние и изменение числа сконструированных характеристик.

Обратите внимание на то, что модель фактически не строится, и поэтому нет какого-либо показателя или диаграммы для изменения общей точности прогноза в результате подготовки данных; вместо этого доступны диаграммы точности прогноза отдельных рекомендуемых предикторов.

Таблица содержит следующую информацию:

- Число полей назначения.
- Число первоначальных (входных) предикторов.
- Предикторы, рекомендуемые для анализа и моделирования. Включает в себя общее число рекомендуемых полей; число первоначальных, непреобразованных рекомендуемых полей; число преобразованных рекомендуемых полей (исключая промежуточные версии полей, поля, производные от предикторов даты/времени и сконструированные предикторы); число рекомендуемых полей, производных от полей даты/времени; число рекомендуемых сконструированных предикторов.
- Число входных предикторов, не рекомендуемых для использования ни в какой форме - ни в первоначальной, ни как производное поле, ни как входное поле для сконструированного предиктора.

Там, где какая-либо часть информации **Полей** подчеркнута, можно щелкнуть, чтобы вывести панель дополнительных подробностей. Подробности о **Поле назначения**, **Входных характеристиках** и **Не используемых входных характеристиках** выводятся на дополнительной панели таблицы полей. Дополнительную информацию смотрите в разделе “Таблица полей” на стр. 123. **Характеристики, рекомендуемые для анализа**, выводятся на дополнительной панели Точность прогноза. Дополнительную информацию смотрите в разделе “Точность прогноза” на стр. 122.

Поля

Основная панель полей содержит обрабатываемые поля и рекомендации автоматической подготовки данных, какие из них использовать в нисходящей модели. Для любого поля рекомендацию можно переопределить, например, чтобы исключить сконструированные показатели или включить показатели, которые автоматическая подготовка данных рекомендует исключить. Если поле было преобразовано, можно принять предложенное преобразование или использовать первоначальную версию.

Панель полей состоит из двух таблиц, одна для поля назначения, другая для обработанных или созданных предикторов.

Таблица Поле назначения

Таблица **Поле назначения** выводится, только если в данных определено поле назначения.

Таблица содержит два столбца:

- **Имя.** Это имя или метка поля назначения; используется всегда первоначальное имя, даже если поле было преобразовано.
- **Уровень измерения.** Содержит значок, представляющий тип измерений; поместите указатель мыши на значок, чтобы вывести метку, описывающую данные - непрерывные, порядковые, номинальные и так далее.

  Если поле назначения было преобразовано, в столбце **Тип измерений** отражается конечная, преобразованная версия. *Примечание*: нельзя отключить преобразования для поля назначения.

Таблица Предикторы

Таблица **Предикторы** выводится всегда. Каждая строка таблицы представляет одно поле. По умолчанию поля сортируются в убывающем порядке точности прогноза.

Для обычных характеристик в качестве имени строки всегда используется первоначальное имя. Для полей даты/времени таблица содержит (в отдельных строках) как первоначальную, так и производную версии; кроме того, таблица содержит сконструированные предикторы.

Учтите, что преобразованные версии полей в таблице всегда представляют конечные версии.

По умолчанию таблица Предикторы содержит только рекомендуемые поля. Чтобы вывести на экран остальные поля, включите переключатель **Включить нерекомендованные поля в таблицу** над таблицей; эти поля появятся в нижней части таблицы.

Таблица содержит следующие столбцы:

- **Вариант для использования.** Выводит выпадающий список, управляющий нисходящим использованием поля и использованием предложенных преобразований. По умолчанию в выпадающем списке отражаются рекомендации.

  Для обычных предикторов, которые были преобразованы, выпадающий список содержит три варианта: **Преобразованное**, **Первоначальное** и **Не использовать**.

  Варианты для непреобразованных обычных предикторов: **Первоначальное** and **Не использовать**.

  Для производных полей даты/времени и сконструированных предикторов варианты такие: **Преобразованное** и **Не использовать**.

  Для первоначальных полей выпадающий список недоступен, и в нем выбрано **Не использовать**.

  Примечание: Когда предиктор, у которого есть и первоначальная, и преобразованная версии, вариант с **Первоначальное** меняется на **Преобразованное** или наоборот, автоматически изменяются значения **Тип измерений** и **Точность прогноза** для этих характеристик.
- **Имя.** Каждое имя поля - это ссылка. Щелкните по имени, чтобы вывести дополнительную информацию о поле на дополнительной панели. Дополнительную информацию смотрите в разделе “Свойства поля” на стр. 123.
- **Уровень измерения.** Содержит значок, представляющий тип данных; поместите указатель мыши на значок, чтобы вывести метку, описывающую данные - непрерывные, порядковые, номинальные и так далее.
- **Точность прогноза.** Точность прогноза выводится на экран только для полей, которые рекомендуются автоматической подготовкой данных. Этого столбца нет, если не определены поля назначения. Точность прогноза изменяется от 0 до 1, причем для "лучших" предикторов значения близки к 1. Вообще говоря, точность прогноза полезна для сравнения предикторов в одном анализе автоматической подготовки данных, но бесполезна для сравнения предикторов из разных анализов.

Сводка действий

При каждом действии, выполненном в ходе автоматической подготовки данных, входные предикторы преобразуются и/или отфильтровываются; оставшиеся поля используются в следующем действии. Те поля, которые остались после последнего шага, рекомендуются для моделирования; те поля, которые послужили входными для преобразованных и сконструированных предикторов, отфильтровываются.

Итоги по действиям - простая таблица, содержащая действия обработки, выполненные автоматической подготовкой данных. Там, где какое-либо **Действие** подчеркнуто, можно щелкнуть, чтобы вывести панель дополнительных подробностей о выполненных действиях. Дополнительную информацию смотрите в разделе “Свойства действия” на стр. 124.

Примечание: Для каждого поля выводится только первоначальная и окончательная версии, без промежуточных версий, которые использовались в ходе анализа.

Точность прогноза

Эта диаграмма выводится по умолчанию при первом запуске анализа или выборе **Предикторы, рекомендуемые для анализа** на основной панели Сводка по обработке полей; содержит точность прогноза рекомендуемых предикторов. Поля сортируются по точности прогноза, в начале списка - поле с самым высоким значением.

Для преобразованных версий порядковых предикторов имя поля отражает выбор суффикса на панели Имена полей на вкладке Параметры, например: *_transformed*.

Значки типа измерения выводятся после имен отдельных полей.

Точность прогноза каждого рекомендуемого предиктора непрерывного или категориального поля назначения вычисляется соответственно по линейной регрессии или по наивной байесовской модели.

Таблица полей

Таблица полей выводится при выборе **Поле назначения**, **Предикторы** или **Предикторы не использовались** на основной панели Сводка по обработке полей; соответствующие характеристики представлены в виде простой таблицы.

Таблица содержит два столбца:

- **Имя.** Имя предиктора.

  Для полей назначения это первоначальное имя или метка используемого поля, даже если поле назначения было преобразовано.

  Для преобразованных версий порядковых предикторов имя отражает выбор суффикса на панели Имена полей на вкладке Параметры, например: *_transformed*.

  Для полей, производных от дат и времени, используется имя последней преобразованной версии, например: *bdate_years*.

  Для сконструированных предикторов используется имя сконструированного предиктора, например: *Predictor1*.
- **Шкала измерения.** Содержит значок, представляющий тип данных.

  Для поля назначения **Тип измерения** всегда отражает преобразованную версию, если поле назначения было преобразовано, например, переведено из порядкового (упорядоченный набор) в непрерывное (диапазон, шкала) или наоборот.

Свойства поля

Панель Свойства поля появляется при щелчке по любому **Имени** на основной панели полей и содержит, для выбранного поля, диаграммы распределения, отсутствующих значений и точности прогноза (если применимо). Кроме того, показаны хронология обработки для поля и имя преобразованного поля (если применимо).

Для каждого набора диаграмм показаны в ряд две версии, чтобы сравнить поле после применения преобразований и до; если же преобразованной версии поля нет, диаграмма показана только для первоначальной версии. Для производных полей даты или времени и сконструированных предикторов выводятся только диаграммы нового предиктора.

Примечание: Если у поля слишком много категорий и из-за этого оно исключается, на экран выводится только хронология обработки.

Диаграмма распределения

Распределение непрерывного поля показывается как гистограмма с наложением кривой нормального распределения и вертикальной опорной линией на среднем значении; для категориальных полей выводятся полосчатые диаграммы.

Гистограммы снабжаются метками, которые показывают стандартное отклонение и асимметрию; однако асимметрия не выводится, если значений меньше трех или если первоначальное поле имеет дисперсию ниже 10-20.

Поместите указатель мыши на диаграмму, чтобы вывести среднее значение гистограммы или число и процент записей той или иной категории от общего числа записей полосчатой диаграммы.

Диаграмма пропущенных значений

На круговых диаграммах сравнивается процент отсутствующих значений при применении преобразований и без применения; процентные значения выводятся как метки диаграммы.

Кроме того, если автоматическая подготовка данных обработала отсутствующие значения, круговая диаграмма после преобразования включает в себя и такую метку, как значение замены, то есть значение, используемое вместо отсутствующих значений.

Поместите указатель мыши на диаграмму, чтобы вывести число и процент отсутствующих значений от общего числа записей.

Диаграмма Точность прогноза

Для рекомендуемых полей на полосчатых диаграммах выводится точность прогноза до и после преобразования. Если поле назначения было преобразовано, расчетная точность предсказания соотносится с преобразованным полем назначения.

Примечание: Диаграммы точности прогноза не выводятся, если поля назначения не определены или если по полю назначения щелкнули на основной панели.

Поместите указатель мыши на диаграмму, чтобы вывести значение точности прогноза.

Обработка таблицы хронологии

Таблица показывает, как была произведена преобразованная версия поля. Действия автоматической подготовки запросов перечислены в порядке их выполнения, однако для некоторых шагов многократные действия могли выполняться над одним конкретным полем.

Примечание: Эта таблица не выводится для полей, которые не были преобразованы.

Информация в таблице разбита на два или три столбца:

- **Действия.** Имя действия. Например, Непрерывные предикторы. Дополнительную информацию смотрите в разделе “Свойства действия”.
- **Подробности.** Список выполненных действий. Например, Преобразовать к стандартным единицам.
- **Функция.** Выводится только для сконструированных предикторов; содержит линейную комбинацию входных полей, например, 0,06*возраст + 1,21*высота.

Свойства действия

Дополнительная панель Свойства действия выводится при выборе любого подчеркнутого **Действия** на основной панели Итоги по действиям и содержит для каждого выполненного шага обработки общие сведения и сведения, которые выводятся только для конкретного действия; те сведения, которые выводятся только для конкретного действия, показаны в начале.

Для каждого действия как заголовок в верхней части дополнительной панели выводится описание действия. После заголовка показаны сведения, которые выводятся только для конкретного действия; они могут включать в себя подробности о числе производных предикторов, преобразованных полей, преобразованных полей назначения, объединенных и переупорядоченных категориях, а также сконструированных или исключенных предикторах.

После выполнения очередного действия число используемых в обработке предикторов может изменяться; например, предикторы могут исключаться или объединяться.

Примечание: Если действие было отключено или если не было указано поле назначения, вместо подробностей при щелчке по действию на основной панели Свойства действия выводится сообщение об ошибке.

Есть девять возможных действий, но не обязательно все они активны при анализе.

Таблица Текстовые поля

Эта таблица содержит количество:

- Значений с усеченными хвостовыми пробельными символами.
- Предикторов, исключенных из анализа.

Таблица Предикторы дат и времени

Эта таблица содержит количество:

- Интервалов, производные от предикторов даты и времени.
- Элементов даты и времени.
- Всех производных предикторов даты и времени, суммарно.

Если есть вычисленные интервалы дат, соответствующие дата или время показаны как сноска.

Таблица Скрининг предикторов

Таблица содержит количество указанных ниже предикторов, исключенных из обработки:

- Константы.
- Предикторы со слишком большим количеством пропущенных значений.
- Предикторы со слишком большим количеством случаев в одной категории.
- Номинальные поля (наборы) со слишком большим числом категорий.
- Все отвергнутые при скрининге предикторы, вместе взятые.

Таблица Проверка типа измерений

Таблица содержит, с указанным ниже разбиением, количества полей, тип которых был преобразован:

- Порядковые поля (упорядоченные наборы), преобразованные в непрерывные поля.
- Непрерывные поля, преобразованные в порядковые поля.
- Общее число преобразований типа.

Если среди входных полей (полей назначения или предикторов) не было непрерывных или порядковых, это указывается в сноске.

Таблица выбросов

Таблица содержит число случаев той или иной обработки выбросов.

- Либо число непрерывных полей, для которых были обнаружены и отсечены выбросы, либо число непрерывных полей, для которых выбросы были обнаружены и перезаданы как отсутствующие значения - в зависимости от значений на панели Подготовить входные поля и поля назначения на вкладке Параметры.
- Число непрерывных полей, исключенных, потому что после обработки выбросов эти поля стали константами.

Одна сноска содержит значение отсечения выбросов, и еще одна выводится, когда нет непрерывных входных полей (полей назначения или предикторов).

Таблица пропущенных значений

Таблица содержит, с указанным ниже разбиением, количества полей, в которых были заменены отсутствующие значения:

- Цель. Эта строка не выводится, если поле назначения не задано.
- Предикторы. Эта информация, в свою очередь, разбивается на число номинальных (наборы), порядковых (упорядоченные наборы) и непрерывных.
- Общее число замененных отсутствующих значений.

Таблица Поле назначения

В этой таблице сообщается, было ли поле назначения преобразовано; это показано так:

- Преобразование Бокса-Кокса к нормальному распределению. Эта информация, в свою очередь, разбивается на столбцы, содержащие заданные критерии (среднее и среднеквадратичное отклонение) и Лямбду.
- Целевые категории переупорядочены для повышения устойчивости.

Таблица Категориальные предикторы

Таблица содержит количество категориальных предикторов:

- У которых категории были переупорядочены от наименьшей до наибольшей, чтобы повысить стабильность.
- У которых категории были объединены, чтобы максимизировать связь с полем назначения.
- У которых категории были объединены, чтобы обработать малочисленные категории.
- Исключены, поскольку слабо связаны с целевым полем.
- Исключены, так как стали постоянными после слияния.

Если категориальных предикторов не было, это указывается в сноске.

Таблица Непрерывные предикторы

Существует две таблицы. Первая содержит один из следующих показателей числа преобразований:

- Значения предикторов, преобразованные к стандартным единицам. Кроме того, здесь показано число преобразованных предикторов, заданное среднее и среднеквадратичное отклонение.
- Значения предикторов, отображенные в общий диапазон. Кроме того, здесь показано число предикторов, преобразованных при помощи преобразования минимакса, а также заданные минимальные и максимальные значения.
- Категориальные значения предикторов и число таких значений.

Вторая таблица содержит подробную информацию о конструировании пространства предикторов с указанием, сколько предикторов:

- Построены.
- Исключены, поскольку слабо связаны с полем назначения.
- Исключены, так как стали постоянными после категоризации.
- Исключены, так как стали постоянными после построения.

Если на входе не было непрерывных предикторов, это указывается в сноске.

Генерирование узла извлечения

Когда вы генерируете узел извлечения, он применяет обратное преобразование назначения к полю оценки. По умолчанию этот узел вводит имя поля оценки, которое может быть создано автоматическим средством

моделирования (таким как автоклассификатор или автонумератор) или узлом Ансамбль. Если масштабное (диапазона) поле назначение было преобразовано, поле оценки показывается в преобразованных единицах; например, log($) вместо $. Чтобы понимать и использовать эти результаты, нужно преобразовать предсказанное значение обратно к первоначальной шкале.

Примечание: Когда узел ADP содержит результаты анализа, которые перемасштабируют поле назначения диапазона (то есть перемасштабирование Бокса-Кокса выбирается на панели Подготовка входных полей и полей назначения), можно сгенерировать только узел Извлечение. Нельзя сгенерировать узел извлечения, если поле назначения - не диапазон или если не выбрано изменение шкалы Бокса-Кокса.

Узел извлечения создается в множественном режиме и использует @FIELD в выражении, так что при необходимости вы можете выбрать преобразованное поле назначения. Для примера используются следующие подробности:

- Имя поля назначения: response
- Имя преобразованного поля назначения: response_transformed
- Имя поля назначения: $XR-response_transformed

Узел Извлечение может создать новое поле: $XR-response_transformed_inverse.

Примечание: Если вы не используете автомоделирование или узел Ансамбль, потребуется изменить узел Извлечение для преобразования правильного поля оценки в вашей модели.

Нормализованные количественные поля назначения

По умолчанию, если вы включаете переключатель **Перемасштабировать количественное поле назначения при помощи преобразования Бокса-Кокса** на панели Подготовка входных полей и полей назначения, это преобразует поле назначения и вы создаете новое поле, которое будет назначением вашего построения модели. Например, если ваше исходное поле назначения - это *response*, новым назначением будет *response_transformed*; модели уровнем ниже узла ADP будут принимать это новое поле назначения автоматически.

Однако это может вызвать некоторые проблемы, что зависит от исходного поля назначения. Например, если полем назначения было *Age*, новыми значениями назначения будут не *Years*, а преобразованная версия *Years*. Это означает, что вы не сможете просмотреть оценки и интерпретировать их, так как они не представлены в распознаваемых единицах. В этом случае можно применить обратное преобразование, так что преобразованные единицы окажутся теми, на которые вы рассчитывали. Для этого:

1. После нажатия кнопки **Анализировать данные** для запуска анализа ADP выберите *Узел извлечения* в меню *Генерировать*.
2. Разместите узел Извлечение после вашего слепка на холсте модели.

Узел Извлечение восстановит поле оценки к исходным измерениям, так что предсказания будут в исходных значениях *Years*.

По умолчанию узел Извлечение преобразует поле оценки, сгенерированное средством автомоделирования или моделью ансамбля. Если вы строите индивидуальную модель, нужно изменить узел Извлечение, чтобы извлечь данные из вашего фактического поля оценки. Если нужно оценить вашу модель, необходимо добавить преобразованное значение назначения в поле **Извлечь из** на узле Извлечение. При этом к назначению применяется то же обратное преобразование и любой узел Оценка или Анализ уровнем ниже будет правильно использовать преобразованные данные, если переключить эти узлы на использование имен полей вместо метаданных.

Если вы хотите восстановить также исходное имя, можно использовать узел Фильтр для удаления исходного поля назначения, если оно еще существует, и переименовать поля назначения и оценки.

Узел Тип

Свойства полей можно задать на узле источника или отдельном узле типа. Обработка аналогична на обоих узлах. Доступны следующие свойства:

- **Поле** Щелкните дважды кнопкой мыши по имени поля, чтобы задать метки значений и полей для данных в IBM SPSS Modeler. Например, здесь можно просмотреть поля метаданных, импортированные из IBM SPSS Statistics. Таким же образом можно создать новые метки для полей и их значений. Задаваемые здесь метки выводятся повсеместно в IBM SPSS Modeler в зависимости от вариантов, выбранных вами в диалоговом окне Свойства потока.
- **Измерение** Эта шкала измерений используется для описания характеристик данных в указанном поле. Если все подробности поля известны, оно называется **полностью инстанциированным**. Дополнительную информацию смотрите в разделе “Уровни измерения” на стр. 129.

  Примечание: Шкала измерений поля отличается от типа хранения, который указывает, хранятся ли данные как строки, целые, действительные числа, даты, время, отметки времени или списки.
- **Значения** Этот столбец позволяет задать опции для чтения значений данных из набора данных; можно также использовать опцию **Задать**, чтобы задать шкалы и значения измерений в отдельном диалоговом окне. Можно также выбрать вариант передачи полей полей без чтения их значений. Дополнительную информацию смотрите в разделе “Значения данных” на стр. 133.

  Примечание: Нельзя изменить ячейку в этом столбце, если соответствующая запись **Поле** содержит список.
- **Отсутствующие** Используется для задания способа, которым будут обрабатываться пропущенные значения для поля. Дополнительную информацию смотрите в разделе “Определение пропущенных значений” на стр. 138.

  Примечание: Нельзя изменить ячейку в этом столбце, если соответствующая запись **Поле** содержит список.
- **Проверка** В этом столбце можно задать опции, гарантирующие, что значения полей будут соответствовать заданным значениям или диапазонам. Дополнительную информацию смотрите в разделе “Проверка значений типа” на стр. 138.

  Примечание: Нельзя изменить ячейку в этом столбце, если соответствующая запись **Поле** содержит список.
- **Роль** Используется для указания узлам моделирования, будут ли поля представлять собой **Входные поля** (поля предикторов) or **Поля назначения** (предсказанные поля) для процесса машинного обучения. Доступны также роли **Двойного назначения** и **Нет**, наряду с ролью **Разделение**, указывающей поле, используемое для разделения записей на отдельные выборки для обучения, испытания и проверки. Значение **Разбиение** задает, что будет выполняться построение отдельных моделей для каждого возможного значения поля. Дополнительную информацию смотрите в разделе “Задание роли поля” на стр. 139.

Используя окно узла Тип можно задать несколько разных опций:

- Используя кнопку меню панели инструментов, можно выбрать опцию **Игнорировать уникальные поля**, если узел Тип был инстанциирован (через ваши спецификации, прочитанные значения или при запуске потока). Игнорирование уникальных полей автоматически приводит к игнорированию полей с одним значением.
- Используя кнопку меню панели инструментов, можно выбрать опцию **Игнорировать большие наборы**, если узел Тип был инстанциирован. Игнорирование больших наборов автоматически приводит к игнорированию наборов с большим числом элементов.
- Используя кнопку меню панели инструментов, можно выбрать опцию **Преобразовывать количественные целые в порядковые**, если узел Тип был инстанциирован. Дополнительную информацию смотрите в разделе “Преобразование количественных данных” на стр. 132.

- Используя кнопку меню панели инструментов, можно сгенерировать узел Фильтр для отбрасывания выбранных полей.
- Используя переключающие кнопки с темными очками, можно задать для всех полей значения по умолчанию Читать или Передать. Вкладка Типы на узле источника передает поля по умолчанию, а сам узел Тип читает значения по умолчанию.
- Используя кнопку **Очистить значения**, можно очистить изменения значений полей, выполненные на этом узле (не наследуемые значения), и повторно прочесть значения из операций более высокого уровня. Эта опция полезна для сброса изменений, которые вы могли выполнить для конкретных вышележащих полей.
- Используя кнопку **Очистить все значения** , можно сбросить значения **всех** полей, прочитанных на узел. Эта опция эффективно задает для столбца **Значение** настройку **Читать** для всех полей. Данная опция полезна для сброса значений во всех полях и повторного чтения значений и типов из операций более высокого уровня.
- Используя контекстное меню, можно выбрать **Копирование** атрибутов из одного поля в другое. Дополнительную информацию смотрите в разделе “Копирование атрибутов типа” на стр. 139.
- Используя опцию **Просмотреть параметры неиспользуемых полей**, можно просмотреть параметры типов для полей, которых больше нет в данных или которые однократно соединялись с этим узлом Тип. Это полезно при повторном использовании узла Тип для измененных наборов данных.

Уровни измерения

Уровень измерения (ранее известный как "тип данных" или "тип использования") описывает использование полей данных в IBM SPSS Modeler. Уровень измерения можно задать на вкладке Типы узла назначения или узла Тип. Например, вы можете захотеть задать шкалу измерения для целочисленного поля со значениями 1 и 0 как для *флага*. Обычно это означает, что 1 = *True*, а 0 = *False*.

Тип хранения в сравнении с уровнем измерения. Обратите внимание, что уровень измерения поля отличается от его типа хранения, который обозначает форму хранения данных - в виде строки, целых чисел, действительных чисел, даты, времени или отметки времени. В то время как типы данных могут изменяться в любой точке потока с использованием узла Тип, тип хранения должен определяться на источнике при чтении данных в IBM SPSS Modeler (хотя его можно впоследствии изменить, используя функцию преобразования). Дополнительную информацию смотрите в разделе “Задание форматирования и системы хранения для полей” на стр. 8.

Некоторые узлы моделирования обозначают разрешенные для своих входных полей и полей назначения уровни измерений с помощью значков на вкладке Поля.

Значки уровня измерения

Таблица 19. Значки уровня измерения

Значок	Шкала измерения
	По умолчанию
	Количественная
	Категориальное
	Флаг
	Номинальная
	Порядковая

Таблица 19. Значки уровня измерения (продолжение)

Значок	Шкала измерения
	Без типа
[ ]	Собрание
	Геопространственное

Доступны следующие шкалы измерений:

- **По умолчанию** Данные, тип хранения и значения которых неизвестны (например, из-за того, что они еще не были прочитаны), выводятся как данные **<По умолчанию>**.
- **Непрерывный** Используется для описания числовых значений, таких как диапазон от 0 до 100 или диапазон от 0,75 до 1,25. Непрерывное значение может быть целочисленным, действительным числом или датой-временем.
- **Категориальный** Используется для строковых значений, когда точное число различных значений неизвестно. Это **неинстацированный** тип данных, означающий, что вся возможная информация о хранении и использовании этих данных еще неизвестна. После того, как данные прочитаны, типом шкалы измерений будет *Флаг*, *Номинальный* или *Без типа*, в зависимости от максимального числа элементов для номинальных полей, заданных в диалоговом окне Свойства потока.
- **Флаг** используется для данных с двумя различными значениями, указывающими на наличие или отсутствие той или иной особенности, например: `true` и `false`, `Да` и `Нет` или `0` и `1`. Используемые значения могут изменяться, но одно всегда должно определяться как значение "true", а другое - как значение "false". Данные могут быть представлены как текст, целочисленное значение, действительное число, дата, время или отметка времени.
- **Номинальный** Используется для описания данных с несколькими отличительными значениями (каждое из которых обрабатывается как элемент набора), например: `малый/средний/большой`. У номинальных данных может быть любой тип хранения: числовой, строковый или даты-времени. Имейте в виду, что задание шкалы измерений *Номинальная* не изменяет автоматически значения на строковый тип хранения.
- **Ординальный** Используется для описания данных с несколькими отличительными значениями, у которых есть естественный порядок. Например, к порядковому типу данных можно отнести категории заработной платы или ранжирование удовлетворенности. Порядок определяется естественным порядком сортировки элементов данных. Например `1, 3, 5` - порядок сортировки по умолчанию для набора целых чисел, а `ВЫСОКИЙ, НИЗКИЙ, ОБЫЧНЫЙ` (возрастающий по алфавиту) - порядок для набора строковых значений. Порядковая шкала измерений позволяет определить набор категориальных данных в качестве порядковых для целей визуализации, построения моделей и экспорта в другие прикладные программы (такие как IBM SPSS Statistics), распознающие порядковые данные как отличительный тип. Поле порядковых данных можно использовать всюду, где может использоваться номинальное поле. Кроме того, поля с хранением любого типа (хранением действительных чисел, целочисленных значений, дат, времени и так далее) можно определить как порядковые.
- **Без типа** Используется для данных, не соответствующих никакому из вышеуказанных типов, для полей с одним значением или для номинальных данных, число элементов в наборе которых превышает заданный максимум. Эта опция полезна также там, где в противном случае шкала измерений представляла бы собой бы набор с большим числом элементов (таких как номер учетной записи). При выборе для поля **Без типа** роль автоматически задается как **Нет**, с **ID записи** в качестве единственно возможного варианта. Максимальный размер по умолчанию для наборов - 250 уникальных значений. Это число можно скорректировать или отключить на вкладке Опции диалогового окна Свойства потока, доступного в меню Инструменты.
- **Собрание** Используется для определения записываемых в список данных, отличных от геопространственных. Эффективно собрание - это поле списка нулевой глубины, где у элементов списка одна из других шкал измерений.

Дополнительную информацию о списках смотрите в теме Система хранения списков и связанные шкалы измерений раздела Узлы источников в Руководстве по узлам источников, процессов и вывода для SPSS Modeler.

- **Геопространственная** Используется с типом хранения Список для определения геопространственных данных. Списки могут быть полями Список целых чисел или Список действительных чисел с глубиной списка от нуля до двух включительно.

  Дополнительную информацию смотрите в теме Геопространственные подуровни измерений раздела Узел типов в Руководстве по узлам источников, процессов и вывода для SPSS Modeler.

Уровни измерения можно задать вручную или разрешить программам читать данные и определять уровень измерения на основе прочитанных значений.

Как вариант, если у вас есть несколько количественных полей, которые нужно рассматривать как категориальные данные, можно выбрать опцию для их преобразования. Дополнительную информацию смотрите в разделе “Преобразование количественных данных” на стр. 132.

Использовать автоматическое назначение типов

1. На узле Тип или на вкладке Типы узла источника задайте для столбца *Значения* настройку **<Читать>** для нужных полей. Это сделает метаданные доступными для всех узлов ниже уровнем. Используя кнопки с темными очками в диалоговом окне, можно быстро задать **<Читать>** или **<Передать>** для всех полей.
2. Нажмите кнопку **Читать значения**, чтобы немедленно прочесть значения со всех источников данных.

Вручную задать уровень измерения для поля

1. Выбрать поле в таблице.
2. Из раскрывающегося списка в столбце *Измерение* выберите уровень измерения для поля.
3. Как вариант, можно использовать клавиши Ctrl-A или щелкнуть по полям при нажатой клавише Ctrl, чтобы выбрать несколько полей перед использованием раскрывающегося списка для выбора уровня измерения.

Подуровни геопространственных измерений

У шкалы измерений Геопространственная, используемой с типом хранения Список, есть шесть подуровней, используемых для определения различных типов геопространственных данных.

- **Точка** - определяет конкретное положение, например, центр города.
- **Многоугольник** - ряд точек, определяющих одну границу региона и его положение, например, район.
- **Ломаная** - также называется линией, набор точек, определяющих маршрут. Например, ломаная может быть фиксированным элементом (шоссе, рекой, железной дорогой) или отслеживать путь объекта, такого как самолет или судно.
- **Несколько точек** - используется, когда в каждой строке ваших данных есть несколько точек на район. Например, если каждая строка представляет улицу города, несколько точек на каждой улице могут соответствовать фонарным столбам.
- **Мультиполигон** - используется, когда в каждой строке ваших данных есть несколько многоугольников. Например, если каждая строка представляет контуры страны, для США могут использоваться несколько многоугольников, определяющих разные части страны, а именно, основную континентальную, Аляску и Гавайи.
- **Мультиломаная** - используется, когда в каждой строке ваших данных есть несколько линий. Так как линии не могут ветвиться, мультиломаную можно использовать для определения группы линий. Примерами могут быть фарватеры навигации или железнодорожные сети в каждой стране.

Эти подуровни шкалы измерений используются с типом хранения Список. Дополнительную информацию смотрите в разделе “Система хранения списков и связанные шкалы измерений” на стр. 11.

Ограничения

При использовании геопространственных данных нужно учитывать следующие ограничения.

- Система координат может воздействовать на формат данных. Например, проективная система координат использует значения x, y и (при необходимости) z, а географическая система координат - долготу, широту и (при необходимости) высоту/глубину.

  Дополнительную информацию о системе координат смотрите в теме Определение геопространственных опций раздела Работа с потоками в Руководстве пользователя SPSS Modeler.
- Ломаная не может пересекать саму себя.
- Многоугольник не замыкается автоматически; для каждого многоугольника необходимо убедиться, что его первая и последняя точки совпадают.
- Важно направление данных в мультиполигоне; движение по часовой стрелке означает сплошную форму, а против часовой стрелки - дыру. Например, при описании области страны с озерами граница суши описывается по часовой стрелке, а каждое озеро - против часовой стрелки.
- Многоугольник не может пересекать сам себя. Примером такого самопересечения может быть рисование непрерывной линией фигуры в форме цифры 8.
- Мультиполигоны не могут перекрываться.
- Для геопространственных полей единственные релевантные типы хранения - это **Действительное** и **Целое** (параметр по умолчанию - **Действительное**).

Значки подуровней геопространственных измерений

Таблица 20. Значки подуровней геопространственных измерений

Значок	Шкала измерения
	Точки
	Многоугольник
	Ломаная
	Несколько точек
	Мультиполигон
	Мультиломаная

Преобразование количественных данных

Рассмотрение категориальных данных как количественных может существенно воздействовать на качество модели, особенно если это данные поля назначения; например, может быть создана модель регрессии вместо бинарной модели. Для предотвращения этого можно преобразовать целочисленные диапазоны в категориальные типы, такие как *Порядковый* или *Флаг*.

1. Нажмите кнопку меню Операции и Генерировать (со знаком инструмента) и выберите пункт **Преобразовать количественные целые в порядковые**. Появится диалоговое окно преобразования.
2. Задайте размер диапазона, который будет автоматически преобразован; преобразование будет применяться к любому диапазону, заключенному (включительно) в определенных вами границах.
3. Щелкните по **OK**. Затронутые диапазоны будут преобразованы к типу *Флаг* или *Порядковый* и выведены на вкладке Типы узла Тип.

Результаты преобразования

- Там где поле *Количественное* с целочисленной системой хранения изменено на *Порядковое*, минимальное и максимальное значение будут дополнены всеми целочисленными значениями между ними. Например, если диапазон лежит от 1 до 5, набором значений будет 1, 2, 3, 4, 5.
- Если поле *Непрерывное* изменяется на *Флаг*, минимальное и максимальное значение диапазона становятся значениями false и true флагового поля.

Что такое инстанциация?

Инстанциация (полное определение) - это процесс чтения или задания информации, такой как тип хранения и значения для поля данных. Для оптимизации системных ресурсов инстанциация реализуется как направляемый пользователем процесс; вы даете указания программному обеспечению на чтение значений, задавая опции на вкладке Типы узла источника или пропуская данные через узел Тип.

- Данные с неизвестными типами называются также **не инстанциированными**. Данные, тип хранения и значения которых неизвестны, выводятся в столбце *Измерение* вкладки Типы, как задано **<По умолчанию>**.
- Когда есть какая-то информация о типе хранения поля, например, о строковом или числовом, данные называются **частично инстанциированными**. **Категориальные** или **Количественные** уровни измерения - частично инстанциированные. Например, значение **Категориальное** задает, что данное поле символическое, но вы не знаете, какое оно точно - номинальное, порядковое или флаговое.
- Когда известны все подробности о типе, в том числе значения, в этом столбце выводится **полностью инстанциированный** уровень измерений - номинальный, порядковый или флаговый. *Примечание*: *Количественный* тип используется и для полностью, и для частично инстанциированных полей данных. Количественные данные могут быть целыми или действительными числами.

Во время выполнения потока данных с узлом Тип неинстанциированные данные сразу становятся частично инстанциированными на основе начальных значений данных. Поле того как все данные проходят через узел, они становятся полностью инстанциированными, кроме тех значений, для которых задано **<Передать>**. Если выполнение прервано, данные остаются частично инстанциированными. После инстанциации вкладки Типы значения поля остаются статичными в данной точке потока. Это означает, что никакие изменения на более высоком уровне не воздействуют на значения конкретного поля даже при перезапуске потока. Для изменения или обновления значений на основании новых данных или дополнительных манипуляций нужно изменить их на самой вкладке Типы или задать для значения поля **<Читать>** или **<Читать +>**.

Когда проводить инстанциацию

В общем случае, если набор данных не слишком велик и вы не собираетесь добавлять поля в поток позднее, инстанциация на узле источника будет более удобным способом. Однако инстанциация на отдельном узле Тип полезна в следующих случаях:

- Большой набор данных, поток отфильтровывает подмножество до узла Тип.
- Данные были отфильтрованы в потоке.
- Данные были слиты с потоком или присоединены к нему.
- При обработке получены новые поля данных.

Значения данных

Используя столбец **Значения** вкладки Типы, можно автоматически прочесть значения из данных или задать уровни измерения и значения в отдельном диалоговом окне.

Опции, доступные в раскрывающемся списке Значения, обеспечивают инструкции для автоматического назначения данных, как показано в следующей таблице.

Таблица 21. Инструкции для автоматического назначения данных

Опция	**Функция**
<Read>	Данные читаются при выполнении узла.

Таблица 21. Инструкции для автоматического назначения данных (продолжение)

Опция	Функция
<Read+>	Данные читаются и присоединяются к текущим данным (если они есть).
<Pass>	Данные не читаются.
<Current>	Сохранить текущие значения данных.
Задать...	Для задания значений и опций уровня измерения открывается отдельное диалоговое окно.

При выполнении узла Тип или нажатии кнопки **Читать значения** автоматически вводятся и читаются значения из ваших данных на основании сделанного выбора. Эти значения можно задать также вручную, используя опцию Задать или дважды щелкнув по ячейке в столбце **Поле**.

После того, как сделаны изменения для полей на узле Тип, вы можете сбросить информацию о значениях, используя следующие кнопки на панели инструментов диалогового окна:

- Нажав кнопку **Очистить значения**, можно очистить изменения значений полей, выполненные на этом узле (не наследуемые значения), и повторно прочесть значения из операций более высокого уровня. Эта опция полезна для сброса изменений, которые вы могли выполнить для конкретных вышележащих полей.
- Используя кнопку **Очистить все значения**, можно сбросить значения **всех** полей, прочитанных на узел. Эта опция эффективно задает для столбца *Значение* настройку **Читать** для всех полей. Данная опция полезна для сброса значений во всех полях и повторного чтения значений и типов из операций более высокого уровня.

Затененный текст в столбце Значения

Если на узле Тип или на узле Источник данные в столбце **Значения** представлены черным шрифтом, это означает, что значения данного поля были прочтены и сохранены на этом узле. Если в данном поле нет текста черным шрифтом, значения этого поля не были прочтены и будут определены в потоке после этого узла в потоке.

В некоторых случаях данные будут представлены затененным текстом. Это происходит, когда SPSS Modeler может идентифицировать или подразумевать допустимое значение поля без фактического чтения и сохранения данных. Это может произойти при использовании одного из следующих узлов:

- Узел пользовательского ввода. Так как данные определены на узле, диапазон значений для поля всегда известен, даже если эти значения не были сохранены на узле.
- Узел источника Файл статистики. Если существуют метаданные для типов данных, это позволяет SPSS Modeler сделать вывод о возможном диапазоне значений, не читая и не сохраняя данные.

На любом узле данные будут показываться затененным текстом, пока вы не щелкнете по **Читать значения**.

Прим.: Если данные в потоке не конкретизированы и значения ваших данных показаны затененными, никакая проверка значений типов, заданных в столбце **Проверка** не применяется.

Использование диалогового окна Значения

При щелчке кнопкой мыши по столбцу **Значения** или **Пропущенные** вкладки Типы выводится выпадающий список заранее определенных значений. При выборе в этом списке опции **Задать...** открывается отдельное диалоговое окно, где можно задать опции для чтения, определения, задания меток и обработки значений для выбранного поля.

Многие из этих элементов управления являются общими для всех типов данных. Здесь описаны общие элементы управления.

Измерение Выводит выбранную в текущий момент шкалу измерений. Это значение можно изменить, чтобы отразить способ, которым вы собираетесь использовать данные. Например, если поле `день_недели` содержит

числа, представляющие отдельные дни, их можно изменить на номинальные данные, чтобы создать узел распределения, исследующий каждую категорию отдельно.

Хранение Выводит тип хранения, если он известен. На типы хранения выбираемая вами шкала измерений не влияет. Для изменения типа хранения можно использовать вкладку Данные на узлах источника фиксированных файлов или файлов переменных.

Поле модели Для полей, сгенерированных в результате скоринга слепка модели, можно также просмотреть подробности поля модели. К ним относятся имя поля назначения, а также роль поля при моделировании (например, предсказанное значение, вероятность, склонность и так далее).

Значения Выберите метод определения значений для выбранного поля. Выбранные здесь опции переопределяют все опции, выбранные вами ранее в столбце **Значения** диалогового окна узла Тип. Для чтения значений можно выбрать следующие опции:

- **Читать из данных** Выберите эту опцию для чтения значений при вызове узла. Эта опция аналогична опции **<Читать>**.
- **Передавать** Выберите эту опцию, чтобы не читать данные из текущего поля. Эта опция аналогична опции **<Передать>**.
- **Задать значения и метки** Опции отсюда используются для задания значений и меток для выбранного поля. Используемая в сочетании с проверкой значений, эта опция позволяет задать значения на основе ваших знаний о текущем поле. Эта опция активирует уникальные элементы управления для каждого типа поля. Опции для значений и меток описаны по отдельности в последующих разделах.

  Примечание: Для поля со шкалой измерений Без типа или <По умолчанию> ни метки, ни значения задать нельзя.
- **Расширить значения по данным** Выберите эту опцию для добавления текущих данных с введенными здесь значениями. Например, если у `поля_1` диапазон (0,10), а вы выберете диапазон значений (8,16), диапазон будет расширен до 16, без удаления исходного минимума. Новый диапазон будет (0,16). При выборе этой опции автоматически будет задана опция автоввода как **<Читать+>**.

Максимальная длина списка Доступно только для данных со шкалой измерений Геопространственная или Собрание. Задайте максимальную длину списка, указав количество элементов, которые может содержать этот список.

Проверить значения Выберите метод принудительного преобразования типа значений для приведения в соответствие с заданными непрерывными, флаговыми или номинальными значениями. Эта опция соответствует столбцу **Проверить** в диалоговом окне узла Тип. Используемая в сочетании с опцией **Задать значения и метки**, проверка значений позволяет привести значения в данных в соответствие с ожидаемыми значениями. Например, если задать значения как 1, 0, а затем применить опцию **Отбросить**, то можно отбросить все записи, значения которых отличаются от 1 или 0.

Определить пустые значения Выберите для активирования следующих элементов управления, используемых для объявления пропущенных или пустых значений в ваших данных.

- **Пропущенные значения** Используйте эту таблицу для определения конкретных значений (таких как 99 или 0) как пустых. Значение должно быть совместимо с типом хранения поля.
- **Диапазон** Используется для задания диапазона пропущенных значений, например, возраста от 1 до 17 или старше 65. Если граничное значение оставить пустым, диапазон будет неограниченным; например, если указать нижнюю границу 100 без верхней границы, все значения от 100 и выше будут определены как пропущенные. Граничные значения включительны; например, диапазон с нижней границей 5 и верхней границей 10 содержит в своем определении 5 и 10. Диапазон пропущенных значений может быть определен для любого типа хранения, включая тип дата-время и строковый (в последнем случае для определения попадания значения в диапазон используется алфавитный порядок сортировки).
- **Пустые и пробельные значения** Можно также определить системные пустые значения (выводимые в данных как `$null$`) и пробельные значения (строковые значения без видимых символов) как пробельные символы.

Примечание: Узел Тип также обрабатывает пустые строки как пробельные значения для целей анализа, хотя они хранятся другим, внутренним способом и в определенных случаях могут обрабатываться по-другому.

Примечание: Для кодирования пробельных символов как неопределенных значений или $null$ используйте узел заполнения.

Описание В этом текстовом поле можно задать метку поля. Заданные здесь метки выводятся в разнообразных положениях, таких как диаграммы, таблицы, вывод и браузеры моделей в зависимости от вариантов, выбранных вами в диалоговом окне Свойства потока.

Задание значений и меток для количественных данных

Уровень измерений *Количественный* используется для числовых полей. Для количественных данных есть три типа хранения:

- Действительное число
- Целое число
- Дата/Время

Для изменения всех количественных данных используется одинаковое диалоговое окно; тип хранения показывается только для справки.

Задание значений

Следующие элементы управления уникальны для количественных полей и используются для задания диапазона значений:

Нижняя. Задать нижнюю границу диапазона значений.

Верхняя. Задать верхнюю границу диапазона значений.

Задание меток

Метку можно задать для любого значения в поле диапазона. Нажмите кнопку **Метки**, чтобы открыть отдельное диалоговое окно для задания меток значений.

Вспомогательное диалоговое окно Значения и метки: Нажатием кнопки **Метки** в диалоговом окне Значения для поля диапазона открывается новое диалоговое окно, в котором можно задать метки для любого значения в указанном диапазоне.

Используя *Значения* и *Метки* в этой таблице, можно определить пары значение-метка. Здесь показаны пары, определенные на текущий момент. Добавить новые пары меток можно, щелкнув в пустой ячейке и введя в ней значение и соответствующую ему метку. *Примечание*: Добавление в эту таблицу пар значение/метка значения не приводит к добавлению в поле каких-либо новых значений. Вместо этого просто создаются метаданные для значения поля.

Метки, задаваемые на узле типа, выводятся во многих местах (таких как подсказки, таблицы вывода и так далее) в зависимости от вариантов, выбранных вами в диалоговом окне свойств потока.

Задание значений и меток для номинальных и порядковых данных

Уровни измерений Номинальный (набор) и Порядковый (упорядоченный набор) обозначают, что значения данных используются дискретно, как элементы набора. Типами хранения для набора может быть строка, целые числа, действительные числа или дата/время.

Следующие элементы управления уникальны для номинальных и порядковых полей и используются для задания значений и меток:

Значения. Столбец *Значения* в таблице позволяет задать значения на основе вашей информации о текущем поле. Используя эту таблицу, можно ввести для поля ожидаемые значения и проверить согласованность набора данных с этими значениями при помощи раскрывающегося списка Проверить значения. Используя кнопку удаления и кнопки со стрелками, можно изменить существующие значения, а также переупорядочить или удалить их.

Метки. Столбец *Метки* позволяет задать метки для каждого значения в наборе. Эти метки появляются в разных положениях, например, на графиках, в таблицах, в выходных данных и в браузерах моделей в зависимости от выбора, сделанного в диалоговом окне свойств потока.

Задание значений для флага

Флаговые поля используются для вывода данных, у которых может быть два отличающихся значения. Типом хранения для флагов не может быть строка, целое число, действительное число или дата/время.

True. Задать значение флага для поля при выполнении условия.

False. Задать значение флага для поля при невыполнении условия.

Метки. Задать метки для каждого значения в флаговом поле. Эти метки появляются в разных положениях, например, на графиках, в таблицах, в выходных данных и в браузерах моделей в зависимости от выбора, сделанного в диалоговом окне свойств потока.

Задание значений для данных собрания

Поля Собрание используются для вывода данных, отличных от геопространственных, которые присутствуют в этом списке.

Для шкалы **измерений** Собрание можно задать только **Меру списка**. По умолчанию этот показатель задается с типом Без типа, но можно выбрать другое значение, чтобы задать шкалу измерений элементов в списке. Вы можете выбрать одну из следующих опций:

- Без типа
- Количественная
- Номинальная
- Порядковая
- Флаг

Задание значений для геопространственных данных

Геопространственные поля используются для вывода геопространственных данных, присутствующих в списке.

Для шкалы **измерений** Геопространственная можно задать следующие опции определения шкалы измерений элементов из списка:

Тип Выберите подшкалу измерений геопространственного поля. Доступные подуровни определяются глубиной поля списка; значения по умолчанию - это Точка (нулевая глубина), Ломаная (глубина 1) и Многоугольник (глубина 2).

Дополнительную информацию о подуровнях смотрите в разделе “Подуровни геопространственных измерений” на стр. 131.

Дополнительную информацию о глубине списков смотрите в разделе “Система хранения списков и связанные шкалы измерений” на стр. 11.

Система координат Эта опция доступна только в том случае, если вы изменили шкалу измерений на геопространственную с какой-то другой. Включите этот переключатель, чтобы применить систему координат к вашим геопространственным данным. По умолчанию показывается система координат,

заданная на панели **Инструменты > Свойства потока > Опции > Геопространственные**. Чтобы использовать другую систему координат, нажмите кнопку **Изменить** для вывода диалогового окна Выбрать систему координат и выберите нужную систему.

Дополнительную информацию о системе координат смотрите в теме Определение геопространственных опций раздела Работа с потоками в Руководстве пользователя SPSS Modeler.

Определение пропущенных значений

Столбец **Пропущенное** на вкладке Типы обозначает, определен ли для поля способ обработки пропущенных значений. Возможны следующие настройки:

Вкл (*). Обозначает, что для этого поля определена обработка пропущенных значений. Это можно сделать с использованием нижележащего узла Фильтр или через непосредственную спецификацию с использованием опции Задать (см. ниже).

Выкл. Для поля не определена обработка пропущенных значений.

Задайте. Выберите эту опцию, чтобы вывести диалоговое окно, где можно объявить явные значения, которые будут рассматриваться как значения отсутствия для этого поля.

Проверка значений типа

При включении опции Проверка для каждого поля изучаются все значения в этих полях для определения, согласуются ли они с текущими параметрами типа или со значениями, заданными в диалоговом окне Задать значения. Это полезно для очистки наборов данных и сокращения их размеров в одной операции.

Задание значения столбца *Проверка* в диалоговом окне узла Тип определяет, что происходит при обнаружении значения вне пределов типа. Чтобы изменить параметры Проверки для поля, используйте раскрывающийся список для этого поля в столбце *Проверка*. Чтобы задать параметры Проверки для всех полей, щелкните по столбцу *Поле* и нажмите Ctrl-A. Затем используйте раскрывающийся список для любого поля в столбце *Проверка*.

Доступны следующие параметры Проверки:

Нет. Значения будут передаваться без проверки. Этот параметр выбран по умолчанию.

Аннулировать. Изменить значения вне пределов на системный null (`$null$`).

Принуждать. Поля, уровни измерений которых полностью инстанциированы, будут проверяться на значения, которые попадают вне заданных диапазонов. Незаданные значения будут преобразовываться в разрешенное значение для этого уровня измерений при помощи следующих правил:

- Для флагов - любое значение, отличное от true и false, будет преобразовано в значение false.
- Для наборов (номинальных и порядковых) - любое неизвестное значение будет преобразовываться в первое из значений набора.
- Числа, большие верхнего предела диапазона, заменяются на значение верхнего предела.
- Числа, меньшие нижнего предела диапазона, заменяются на значение нижнего предела.
- Значениям null из диапазона присваивается значение средней точки для этого диапазона.

Отклонить. Если обнаруживаются запрещенные значения, отбрасывается вся запись.

Предупреждение. Количество запрещенных элементов подсчитывается и выводится в отчете в диалоговом окне свойств потока после прочтения всех данных.

Прервать. При встрече первого запрещенного значения выполнение потока прерывается. Сообщение об ошибке выводится в диалоговом окне свойств потока.

Задание роли поля

Роль поля задает, как оно используется для построения модели, например, это входное поле или поле назначения (значение которого предсказывается).

Пример: Роли Раздел, Частота и ID записи можно применить только к одному полю.

Доступны следующие роли:

Входная. Поле будет использоваться для машинного обучения как входное (предикторное поле).

Цель. Поле будет использоваться для машинного обучения как выходное, то есть поле назначения (одно из полей, значения которых будет предсказывать модель).

И те, и другие. Поле будет использоваться узлом Априори и как входное, и как выходное. Все другие узлы моделирования будут игнорировать это поле.

Нет. При машинном обучении поле будет игнорироваться. Поля с заданным уровнем измерения **Без типа** автоматически преобразуются в **Нет** в столбце *Роль*.

Подмножества. Обозначает поле, используемое для разделения данных на отдельные выборки для целей обучения, испытания и (необязательно) проверки. У этого поля должен быть задан окончательно определенный тип с двумя или тремя возможными значениями (как определено в диалоговом окне Значения поля). Первое значение представляет выборку обучения, второе - испытания, а третье (если оно есть) - проверочную выборку. Любые дополнительные значения игнорируются, и флаговые поля использовать нельзя. Обратите внимание, что для использования поля раздела при анализе разделение должно быть включено на вкладке Опции модели соответствующего узла построения моделей или анализа. При включении разделения записи со значением null для поля раздела исключаются из анализа. Если в потоке было определено несколько полей раздела, на вкладке Поля каждого применимого узла моделирования должно быть задано одно поле раздела. Если в ваших данных еще нет подходящего поля, его можно создать на узле Раздел или на узле Получение. Дополнительную информацию смотрите в разделе “Узел раздела” на стр. 167.

Расщепление. (Только для номинальных, порядковых и флаговых полей) Задает, что для каждого возможного значения этого поля должна быть создана модель.

Частота. (Только для числовых полей) При задании этой роли значение поля может использоваться для записи как фактор взвешивания по частоте. Эта возможность поддерживается только моделями дерева C&R, CHAID, QUEST и линейными моделями; все остальные узлы эту роль игнорируют. Взвешивание по частоте включается посредством опции **Использовать вес по частоте** на вкладке Поля узлов моделирования, поддерживающих данную возможность.

ID записи. Это поле будет использоваться как уникальный идентификатор записи. Большинством узлов данная возможность игнорируется; однако она поддерживается линейными моделями и обязательна для узлов исследования данных IBM Netezza In-database.

Копирование атрибутов типа

Атрибуты типа, такие как значения, опции проверки и пропущенные значения, можно легко скопировать из одного поля в другое:

1. Щелкните правой кнопкой мыши по полю, атрибуты которого вы хотите скопировать.
2. В контекстном меню выберите **Копировать**.
3. Щелкните правой кнопкой мыши по одному или нескольким полям, атрибуты которых вы хотите изменить.
4. В контекстном меню выберите **Специальная вставка**. *Примечание*: Можно выбрать несколько полей при помощи клавиши Ctrl+щелчок мыши или опции **Выбрать поля** в контекстном меню.

Откроется новое диалоговое окно, где можно выбрать конкретные атрибуты, которые вы хотите вставить. При вставке в несколько полей выбранные вами здесь опции будут применены ко всем полям назначения.

Вставить следующие атрибуты. Выберите в списке ниже атрибуты для вставки из одного поля в другое.

- **Тип.** Выберите, чтобы вставить шкалу измерений.
- **Значения.** Выберите, чтобы вставить значения поля.
- **Пропущенные.** Выберите, чтобы вставить параметры обработки пропущенных значений.
- **Проверить.** Выберите, чтобы вставить опции проверки значений.
- **Роль.** Выберите, чтобы вставить роль поля.

Вкладка Параметры форматов полей

На вкладке Формат узлов Таблица и Тип выводится список текущих и неиспользуемых полей и опций форматирования для каждого поля. Ниже приведены описания каждого из столбцов в таблице форматирования полей:

Поле. В этом столбце выводится имя выбранного поля.

Формат. Щелкнув дважды по ячейке в этом столбце, можно задать форматирование для полей на индивидуальной основе при помощи диалогового окна, которое откроется. Дополнительную информацию смотрите в разделе “Задание опций форматирования полей” на стр. 141. Заданное здесь форматирование переопределяет форматирование, заданное в общих свойствах потока.

Примечание: узлы экспорта Statistics и вывода Statistics экспортируют файлы *.sav*, в которые включается форматирование на уровне отдельных полей. Если задаваемый формат на уровне отдельных полей не поддерживается форматом файла *.sav* IBM SPSS Statistics, узел будет использовать формат IBM SPSS Statistics по умолчанию.

Выравнивание. Укажите при помощи этого столбца, каким образом должны выравниваться значения в столбце таблицы. Значение этой опции по умолчанию - **Авто**, выравнивающее символические значения по левому краю, а числовые значения по правому краю. Это поведение по умолчанию можно переопределить, выбрав **По левому краю**, **По правому краю** или **По центру**.

Ширина столбца. По умолчанию ширина столбцов вычисляется автоматически на основе значений поля. Чтобы переопределить автоматическое определение ширины, щелкните в ячейке таблицы и при помощи выпадающего списка выберите новую ширину. Чтобы ввести пользовательскую ширину, здесь не указанную, откройте вспомогательное диалоговое окно Форматы полей, щелкнув дважды в ячейке таблицы в столбце Поле или Формат. Можно также щелкнуть дважды кнопкой мыши в ячейке и выбрать **Задать формат**.

Просмотр текущих полей. По умолчанию в этом диалоговом окне выводится список текущих активных полей. Чтобы просмотреть список неиспользуемых полей, выберите **Просмотреть параметры неиспользуемых полей**.

Контекстное меню. Контекстное меню для этой вкладки предоставляет различные опции вариантов выбора и изменения значений параметров. Щелкните дважды кнопкой мыши в столбце, чтобы вывести это меню.

- **Выделить все.** Выбирает все поля.
- **Не выделять.** Сбрасывает выбор.
- **Выбрать поля.** Выбирает поля на основе характеристик типа или хранения. Есть опции **Выбрать категориальные**, **Выбрать непрерывные** (числовые), **Выбрать без типа**, **Выбрать строковые**, **Выбрать числа** и **Выбрать дату-время**. Дополнительную информацию смотрите в разделе “Уровни измерения” на стр. 129.
- **Задать формат.** Открывает вспомогательное диалоговое окно для задания опций дат, времени и десятичных чисел на уровне отдельных полей.

- **Задать выравнивание.** Задает выравнивание для выбранных полей. есть опции **Авто**, **По центру**, **По левому краю** и **По правому краю**.
- **Задать ширину столбца.** Задает ширину полей выбранных полей. Задайте **Авто**, чтобы читать ширину из данных. Можно также задать ширину полей 5, 10, 20, 30, 50, 100 или 200.

Задание опций форматирования полей

Форматирование полей задается во вспомогательном диалоговом окне, доступном с вкладки Формат на узлах Тип и Таблица. Если перед открытием этого узла выбрано несколько полей, для всех них используются значения параметров из первого выбранного поля. При нажатии кнопки **OK** после задания здесь спецификаций эти параметры будут применены ко всем полям, выбранным на вкладке Формат.

На уровне отдельных полей доступны следующие опции. Многие из них можно также задать в диалоговом окне свойств потока. Любые значения параметров, заданные на уровне поля, переопределяют значения по умолчанию, заданные для потока.

Формат даты. Выберите формат даты, чтоб использовать его для полей хранения дат или в случае, когда строки интерпретируются функциями дат CLEM как даты.

Формат времени. Выберите формат времени, чтоб использовать его для полей хранения времени или в случае, когда строки интерпретируются функциями времени CLEM как время.

Формат вывода чисел. Можно выбрать стандартный формат вывода (**####.###**), экспоненциальный (**#.###E+##**) или формат вывода валюты (**$###.##**).

Десятичный разделитель. Выберите запятую (,) точку (.) в качестве десятичного разделителя.

Символ группировки. Для форматов вывода чисел выберите символ, используемый для группирования значений (такой как запятая в значении *3,000.00*). Доступны следующие опции: ничего, точка, запятая, пробел и символ, определяемый языковым стандартом (в этом случае будет использоваться символ по умолчанию для текущей локали).

Десятичные знаки (стандартные, экспоненциальные, валюты, экспорта). Для форматов вывода чисел задает число десятичных разрядов, которые будут использоваться при выводе действительных чисел. Эта опция задается отдельно для каждого формата вывода.

Выравнивание. Задает, каким способом должны выравниваться значения в столбце. Значение этой опции по умолчанию - **Авто**, выравнивающее символические значения по левому краю, а числовые значения по правому краю. Это поведение по умолчанию можно переопределить, выбрав По левому краю, По правому краю или По центру.

Ширина столбца. По умолчанию ширина столбцов вычисляется автоматически на основе значений поля. Можно задать пользовательскую ширину столбцов в интервалах по пять при помощи кнопок со стрелками справа от окна списка.

Фильтрация или переименование полей

Поля можно переименовать или исключить в любой точке в потоке. Например, как медик-исследователь, возможно, вы не исследуете уровень калия (данные на уровне полей) пациентов (данные на уровне записей); поэтому вы можете отфильтровать поле *K* (калий). Это можно сделать при помощи узла фильтра или вкладки Фильтр на узле источника или вывода. Функции будут одинаковы независимо от того, с какого узла к ним обратиться.

- С узлов источников, таких как Файл переменных, Фиксированный файл, Файл Statistics или XML, можно переименовывать или фильтровать поля по мере считывания данных в IBM SPSS Modeler.
- При помощи узла фильтра можно переименовать или отфильтровать поля в любой точке в потоке.

- С узлов экспорта Statistics, преобразования Statistics, модели Statistics и вывода Statistics можно отфильтровать или переименовать поля для согласования со стандартами именования IBM SPSS Statistics. Дополнительную информацию смотрите в разделе “Переименование или фильтрация полей для IBM SPSS Statistics” на стр. 361.
- При помощи вкладки Фильтр на любом из указанных выше узлов можно определить или отредактировать наборы множественных ответов. Дополнительную информацию смотрите в разделе “Редактирование наборов множественных ответов” на стр. 143.
- И наконец, при помощи узла фильтра вы можете отобразить поля с одного узла источника на другой.

Задание опций фильтрации

В таблице, используемой на вкладке Фильтр, для каждого поля выводится имя в том виде, в котором оно поступает на узел, и в том виде, в котором оно его покидает. Опции в этой таблице позволяют переименовать или отфильтровать поля, которые повторяются или не требуются для операций нисходящего потока.

- **Поле.** Выводит входные поля из подключенных в текущий момент источников данных.
- **Фильтр.** Выводит состояние фильтров всех входных полей. Отфильтрованные поля содержат в этом столбце красный крестик (X), указывающий, что это поле не будет передано вниз по потоку. Щелкните кнопкой мыши в столбце *Фильтр* для выбранного поля, чтобы включить или выключить фильтрацию. Можно также выбрать опции для нескольких полей одновременно с помощью метода выбора щелчком при нажатой клавише Shift.
- **Поле.** Выводит поля в виде, в котором они покидают узел. Дубликаты имен выводятся красным. Имена полей можно отредактировать, щелкнув в этом столбце и введя новое имя. Можно также, щелкнув в столбце *Фильтр*, отключить дубликаты полей.

Все столбцы в таблице можно отсортировать, щелкнув по заголовку столбца.

Просмотр текущих полей. Выберите эту опцию для просмотра полей наборов данных, соединение которых с узлом фильтра активно. Эта опция выбирается по умолчанию и представляет собой самый распространенный способ использования узлов фильтра.

Просмотр неиспользуемых параметров полей. Выберите эту опцию для просмотра полей наборов данных, которые были соединены, но более не соединены с узлом фильтра. Эта опция полезна при копировании узлов фильтра из одного потока в другой или при сохранении узлов фильтра и их перезагрузке.

Меню кнопок фильтра

Нажмите кнопку Фильтр в правом верхнем углу диалогового окна, чтобы обратиться к меню, предоставляющему ряд ярлыков и других опций.

Вы можете:

- Удалить все поля.
- Включить все поля.
- Переключить все поля.
- Удалить дубликаты. *Примечание*: Выбор этой опции удаляет все вхождения повторяющегося имени, включая первое.
- Удалить поля и наборы множественных ответов для согласования с другими прикладными программами. Дополнительную информацию смотрите в разделе “Переименование или фильтрация полей для IBM SPSS Statistics” на стр. 361.
- Усечь имена полей.
- Анонимизировать имена полей и наборов множественных ответов.
- Использовать имена входных полей.

- Отредактировать наборы множественных ответов. Дополнительную информацию смотрите в разделе “Редактирование наборов множественных ответов”.
- Задать состояние фильтра по умолчанию.

Можно также при помощи переключающих кнопок со стрелками в верхней части диалогового окна указать, хотите ли вы включить или отбросить поля по умолчанию. Эта возможность полезна для больших наборов данных, где всего лишь несколько полей нужно будет включить в нисходящий поток. Например, можно выбрать только те поля, которые вы хотите сохранить, и указать, что все остальные поля следует отбросить (вместо того, чтобы выбирать все поля для отбрасывания по отдельности).

Усечение имен полей

В меню кнопок Фильтр (в верхнем левом углу вкладки Фильтр) можно выбрать усечение имен полей.

Максимальная длина. Задайте число символов, чтобы ограничить длину имен полей.

Количество символов. Неуникальные после усечения имена полей будут усечены дополнительно и сделаны однозначными посредством добавления цифр к имени. Можно задать число используемых цифр. Для настройки этого числа используйте кнопки со стрелками.

Например, в следующей таблице показано, как имена полей в наборе медицинских данных усекаются при помощи параметров по умолчанию (максимальная длина = 8 и число цифр = 2).

Таблица 22. Усечение имен полей

Имена полей	Усеченные имена полей
Patient Input 1	Patien01
Patient Input 2	Patien02
Heart Rate	HeartRat
BP	BP

Анонимизация имен полей

Анонимизация имен полей возможна с любого узла, содержащего вкладку Фильтр; щелкните по меню кнопок Фильтр в верхнем левом углу и выберите **Анонимизировать имена полей**. Анонимизированные имена полей состоят из префикса плюс уникальное числовое значение.

Анонимизировать имена. Выберите **Только выбранные поля**, чтобы анонимизировать только имена полей, уже выбранных на вкладке Фильтр. Опция по умолчанию - **Все поля**, анонимизирующая все имена полей.

Префикс имен полей. Для анонимизированных имен полей используется префикс по умолчанию **anon_**; выберите **Пользовательский** и введите свой собственный префикс, если хотите использовать другой.

Анонимизировать наборы множественных ответов. Анонимизирует наборы множественных ответов тем же способом, что и поля. Дополнительную информацию смотрите в разделе “Редактирование наборов множественных ответов”.

Чтобы восстановить исходные имена полей, в меню кнопок Фильтр выберите **Использовать имена входных полей**.

Редактирование наборов множественных ответов

Наборы множественных ответов можно добавить или отредактировать с любого узла, содержащего вкладку Фильтр; щелкните по меню кнопок Фильтр в верхнем левом углу и выберите **Редактировать наборы множественных ответов**.

Наборы множественных ответов используются для записи данных, содержащих несколько значений для каждого наблюдения, например, при вопросе к респондентам опроса, какие музеи они посетили или какие

журналы они читают. Наборы множественных ответов можно импортировать в IBM SPSS Modeler при помощи узла источника Data Collection или узла источника Statistics File, а затем определить в IBM SPSS Modeler при помощи узла Фильтр.

Нажмите кнопку **Новый**, чтобы создать новый набор множественных ответов, или кнопку **Правка**, чтобы изменить существующий набор.

Имя и метка. Задает имя и описание для набора.

Тип. Обработка наборов множественных ответов возможна одним из двух способов:

- **Дихотомический множественный набор.** Для каждого возможного набора создается поле флага; таким образом, при наличии 10 журналов создается 10 полей флагов, у каждого из которых могут быть значения, такие как 0 или 1 для *true* или *false*. Подсчитываемое значение позволяет указать, какое значение будет подсчитываться как true. Этот метод полезен, если вы хотите разрешить респондентам выбирать все применяемые опции.
- **Категориальный множественный набор.** Для каждого ответа создается по одному номинальному полю, до максимального числа ответов от данного респондента. Для каждого номинального поля предусмотрены значения, представляющие возможные ответы, например: *1* для *Time*, *2* для *Newsweek*, и *3* для *PC Week*. Этот метод наиболее полезен, если вы хотите ограничить число ответов, например, когда респондентов просят указать три журнала, читаемые ими чаще всего.

Поля в наборе. Значки в правой части окна позволяют добавить или удалить поля.

Комментарии

- У всех полей, включаемых в набор множественных ответов, должно быть одно и то же хранение.
- Наборы отличимы от включаемых в них полей. Например, удаление набора не приведет к удалению включенных в него полей; будут удалены всего лишь связи между ними. Набор останется видим из точки удаления вверх по поту, но станет невидим вниз по потоку.
- В случае переименования полей при помощи узла Фильтр (непосредственно на вкладке или посредством выбора Переименовать для опций IBM SPSS Statistics, **Усечение** или **Анонимизировать** в меню Фильтр) все ссылки на эти поля, используемые в наборах множественных ответов, будут также изменены. Однако никакие поля в наборе множественных ответов, отброшенные узлом Фильтр, из набора множественных ответов удалены не будут. На такие поля, хотя они больше и не будут видимы в потоке, все еще будет ссылаться набор множественных ответов, на что можно обратить внимание, например, при экспорте.

Узел извлечения

Одна из самых мощных функциональных возможностей в IBM SPSS Modeler - это возможность изменять значения данных и получать новые поля из существующих данных. В течение продолжительных проектов исследования данных часто бывает нужно выполнить несколько операций по извлечению, таких как извлечение ID покупателя из строки данных веб-журнала или создание значения жизненного цикла пользователя на основе данных транзакций и демографических данных. Все эти преобразования можно выполнить при помощи ряда узлов операций с полями.

Несколько узлов поддерживают возможность излечения новых полей:

Узел извлечения изменяет значения данных или создает новые поля из одного или нескольких существующих полей. Он создает поля формулы типа, флага, номинала, состояния, количества и условного выражения.

Узел переклассификации преобразует один набор категориальных значений в другой. Переклассификация полезна для сворачивания категорий или для перегруппировки данных для анализа.

Узел разделения на интервалы автоматически создает новые номинальные поля на основе значений одного или нескольких существующих количественных полей (числового диапазона). Например, можно преобразовать количественное входное поле в новое категориальное поле, содержащее группы входных данных, как отклонения от среднего. После создания интервалов для нового поля вы можете сгенерировать узел извлечения на основе точек деления.

Узел Задать как флаг извлекает несколько полей флагов на основании категориальных значений, определенных для одного или нескольких номинальных полей.

Узел реструктуризации преобразует номинальное или флаговое поле в группу полей, которые можно заполнить значениями еще одного поля. Например, если задано поле с именем *тип_платежа*, у которого могут быть значения *кредит*, *наличные* и *дебет*, могут быть заданы три новые поля (*кредит*, *наличные*, *дебет*), каждое из которых может содержать значение фактического выполненного платежа.

Узел Хронология создает новые поля, содержащие данные из полей в предыдущих записях. Хронологические узлы чаще всего используются для последовательных данных, таких как данные временных рядов. Перед использованием узла Хронология может потребоваться отсортировать данные с использованием узла Сортировка.

Использование узла извлечения

При помощи узла извлечения можно создать поля шести типов из одного или нескольких существующих полей.

- **Формула.** Новое поле будет результатом произвольного выражения CLEM.
- **Флаг.** Новое поле будет флаговым, представляющим заданное условие.
- **Номинальное.** Новое поле будет номинальным, означающим, что все элементы будут группой заданных значений.
- **Состояние.** Новое поле будет представлять одно из двух состояний. Переключение между этими состояниями будет инициироваться заданным условием.
- **Частота.** Новое поле будет основано на числе имевших место соблюдений условия.
- **Условие.** Новое поле будет представлять значение одного из двух выражений, в зависимости от значения условия.

Каждый из этих узлов содержит набор специальных опций в диалоговом окне узла извлечения. Эти опции описаны в последующих темах.

Задание базовых опций для узла извлечения

В верхней части диалогового окна для узлов извлечения выводится есть ряд опций для выбора нужного типа узла извлечения.

Режим. Выберите **Одно** или **Несколько**, в зависимости от того, хотите ли вы получить одно или несколько полей. Если выбрать**Несколько** диалоговое окно изменится; в него будут включены опции для нескольких полей извлечения.

Производное поле. Для простых узлов извлечения задайте имя поля, которое вы хотите извлечь и добавить в каждую запись. Имя по умолчанию - Derive*N*, где *N* - число узлов извлечения, созданных вами к этому моменту во время текущего сеанса.

Вывести как. Выберите в выпадающем списке тип узла извлечения, такой как Формула или Номинальное. Для каждого типа создается новое поле на основе условий, заданных вами в диалоговом окне, представляющем конкретный тип.

После выбора опции в выпадающем списке в главное диалоговое окно добавляется новый набор опций в соответствии со свойствами каждого типа узла извлечения.

Тип поля. Выберите для вновь полученного узла извлечения шкалу измерений, например, непрерывную категориальную или флаговую. Эта опция обычна для всех форм узлов извлечения.

Примечание: Для извлечения новых полей часто требуется использование специальных функций или математических выражений. Создать эти выражения помогает доступный в диалоговом окне для всех типов узлов извлечения Построитель выражений, обеспечивающий проверку правил и предоставляющий полный список выражений CLEM.

Извлечение нескольких полей

Задание на узле извлечения режима **Несколько** позволяет извлечь на одном узле сразу несколько полей на основе одного и того же условия. Эта возможность экономит время, если требуется выполнить одинаковые преобразования для нескольких полей в наборе данных. Например, если вы хотите построить регрессионную модель, предсказывающую текущую заработную плату на основе начальной заработной платы и послужного списка, может оказаться выгодным применить логарифмическое преобразование ко всем этим трем несимметричным переменным. Вместо того, чтобы добавлять новый узел извлечения для каждого преобразования, одну м ту же функцию можно применить сразу ко всем полям. Просто выберите все поля, из которых следует получить новое поле, а затем введите выражение извлечения при помощи функции `@FIELD` в круглых скобках поля.

Примечание: Функция `@FIELD` представляет собой важный инструмент для извлечения сразу нескольких полей. Она позволяет ссылаться на содержание одного или нескольких текущих полей, не указывая точное имя поля. Например, выражение CLEM, используемое для применения логарифмического преобразования к нескольким полям, выглядит так: `log(@FIELD)`.

При выборе режима **Несколько** в диалоговое окно добавляются следующие опции:

Производное от. При помощи средства выбора полей выберите поля, из которых нужно получить новые поля. Для каждого выбранного поля будет сгенерировано по одному выходному полю. *Примечание*: Выбранные поля не обязательно должны быть одного типа хранения, однако если заданное условие будет допустимо не для *всех* полей, операция извлечения завершится неудачно,

Расширение имени поля. Введите расширение, которое вы хотели бы добавить к имени нового поля. Например, к имени нового поля, содержащего логарифм *текущей заработной платы*, можно добавить расширение *log_*, образовав имя *log_Текущая заработная плата*. При помощи радиокнопок выберите, добавить ли расширение как префикс (в начало) или как суффикс (в конец) имени поля. Имя по умолчанию - Derive*N*, где *N* - число узлов извлечения, созданных вами к этому моменту во время текущего сеанса.

Как и в случае одномодового узла извлечения, теперь требуется создать выражение, которое будет использоваться для извлечения нового поля. В зависимости от типа выбранной операции извлечения, предусмотрен ряд опций для создания условия. Эти опции описаны в последующих темах. Чтобы создать выражение, его можно просто ввести в поле формулы либо 8 применить Построитель выражений, нажав кнопку калькулятора. Обязательно используйте функцию `@FIELD` при обращении к операциям над несколькими полями.

Выбор нескольких полей

Для всех улов, выполняющих операции с несколькими входными полями, таких как узлы извлечения (множественный режим), агрегации, сортировки, множественных зависимостей и зависимостей от времени, можно простым способом выбрать несколько полей при помощи диалогового окна Выбрать поля.

Сортировать по. Доступные поля можно отсортировать для просмотра, выбрав одну из следующих опций:

- **Естественный.** Просмотр полей в порядке их поступления через поток данных на текущий узел.
- **Имя.** Использовать алфавитный порядок сортировки полей для просмотра.
- **Тип.** просмотр полей, отсортированных по их шкале измерений. Эта опция полезна при выборе полей с конкретной шкалой измерений.

Выберите поля в списке по одному или сразу несколько, удерживая при их выборе нажатой клавишу Shift или Ctrl. При помощи опций под этим списком можно также выбрать группы полей на основе их шкалы измерений, а также выбрать сразу все поля или отменить выбор всех полей в таблице.

Задание опций формулы извлечения

Узлы Формула извлечения создают новое поле для каждой записи в наборе данных на основе результатов выражения CLEM. Это выражение не может быть условным. Для получения значений на основе условного выражения используйте узел извлечения флагового или условного типа.

Формула При помощи языка CLEM задайте формулу для получения значения для нового поля.

Примечание: Так как у SPSS Modeler нет сведений, какая шкала будет использоваться для поля извлеченного списка, для шкал измерения Собрание и Геопространственная можно выбрать опцию **Задать...**, чтобы открыть диалоговое окно Значения и задать нужную шкалу. Дополнительную информацию смотрите в разделе “Задание значений извлеченного списка”.

Для геопространственных полей единственные релевантные типы хранения - это **Действительное** и **Целое** (параметр по умолчанию - **Действительное**).

Задание значений извлеченного списка

Диалоговое окно Значения выводится после выбора опции **Задать...** в выпадающем списке Формула **Тип поля** узла извлечения. В этом диалоговом окне задаются значения шкал, которые будут использоваться для шкал измерения Формулы **Тип поля**, Собрание или Геопространственная.

Измерение Выберите **Собрание** или **Геопространственная**. Если выбрать другую шкалу измерений, в диалоговом окне появится сообщение, что нет доступных для изменений значений.

Собрание

Для шкалы **измерений** Собрание можно задать только **Меру списка**. По умолчанию для этой меры задается значение Без типа, но можно выбрать другое значение, чтобы задать шкалу измерений элементов в списке. Вы можете выбрать одну из следующих опций:

- Без типа
- Категориальная
- Количественная
- Номинальная
- Порядковая
- Флаг

Геопространственное

Для геопространственной шкалы **Измерений** можно выбрать следующие опции, чтобы задать шкалу измерений элементов из списка:

Тип Выберите подшкалу измерений геопространственного поля. Доступные подшкалы определяются глубиной поля списка; значения по умолчанию:

- Точка (глубина 0)
- Ломаная (глубина 1)
- Многоугольник (глубина 1)
- Несколько точек (глубина 1)
- Мультиломаная (глубина 2)
- Мультиполигон (глубина 2)

Дополнительную информацию о подшкалах смотрите в теме Подшкалы геопространственных измерений раздела Узел типов в Руководстве по узлам источников, процессов и вывода для SPSS Modeler.

Дополнительную информацию о глубине списков смотрите в теме Система хранения списков и связанные шкалы измерений SPSS Modeler.

Система координат Эта опция доступна только в том случае, если вы изменили шкалу измерений на геопространственную с какой-то другой. Включите этот переключатель, чтобы применить систему координат к вашим геопространственным данным. По умолчанию показывается система координат, заданная на панели **Инструменты > Свойства потока > Опции > Геопространственные**. Чтобы использовать другую систему координат, нажмите кнопку **Изменить** для вывода диалогового окна Выбрать систему координат и выберите систему, соответствующую вашим данным.

Дополнительную информацию о системе координат смотрите в теме Определение геопространственных опций раздела Работа с потоками в Руководстве пользователя SPSS Modeler.

Извлечение геопространственного поля или поля списка

Бывают случаи, когда данные, которые должны записываться как элемент списка, импортируются в SPSS Modeler в неправильном формате. Например, как отдельные геопространственные поля, такие как поля координаты x и координаты y или поля широты и долготы в виде отдельных строк в файле `.csv`. В таком случае необходимо объединить отдельные поля в одно поле списка; одна из возможностей сделать это - воспользоваться узлом извлечения.

Примечание: При объединении геопространственных данных необходимо знать, какое из полей представляет собой поле координаты x (или долготы), какое - координаты y (или широты). При комбинировании ваших данных необходимо обеспечить их упорядочение в полученном поле списка, как это принято стандартами для геопространственных координат: [x, y] или [долгота, широта].

Описанные ниже шаги показывают простой пример извлечения поля списка.

1. В вашем потоке присоедините узел извлечения к узлу источника.
2. На вкладке Параметры узла извлечения выберите опцию **Формула** из списка **Извлечь как**.
3. Для опции **Тип поля** выберите или значение **Собрание** в случае списка, отличного от геопространственного, или **Геопространственный**. По умолчанию SPSS Modeler использует подход "наиболее вероятного предположения", чтобы задать правильные подробности списка; можно выбрать опцию **Задать...**, чтобы открыть диалоговое окно Значения. В этом диалоговом окне можно ввести дополнительную информацию о данных в вашем списке; например, для геопространственного списка можно изменить шкалу измерений.
4. На панели **Формула** введите формулу для объединения ваших данных в список правильного формата. Другой вариант - нажать кнопку калькулятора, чтобы открыть Построитель выражений.

Простой пример формулы для извлечения списка - это [x, y], где x и y - отдельные поля в источнике данных. Новое создаваемое извлеченное поле - это список, где значение для каждой записи получено конкатенацией значений x и y для этой записи.

Примечание: У объединенных таким образом в список полей должен быть одинаковый тип хранения.

Дополнительную информацию о списках и их глубине смотрите в разделе “Система хранения списков и связанные шкалы измерений” на стр. 11.

Задание опций флагов извлечения

Узлы флагов извлечения используются для указания конкретного условия, такого как артериальное давление или неактивность учетной записи покупателя. Для каждой записи создается поле флага, и при выполнении условия значение флага добавляется в это поле.

Значение True. Задайте значение, которое будет включаться в поле флага для записей, соответствующих заданному ниже условию. Значение по умолчанию - T.

Значение False. Задайте значение, которое будет включаться в поле флага для записей, *не* соответствующих заданному ниже условию. Значение по умолчанию - F.

True, когда. Задайте условие CLEM для оценки определенных значений каждой записи и присвоения записи значения true или false (определенному выше). Имейте в виду, что значение true будет присваиваться записям в случае числовых значений non-false.

Примечание: Для возврата пустой строки нужно ввести открывающую и закрывающую кавычки без всяких символов между ними: "". Пустые строки часто используются в качестве значений false, позволяющих более четко выделять значения true в таблице. Таким же образом, если требуется строчное значение, следует использовать кавычки; в противном случае оно будет обработано как число.

Пример

В выпусках IBM SPSS Modeler до выпуска 12.0 множественные ответы импортировались в одно поле, со значениями, разделяемыми запятыми. Например:

```
museum_of_design,institute_of_textiles_and_fashion
museum_of_design
archeological_museum
$null$
national_art_gallery,national_museum_of_science,other
```

Чтобы подготовить эти данные к анализу, при помощи функции `hassubstring` можно сгенерировать отдельное поле флага для каждого ответа , применив, например, такое выражение:

```
hassubstring(museums,"museum_of_design")
```

Задание номинальных опций извлечения

Номинальные узлы извлечения позволяют, проверив набор условий CLEM, определить, какому условию соответствует каждая запись. При выполнении условия для каждой записи в новое полученное поле будет добавляться значение (указывающее, какой набор условий был выполнен).

Значение по умолчанию. Задайте значение для использования в новом поле, если ни одно из условий не выполняется.

Задать для поля. Задайте значение для ввода в новом поле при выполнении для него конкретного условия. С каждым значением в списке связано условие, задаваемое вами в смежном столбце.

Если выполняется условие. Задайте условие для каждого элемента в поле набора, которое будет выводиться в списке. Выберите его при помощи Построителя выражений в списке доступных функций и полей. Кнопка со стрелкой и кнопка удаления позволяют переупорядочить или удалить условия.

Условие работает посредством проверки значений конкретного поля в наборе данных. При проверке каждого условия значения, заданные выше, будут назначаться в новое поле, указывая, какое условие (если оно есть) выполняется. Если ни одно из условий не выполняется, будет использоваться значение по умолчанию.

Задание опций состояния извлечения

Узлы состояния извлечения отчасти похожи на узлы флагов извлечения. Узел флагов задает значения, зависящие от выполнения *одного* условия для текущей записи, а узел состояния извлечения может изменять значения поля в зависимости от того, каким образом для него выполняются *два независимых* условия. Это означает, что значение будет изменяться (включаться или выключаться) при выполнении каждого условия.

Начальное состояние. Выберите, присваивать ли исходно каждой записи нового поля значение **On** (включить) или **Off** (выключить). Имейте в виду, что это значение может изменяться в случае выполнения каждого из условий.

Значение "On". Задайте значение для нового поля в случае выполнения условия On.

Переключать на "On", когда. Задайте условие CLEM, которое будет изменять состояние на On в случае выполнения условия (true). Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений.

Значение "Off". Задайте значение для нового поля в случае выполнения условия Off.

Переключать на "Off", когда. Задайте условие CLEM, которое будет изменять состояние на Off в случае невыполнения условия (false). Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений.

Примечание: Для задания пустой строки нужно ввести открывающую и закрывающую кавычки без всяких символов между ними: "". Таким же образом, если требуется строчное значение, следует использовать кавычки; в противном случае оно будет обработано как число.

Задание опций счета извлечения

Узел счета извлечения используется для применения ряда условий к значениям числового поля в наборе данных. При соблюдении каждого условия значение полученного поля счета будет увеличиваться на указанное приращение. Этот тип узла извлечения полезен для данных временных рядов.

Начальное значение. Задает значение, которое будет использоваться при выполнении для нового поля. Начальное значение должно быть числовой константой. Для увеличения или уменьшения этого значения используйте кнопки со стрелками.

Увеличивать, когда. Задайте условие CLEM, которое, когда оно выполняется, будет изменять полученное значение на основе числа, заданного в опции Приращение по. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений.

Приращение по. Задайте значение, на которое будет увеличиваться счет. Можно использовать либо числовую константу, либо результат выражения CLEM.

Сбросить, когда. Задайте условие, которое, когда оно выполняется, будет сбрасывать полученное значение до начального значения. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений.

Задание условных опций извлечения

Условные узлы извлечения используют ряд операторов (If-Then-Else) для получения значения нового поля.

If. Задайте условие CLEM, которое будет оцениваться для каждой записи после выполнения. Если условие оценивается как true или (в случае чисел) non-false, новому полю присваивается значение, задаваемое ниже выражением Then. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений.

Then. Задайте для нового поля значение или выражение CLEM, если оператор If выше оценивается как true (или non-false). Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений.

Else. Задайте для нового поля значение или выражение CLEM, если оператор If выше оценивается как false. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений.

Перекодирование значений при помощи узла извлечения

Узлы извлечения можно также использовать для перекодирования значений, например, посредством преобразования строкового поля с категориальными значениями в числовое номинальное поле (набора).

1. Для опции Извлечь как выберите тип поля (номинальное, флаговое и так далее) соответствующим образом.
2. Задайте условия для записи значений. Например, можно задать значение 1, если `Drug='drugA'`, 2, если `Drug='drugB'`, и так далее.

Узел заполнения

Узлы заполнения используются для замены значений полей и изменения хранения. Можно выбрать замену значений на основе заданного условия CLEM, например, `@BLANK(FIELD)`. Как вариант, вы можете выбрать замещение всех пустых значений или значений null на конкретное значение. Узлы заполнения часто используются сочетании с узлом Тип для замены пропущенных значений. Например, можно заполнить пробелы средним значением поля, задав такое выражение, как `@GLOBAL_MEAN`. Это выражение заполнит все пробелы средним значением, вычисленным узлом задания глобальных значений.

Заполнить в полях. При помощи средства выбора полей (кнопки справа от текстового поля) выберите в наборе данных поля, значения которых будут проверены и заменены. Поведение по умолчанию предполагает замену значений в зависимости от условия и указанных ниже выражений замены. Можно также выбрать и альтернативный метод замены при помощи опций замены ниже.

Примечание: При выборе нескольких полей для замены их значений на пользовательское значение важно, чтобы типы полей были подобны (все числовые или все символические).

Заменить. Выберите эту опцию для замены значений выбранных полей при помощи одного из следующих методов:

- **На основе условия.** Эта опция активирует поле Условие и Построитель выражений для создания выражения, используемого в качестве условия выполнения замены на заданное значение.
- **Всегда.** Заменяет все значения выбранного поля. Например, эту опцию можно использовать для преобразования типа хранения прибыли (income) на строковое значение (string) при помощи выражения CLEM: `(to_string(income))`.
- **Пробельные значения.** Заменяет все пользовательские пробельные значения в выбранном поле. Для выбора пробельных значений используется стандартное условие `@BLANK(@FIELD)`. *Примечание*: Пробелы можно определить при помощи вкладки Типы узла источника или на узле Тип.
- **Пустые значения.** Заменяет все системные пустые значения (NULL) выбранном поле. Для выбора пустых значений используется стандартное условие `@NULL(@FIELD)`.
- **Пробельные и пустые значения.** Заменяет в выбранном поле и пробельные значения, и системные пустые значения. Эта опция полезна, если неизвестно, были ли пустые значения определены как пропущенные значения.

Условие. Эта опция доступна, когда выбрана опция **На основе условия**. В этом текстовом поле можно задать выражение CLEM для оценки выбранных полей. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений.

Заменить на. Задайте выражение CLEM, чтобы присвоить выбранным полям новое значение. Можно также заменить значение на пустое значение (NULL), введя в текстовом поле `undef`. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений.

Примечание: Если выбранные поля - строковые, их значения следует заменить на строковое значение. Использование в качестве значения замены значения по умолчанию 0 или другого числового значение приведет к ошибке.

Преобразование хранения при помощи узла заполнения

При помощи условия Заменить узла заполнения можно легко преобразовать хранение для одного или нескольких полей. Например, при помощи функции преобразования `to_integer` можно преобразовать прибыль *(income)* из строкового значения в целочисленное, применив следующее выражение CLEM: `to_integer(income)`.

Можно просмотреть доступные функции преобразования и автоматически создать выражение CLEM при помощи Построителя выражения. В выпадающем списке Функции выберите **Преобразование**, чтобы просмотреть список функций преобразования хранения. Доступны следующие функции преобразования:

- `to_integer(ITEM)`
- `to_real(ITEM)`
- `to_number(ITEM)`
- `to_string(ITEM)`
- `to_time(ITEM)`
- `to_timestamp(ITEM)`
- `to_date(ITEM)`
- `to_datetime(ITEM)`

Преобразование значений дат и времени. Имейте в виду, что функции преобразования (и любые другие функции, которым требуется конкретный тип ввода, такой как значение даты и времени) зависят от текущих форматов, заданных в диалоговом окне опций потока. Например, если вы хотите преобразовать строковое поле со значениями *Jan 2003*, *Feb 2003* и так далее в хранение даты, в качестве формата дат по умолчанию для потока выберите **МЕС ГГГГ**.

Функции преобразования доступны также с узла извлечения, для временного преобразования во время вычисления операции извлечения. Узел извлечения можно также использовать для выполнения и других операций с данными, таких как перекодирование строковых полей с категориальными значениями. Дополнительную информацию смотрите в разделе “Перекодирование значений при помощи узла извлечения” на стр. 151.

Узел переклассификации

Узел переклассификации позволяет выполнить преобразование из одного набора категориальных значений в другой. Переклассификация полезна для свертывания категорий или перегруппировки данных для анализа. Например, значения для категории *Продукт* можно переклассифицировать в три группы, такие как *Кухонная посуда*, *Ванна и белье* и *Приборы*. Часто эта операция выполняется непосредственно с узла распределения путем группирования значений и генерирования узла переклассификации. Дополнительную информацию смотрите в разделе “Использование узла распределения” на стр. 231.

Переклассификацию можно выполнить для одного или нескольких символических полей. Можно также выбрать подстановку новых значений для существующего поля или генерирование нового поля.

Когда использовать узел переклассификации

Перед использованием узла переклассификации посмотрите, не лучше ли подходит для поставленной задачи другой узел операций с полями:

- Для преобразования диапазонов числовых значений в наборы при помощи автоматического метода, такого как применение рангов или процентилей, следует использовать узел разделения на интервалы. Дополнительную информацию смотрите в разделе “Узел разделения на интервалы” на стр. 157.
- Для классификации диапазонов числовых значений в наборы вручную следует использовать узел извлечения. Например, если вы хотите свернуть значения заработной платы в конкретные категории диапазона заработной платы, нужно определить каждую категорию при помощи узла извлечения вручную.
- Для создания одного или нескольких полей на основе значений категориального поля, такого как *Тип_закладной*, следует использовать узел Задать как флаг.
- Для преобразования категориального поля в числовой тип хранения можно использовать узел извлечения. Например, значения *Нет* и *Да* можно преобразовать соответственно в 0 и 1. Дополнительную информацию смотрите в разделе “Перекодирование значений при помощи узла извлечения” на стр. 151.

Задание опций для узла переклассификации

Существует три шага использования узла переклассификации:

1. Во-первых решите, хотите ли вы переклассифицировать несколько полей или только одно поле.
2. Далее выберите, перекодировать ли существующее поле или создать новое поле.
3. Затем при помощи динамических опций в диалоговом окне переклассификации отобразите наборы должным.

   Режим. Выберите опцию **Одно**, чтобы переклассифицировать категории для одного поля. Выберите **Несколько**, чтобы активировать опции, включающие поддержку преобразования сразу нескольких полей.

   Переклассифицировать как. Выберите опцию **Новое поле**, чтобы сохранить исходное номинальное поле и получить дополнительное поле, содержащее переклассифицированные значения. Выберите вариант **Существующее поле**, чтобы перезаписать значения в исходном поле на новые классификации. По сути это операция "заполнения".

   После задания режима и опций замены нужно выбрать поле преобразования и задать новые значения классификаций при помощи динамически опций в нижней половине диалогового окна. Состав этих опций зависит от режима, выбранного вами выше.

   Переклассифицировать поля. При помощи кнопки Средство выбора полей, находящейся справа, выберите одно (режим Одно) или несколько (режим Несколько) категориальных полей.

   Имя нового поля. Задайте имя для нового номинального поля, содержащего перекодированные значения. Эта возможность доступна только в режиме Одно, если выше выбрана опция **Новое поле**. Если выбрана опция **Существующее поле**, исходное поле будет сохранено. При работе в режиме Несколько, эта опция заменяется на элементы управления для задания расширения, добавляемое к имени каждого нового поля. Дополнительную информацию смотрите в разделе “Переклассификация нескольких полей” на стр. 154.

   Переклассифицировать значения. Эта таблица позволяет очистить отображение значений старого набора на заданные здесь значения.

   - **Исходное значение.** Этот столбец содержит существующие значения для выдранных полей.
   - **Новое значение.** Введите при помощи этого столбца новые значения категорий или выберите значение из выпадающего списка. При автоматическом генерировании узла переклассификации с помощью значений из диаграммы распределения эти значения включаются в выпадающий список. Это позволяет быстро отобразить существующие значения на известный набор значений. Например, медицинские организации иногда группируют диагнозы по разному на основе сети или локали. После их слияния или приобретения всем сторонам потребуется переклассифицировать новые или даже существующие данные сообразным способом. Вместо ввода вручную каждого значения назначения из очень длинного списка вы можете передать главный список значений в IBM SPSS Modeler, вызвать диаграмму распределения для поля *Диагноз* и сгенерировать для этого поля узел переклассификации (значений) непосредственно из диаграммы. Этот процесс сделает все значения назначения поля Диагноз доступными в выпадающем списке Новые значения.

4. Нажмите кнопку **Получить**, чтобы прочитать значения для одного или нескольких полей, выбранных выше.
5. Нажмите кнопку **Копировать**, чтобы вставить исходные значения в столбец *Новое значение* для полей, которые еще не были отображены. Отображенные исходные значения будут добавлены в выпадающий список.
6. Нажмите кнопку **Очистить новые**, чтобы стереть все спецификации в столбце *Новое значение*. *Примечание*: Эта опция не стирает значения в выпадающем списке.
7. Нажмите кнопку **Авто**, чтобы автоматически сгенерировать последовательные целые числа для каждого из исходных значений. Могут быть сгенерированы только последовательные целые числа (но не действительные значения, такие как 1,5, 2,5, и так далее).

Например, можно автоматически сгенерировать последовательные номера ID продуктов или номера предлагаемых курсов обучения в университете. Эта функциональная возможность соответствует преобразованию Автоматическое перекодирование для наборов в IBM SPSS Statistics.

Для неспецифицированных значений использовать. Эта опция используется для заполнения незаданных значений в новом поле. Можно либо сохранить исходное значение, выбрав опцию **Исходное значение**, либо задать значение по умолчанию.

Переклассификация нескольких полей

Для отображения значений категорий сразу для нескольких полей задайте режим **Несколько**. При этом в диалоговом окне Переклассифицировать включится поддержка новых параметров, которые описаны ниже.

Переклассифицировать поля. При помощи кнопки Средство выбора полей, находящейся справа, выберите поля, которые вы хотите преобразовать. При помощи средства выбора полей можно выбрать все поля сразу или поля конкретного типа, такие как номинальные или флаговые.

Расширение имени поля. При одновременной записи нескольких полей эффективнее задать общее расширение, добавляемое ко всем новым менам полей, а не к отдельным именам полей. Задайте расширение, такое как `_recode` и выберите, добавлять ли его в конец или в начало исходного имени поля.

Хранение и шкала измерений для переклассифицированных полей

Узел переклассификации всегда создает из операции перекодирования номинальное поле. В некоторых случаях это может изменить шкалу измерений поля, если используется режим переклассификации **Существующее поле**.

Хранение нового поля (способ *хранения* данных, а не их *использования*) вычисляется на основе следующих опций вкладки Параметры.

- Если неспецифицированные значения заданы для использования значения по умолчанию, тип хранения определяется путем проверки как новых значений, так и значения по умолчанию, и определения подходящего типа хранения. Например, если значения можно синтаксически проанализировать как целочисленные, у поля будет целочисленный тип хранения.
- Если неспецифицированные значения заданы для использования исходных значений, тип хранения определяется типом хранения исходного поля. Если все значения могут быть синтаксически проанализированы при помощи типа хранения исходного поля, хранение остается прежним; в противном случае хранение определяется путем нахождения типа хранения, наиболее подходящего и для новых, и для старых значений. Например, переклассификация целочисленного набора { 1, 2, 3, 4, 5 } с переклассификацией 4 => 0, 5 => 0 сгенерирует новый набор целых чисел { 1, 2, 3, 0 }, тогда как с переклассификацией 4 => “больше 3”, 5 => “больше 3” будет сгенерирован набор строчных значений { “1”, “2”, “3”, “больше 3” }.

Примечание: Если исходный тип был неинстанциированным, новый тип тоже будет неинстанциированным.

Узел анонимизации

Узел анонимизации позволяет скрыть имена полей и/или значения полей при работе с данными, которые должны быть включены в нисходящий поток моделей этого узла. Таким образом, обобщенную модель можно беспрепятственно рассылать (например, в службу технической поддержки), не опасаясь, что неавторизованные пользователи смогут просмотреть конфиденциальные данные, такие как записи о сотрудниках или медицинские карты пациентов.

В зависимости от того, куда вы помещаете узел анонимизации в потоке, может потребоваться внести изменения в другие узлы. Например, если вы вставляете узел анонимизации вверх по потоку с узла выбора, критерии выбора на узле выбора нужно будет изменить, если они действуют на значения, которые становятся теперь анонимизированными.

Метод, используемый для анонимизации, зависит от различных факторов. Для имен полей и всех значений полей, кроме случаев шкал измерений непрерывного типа, данные заменяются строкой формы:

*префикс*_S*n*

где *prefix_* - либо пользовательская строка, либо строка по умолчанию `anon_`, а *n* - целочисленное значение, начинающееся нулем и увеличивающееся на 1 для каждого уникального значения (например: `anon_S0`, `anon_S1`, и так далее).

Значения полей непрерывного типа должны быть преобразованы, поскольку числовые диапазоны касаются целочисленных значений или действительных чисел, но не строк. Соответственно, их можно анонимизировать только посредством преобразования диапазона в другой диапазон, поэтому исходные данные будут скрыты. Преобразование значения *x* в диапазон выполняется следующим образом:

`A*(`*x*` + B)`

где:

A - коэффициент масштабирования, который должен быть больше 0.

B - добавляемое к значениям смещение пересчета.

Пример

Для случая поля *AGE*, где задан коэффициент масштабирования *A* 7 и смещение пересчета *B* 3, значения для *AGE* будут преобразованы следующим образом:

```
7*(AGE + 3)
```

Задание опций для узла анонимизации

Здесь можно выбрать, значения каких полей должны быть скрыты в дальнейшем нисходящем потоке.

Имейте в виду, что перед тем, как можно будет выполнять операции анонимизации, поля данных должны быть инстанциированы с узла анонимизации вверх по потоку. Данные можно инстанциировать, нажав кнопку **Прочитать значения** на узле Тип или на вкладке Типы узла источника.

Поле. Возвращает список полей в текущем наборе данных. Если какие-либо имена полей уже были анонимизированы, эти имена появятся здесь.

Измерение. Шкала измерений поля.

Анонимизировать значения. Выберите одно или несколько полей, щелкните по этому столбцу и выберите **Да**, чтобы анонимизировать значение поля при помощи префикса по умолчанию **anon_**; выберите **Задать**, чтобы вывести диалоговое окно, в котором можно ввести свой собственный префикс, или (в случае значений полей *непрерывного* типа) укажите, использовать ли для преобразования значений полей случайные или

пользовательские значения. Имейте в виду, что поля *непрерывного* и *дискретного* типов нельзя указывать в одной и той же операции; это нужно делать по отдельности для каждого типа поля.

Просмотр текущих полей. Выберите эту опцию для просмотра полей наборов данных, соединение которых с узлом анонимизации активно. Эта опция выбирается по умолчанию.

Просмотр неиспользуемых параметров полей. Выберите эту опцию для просмотра полей наборов данных, которые были соединены, но более не соединены с узлом анонимизации. Эта опция полезна при копировании узлов из одного потока в другой или при сохранении узлов и их перезагрузке.

Задание способов анонимизации значений полей

В диалоговом окне Заменить значения можно выбрать, использовать ли для анонимизированных значений полей префикс по умолчанию или пользовательский префикс. Нажатие в этом поле кнопки **OK** изменяет значение параметра Анонимизировать значения на вкладке Параметры на значение **Да** для одного или нескольких выбранных полей.

Префикс значений полей. Для анонимизированных значений полей используется префикс по умолчанию **anon_**; выберите **Пользовательский** и введите свой префикс, если вместо префикса по умолчанию вы хотите использовать другой.

Диалоговое окно Преобразовать значения вводится только для полей непрерывного типа; в нем можно указать, использовать ли для преобразования значений полей случайные или пользовательские значения.

Переменный. Выберите эту опцию, чтобы использовать для преобразования случайные значения. Опция **Задать начальное значение рандомизации** выбрана по умолчанию; задайте значение в поле **Начальное значение** или используйте значение по умолчанию.

Фиксированная. Выберите эту опцию, чтобы задать для преобразования ваши собственные значения.

- **Умножить на.** Число, на которое будут умножаться значения полей при преобразовании. Минимальное значение равно 1; максимальное - обычно 10, но оно может быть уменьшено во избежание переполнения.
- **Увеличить на.** Число, которое будет прибавляться к значениям полей при преобразования. Минимальное значение равно 0; максимальное - обычно 1000, но оно может быть уменьшено во избежание переполнения.

Анонимизация значений полей

Значения полей, выбранных для анонимизации на вкладке Параметры, уже анонимизированы:

- При запуске потока, содержащего узел анонимизации
- При просмотре значений

Для предварительного просмотра значений нажмите кнопку **Анонимизировать значения** на вкладке Анонимизированные значения. Далее выберите имя поля в выпадающем списке.

Если используется шкала измерений непрерывного типа, в выводе появятся:

- Минимальное и максимальное значения исходного диапазона.
- Уравнение, используемое для преобразования значений.

Если тип используемой шкалы измерений отличается от непрерывного, на экране появятся исходное и анонимизированное значения для этого поля.

Если вывод появляется на желтом фоне, это означает, что либо параметры настройки выбранного поля были изменены с момента последней анонимизации значений, либо были внесены изменения в данные вверх по потоку узла анонимизации, вследствие чего анонимизированные значения, возможно, более неверны. Появится текущий набор значений; нажмите кнопку **Анонимизировать значения** еще раз, чтобы сгенерировать новый набор значений, соответствующих текущим параметрам настройки.

Анонимизировать значения. Создает анонимизированные значения для выбранного поля и выводит их в таблице. Если используются случайные начальные значения для поля непрерывного типа, повторное нажатие этой кнопки будет каждый раз создавать другой набор значений.

Очистить значения. Очищает исходные и анонимизированные значения в таблице.

Узел разделения на интервалы

Узел разделения на интервалы автоматически создает новые номинальные поля на основе значений одного или нескольких непрерывных полей (числового диапазона). Например, вы можете преобразовать непрерывное поле доходов в новое категориальное поле, содержащее категории дохода равной ширины или как отклонения от среднего. Альтернативно, можно выбрать категориальное поле "инспектора", чтобы сохранить прочность исходной связи между двумя полями.

Разделение на интервалы может оказаться полезным по ряду причин, включая:

- **Требования к алгоритмам.** Определенным алгоритмам, таким как наивный критерий Байеса или логистическая регрессия, требуются категориальные входные поля.
- **Производительность.** Такие алгоритмы, как полиномиальная логистическая регрессия, могут выполняться лучше, если число отличительных значений входных полей будет сокращено. Например, используйте для каждого интервала значение медианы или среднего вместо исходных значений.
- **Конфиденциальность данных.** Конфиденциальную информацию о персонале, такую как, величина заработной платы, можно сообщать диапазонами значений вместо фактических величин для защиты превратности.

Доступен ряд методов разделения на интервалы. Создав интервалы для нового поля, вы можете сгенерировать узел извлечения на основе точек отсечения.

Когда использовать узел разделения на интервалы

Перед использованием узла разделения на интервалы посмотрите, не лучше ли подходит для поставленной задачи другая методика:

- Чтобы вручную задать точки отсечения для категорий, таких как конкретные заранее заданные диапазоны заработных плат, используйте узел извлечения. Дополнительную информацию смотрите в разделе “Узел извлечения” на стр. 144.
- Чтобы создать новые категории для существующих наборов, используйте узел переклассификации. Дополнительную информацию смотрите в разделе “Узел переклассификации” на стр. 152.

Обработка пропущенных значений

Узел разделения на интервалы обрабатывает пропущенные значения следующими способами:

- **Пользовательские пробелы.** Пропущенные значения, заданные в виде пробелов, исключаются во время преобразования. Например, если вы назначили –99 для указания пробельного значения при помощи узла Тип, это значение будет включено в процесс разделения на интервалы. Чтобы игнорировать пробелы во время разделения на интервалы, вы должны заменить пробельные значения на системное пустое значение при помощи узла заполнения.
- **Системные пропущенные значения ($null$).** Пустые значения игнорируются во время преобразования с разделением на интервалы и остаются пустыми после преобразования.

На вкладке Параметры представлены опции для доступных методик. На вкладке Представление выводятся точки отсечения, установленные для данных, ранее обработанных на узле.

Задание опций для узла разделения на интервалы

С помощью узла разделения на интервалы можно автоматически сгенерировать интервалы (категории), применив следующие методы:

- Разделение на интервалы фиксированной ширины
- Плитка (равное число или сумма)
- Среднее линейное и среднеквадратичное отклонение:
- Ранги
- Оптимизированное относительно категориального поля "инспектор"

Нижняя честь диалогового окна изменяется динамически в зависимости от выбираемого метода разделения на интервалы.

Поля интервалов разделения. Здесь выводятся непрерывные поля (числовых диапазонов), ожидающие преобразования. Узел разделения на интервалы позволяет разделять на интервалы несколько полей одновременно. Поля можно добавить или удалить при помощи кнопок справа.

Метод разбиения на группы. Выберите метод, используемый для определения точек отсечения для новых интервалов (категорий) полей. В последующих темах описаны опции, доступные в каждом случае.

Пороги интервалов разделения. Укажите, как вычислять пороги.

- **Всегда пересчитывать.** Точки отсечения и выделение интервалов всегда будут вычисляться заново при запуске узла.
- **Прочитать со вкладки Значения интервалов разделения, если она доступна.** Точки отсечения и выделение интервалов будут вычисляться только при необходимости (например, если были добавлены новые данные).

В следующих темах обсуждаются опции для доступных методов разделения на интервалы.

Интервалы фиксированной ширины

При выборе в качестве метода разделения на интервалы опции **Фиксированная ширина** в диалоговом окне появляется новый набор опций.

Расширение имен. Задайте расширение, которое следует использовать для сгенерированных полей. *_BIN* - это расширение по умолчанию. Можно также указать, добавлять ли расширение в начало имени поля (**Префикс**) или в его конец (**Суффикс**). Например, вы можете сгенерировать новое поле с именем *income_BIN*.

Ширина хранения. Задайте значение (целое или действительное число), используемое для вычисления “ширины” интервала. Например, для разделения на интервалы поля *Возраст* можно использовать значение по умолчанию 10. При диапазоне *возраста* от 18 до 65 сгенерированные интервалы будут подобны показанным в следующей таблице.

Таблица 23. Интервалы для поля Возраст с диапазоном от 18 до 65

Интервал 1	Интервал 2	Интервал 3	Интервал 4	Интервал 5	Интервал 6
>=13 но <23	>=23 но <33	>=33 но <43	>=43 но <53	>=53 но <63	>=63 но <73

Начало интервалов разделения вычисляется по наименьшему отсканированному значению минус половина ширины интервала (заданной). Например, в интервалах разделения, показанных выше, в качестве начала интервалов используется 13 в соответствии со следующим вычислением: 18 *[наименьшее значение данных]* – 5 *[0,5 × (ширина интервала 10)]* = 13.

Число групп. Эта опция позволяет задать целочисленное значение, определяющее число интервалов (категорий) фиксированной ширины для новых полей.

После выполнения узла разделения на интервалы в потоке можно будет просмотреть сгенерированные пороги интервалов, щелкнув по вкладке **Предварительный просмотр** в диалоговом окне узла разделения на интервалы. Дополнительную информацию смотрите в разделе “Предварительный просмотр обобщенных интервалов” на стр. 162.

Плитка (равное число или сумма)

Метод плиточного разделения на интервалы создает номинальные поля, которые можно использовать для разбиения отсканированных записей на группы процентилей (или квартилей, децилей и так далее) так, чтобы каждая группа содержала одно и то же число записей или чтобы сумма значений в каждой группе была одинаковой. Записи ранжируются по возрастанию на основе значения заданного поля интервала, поэтому записям с самыми низкими значениями для выбранной переменной интервала назначается ранг 1, следующему набору записей назначается ранг 2 и так далее. Значения порогов для каждого интервала генерируются автоматически на основе используемых данных и метода разделения на плитки.

Расширение имен плиток. Задайте расширение, используемое для полей, сгенерированных при помощи стандартного метода p-плитки. Расширение по умолчанию - *_TILE* плюс *N*, где *N* - номер плитки. Можно также указать, добавлять ли расширение в начало имени поля (**Префикс**) или в его конец (**Суффикс**). Например, вы можете сгенерировать новое поле с именем *income_BIN4*.

Пользовательское расширение плиток. Задайте расширение для пользовательского диапазона плиток. Значение по умолчанию - *_TILEN*. Имейте в виду, что *N* в этом случае *не* будет заменено пользовательским номером.

Доступные методы p-плитки:

- **Квартиль.** Генерируются 4 интервала, каждый из которых содержит 25% наблюдений.
- **Квинтиль.** Генерируются 5 интервалов, каждый из которых содержит 20% наблюдений.
- **Дециль.** Генерируются 10 интервалов, каждый из которых содержит 10% наблюдений.
- **Вингтиль.** Генерируются 20 интервалов, каждый из которых содержит 5% наблюдений.
- **Процентиль.** Генерируются 100 интервалов, каждый из которых содержит 1% наблюдений.
- **Пользовательское N.** Выберите эту опцию, чтобы задать число интервалов. Например, значение 3 сгенерирует 3 полосовые категории (и 2 точки отсечения), каждая из которых будет содержать 33,3% наблюдений.

Имейте в виду, что если число дискретных значений в данных будет меньше указанного числа плиток, плитки использоваться не будут. В таких случаях новое распределение, скорее всего, отразит исходное распределение ваших данных.

Метод разделения на плитки. Задает метод, используемый для назначения записей в интервалы.

- **Число записей.** Пытается назначить в каждый интервал равное число записей.
- **Сумма значений.** Пытается назначить в интервалы записи так, чтобы сумма значений в каждом интервале была одинаковой. При планировании направлений работ по продажам с помощью этого метода можно, например, назначить потенциальных покупателей в группы децилей на основе значения для каждой записи, с самыми ценными покупателями в верхнем интервале. Например, фармацевтическая фирма может ранжировать врачей, назначив их в группы децилей на основе числа написанных ими предписаний. Тогда как каждый дециль будет содержать приблизительно равные количества письменных текстов, число лиц, распространяющих эти тексты, будет неодинаковым, причем авторы, пишущие больше всего рукописей, будут сосредоточены в децили 10. Имейте в виду, что этот подход предполагает, что все значения - больше нуля;, в противном случае могут быть возвращены неожиданные результаты.

Совпадающие наблюдения. Результаты условия совпадающих наблюдений, когда значения с любой из стороны от точки отсечения идентичны. Например, если вы назначаете децили и более 10% записей представляют для поля интервала разделения одно и то же значение, то все они не смогут попасть в один и тот же интервал без форсирования порога так или иначе. Совпадающие наблюдения могут быть перенесены

в следующий интервал или сохранены в текущем интервале, но должны быть разрешены, чтобы все записи с идентичными значениями попали в один и тот же интервал, даже если при этом в некоторых интервалах окажется больше совпадающих наблюдений, чем ожидалось. Вследствие этого могут также быть скорректированы пороги последующих интервалов, что приведет к неодинаковому назначению значений для одного и того же набора чисел, на основе метода, используемого для разрешения совпадающих наблюдений.

- **Добавлять в следующий.** Выберите эту опцию для переноса значений совпадающих наблюдений в следующий интервал разделения.
- **Сохранять в текущем.** Сохраняет значения совпадающих наблюдений в текущем (более низком) интервале разделения. Этот метод может привести к тому, что всего будет создано меньше интервалов.
- **Назначать произвольно.** Выберите эту опцию для выделения значений совпадающих наблюдений для интервала случайным образом. Эта опция пытается сохранять записи в каждом интервале в равных количествах.

Пример: Разделение на плитки по числу записей

В следующей таблице показано, как упрощенные значения полей ранжируются в виде квартилей при разделении на плитки по числу записей. Обратите внимание, что результаты отличаются в зависимости от выбранной опции совпадающих наблюдений.

Таблица 24. Пример разделения на плитки по числу записей.

Значения	Добавить в следующий	Сохранять в текущем
10	1	1
13	2	1
15	3	2
15	3	2
20	4	3

Число позиций на один интервал вычисляется следующим образом:

```
Общее число значений / число плиток
```

В упрощенном примере выше нужное число позиций для одного интервала равно 1,25 (5 значений / 4 квартили). Значение 13 (номер значения 2) захватывает в вилку желаемый числовой порог 1,25 и поэтому обрабатывается неодинаково в зависимости от выбранной опции совпадающих наблюдений. В режиме **Добавить в следующий** оно добавляется в интервал 2. В режиме **Сохранить в текущем** оно остается в интервале, с вытеснением диапазона значений для интервала 4 за диапазон существующих значений данных. В результате создаются только три интервала, и пороги для каждого интервала корректируются соответствующим образом, как показано в следующей таблице.

Таблица 25. Результат примера разделения на интервалы.

Группировка	Нижняя	Верхняя
1	>=10	<15
2	>=15	<20
3	>=20	<=20

Примечание: Скорость разделения на интервалы по плиткам может выиграть от включения поддержки параллельной обработки.

Ранжировать наблюдения

При выборе в качестве метода разделения на интервалы опции **Ранги** в диалоговом окне появляется новый набор опций.

При ранжировании создаются новые поля, содержащие ранги, дробные ранги и значения процентилей для числовых полей в зависимости от заданных ниже опций.

Порядок ранга. Выберите **По возрастанию** (наименьшее значение будет помечено 1) или **По убыванию** (наибольшее значение будет помечено 1).

Ранг. Выберите эту опцию для ранжирования наблюдений по возрастанию или по убыванию, как указано выше. Диапазоном значений в новом поле будет 1 - *N*, где *N* - число дискретных значений в исходном поле. Совпадающим значениям присваивается среднее для их ранга.

Дробный ранг. Выберите эту опцию для ранжирования наблюдений, где значение нового поля равно рангу, деленному на сумму весов непропущенных наблюдений. Дробные ранги попадают в диапазон 0 – 1.

Процент дробного ранга. Каждый ранг делится на число записей с непропущенными значениями и умножается на 100. Процент дробных рангов попадают в диапазон 0 – 100.

Добавочный. Для всех опций ранжирования можно также создать пользовательские расширения и указать, добавлять ли это расширение в начало имени поля (**Префикс**) или в его конец (**Суффикс**). Например, вы можете сгенерировать новое поле с именем *income_P_RANK*.

Среднее линейное/среднеквадратичное отклонение

При выборе в качестве метода разделения на интервалы опции **Среднее линейное/среднеквадратичное отклонение** в диалоговом окне появляется новый набор опций.

Этот метод генерирует одно или несколько новых полей с полосовыми категориями на основе значений среднего линейного и среднеквадратичного отклонения распределения заданных полей. Выберите ниже число отклонений для использования.

Расширение имен. Задайте расширение, которое следует использовать для сгенерированных полей. *_SDBIN* - это расширение по умолчанию. Можно также указать, добавлять ли расширение в начало имени поля (**Префикс**) или в его конец (**Суффикс**). Например, вы можете сгенерировать новое поле с именем *income_SDBIN*.

- **+/– 1 среднеквадратичное отклонение.** Выберите эту опцию, чтобы сгенерировать три интервала.
- **+/– 2 среднеквадратичных отклонения.** Выберите эту опцию, чтобы сгенерировать пять интервалов.
- **+/– 3 среднеквадратичных отклонения.** Выберите эту опцию, чтобы сгенерировать семь интервалов.

Например, выбор опции +/–1 среднеквадратичное отклонение приведет к трем интервалам, вычисленным и показанным в следующей таблице.

Таблица 26. Пример интервалов среднеквадратичного отклонения.

Интервал 1	Интервал 2	Интервал 3
x < (среднее линейное – среднеквадратичное отклонение)	(среднее линейное – среднеквадратичное отклонение) <= x <= (среднее линейное + среднеквадратичное отклонение)	x > (среднее линейное + среднеквадратичное отклонение)

В случае нормального распределения 68% всех наблюдений попадают в интервал плюс/минус одно среднеквадратичное отклонение от среднего линейного, 95% - в интервал плюс/минус два среднеквадратичных отклонения и 99% - в интервал плюс/минус три среднеквадратичных отклонения. Однако имейте в виду, что создание полосовых категорий на основе среднеквадратичных отклонений может привести к тому, что некоторые интервалы будут находиться вне диапазона фактических данных и даже вне диапазона возможных значений данных (например, возможны отрицательные значения заработной платы).

Оптимальное разделение на интервалы

Если поле, которое вы хотите разделить на интервалы, прочно связано с другим категориальным полем, можно выбрать категориальное поле в качестве поля "инспектора", чтобы создать интервалы способом, при котором прочность исходной связи между двумя полями сохранится.

Например, при помощи кластерного анализа вы сгруппировали состояния на основе штрафных ставок за просроченные платежи, с наивысшей ставкой в первом кластере. В этом случаев качестве полей интервалов можно выбрать *Процент просроченности* и *Процент лишения права выкупа*, а в качестве поля инспектора - поле принадлежности к кластерам, сгенерированное моделью.

Расширение имени Задайте расширение, которое следует использовать для сгенерированных полей, и укажите, добавлять ли это расширение в начало имени поля (**Префикс**) или в его конец (**Суффикс**). Например, вы можете сгенерировать новое поле с именем `pastdue_OPTIMAL` и еще одно, с именем `inforeclosure_OPTIMAL`.

Поле супервизора Категориальное поле, используемое для построения интервалов.

Предварительно разбитые на категории поля для повышения производительности при работе с большими наборами данных Указывает, следует ли использовать обработку для ускорения оптимального разделения на интервалы. Эта опция распределяет значения шкалы в большое число интервалов при помощи одного метода неконтролируемого разделения на интервалы, представляет значения в каждом интервале средним и корректирует соответствующим образом вес наблюдений перед операцией неконтролируемого разделения на интервалы. В реальных условиях этот метод повышает за счет степени точности скорость и рекомендуется для больших наборов данных. При использовании этой опции можно также задать максимальное количество интервалов для каждой переменной после предварительной обработки.

Объединить категории, содержащие относительно мало наблюдений по сравнению с соседней категорией. Эта опция, если она включена, указывает, что слияние интервала выполняется, если отношение его размера (количества наблюдений) к размеру соседнего интервала меньше указанного порогового значения; большие пороговые значения приводят к увеличению числа слияний.

Параметры точек отсечения

В диалоговом окне Параметры точек отсечения можно задать дополнительные опции для алгоритма оптимальной категоризации. Эти опции указывают алгоритму, как вычислять интервалы разделения при помощи поля назначения.

Конечные точки интервалов разделения. Можно указать, должна ли нижняя или верхняя конечная точка быть включительной (значение <= *x*) или исключительной (значение < *x*).

Первый и последний интервалы. И для первого, и для последнего интервала можно указать, должны ли они быть неограниченными (в направлении положительной или отрицательной бесконечности) или их следует ограничить самой низкой или самой высокой точками данных.

Предварительный просмотр обобщенных интервалов

На вкладке Значения интервалов в диалоговом окне узла разделения на интервалы можно просмотреть пороги для сгенерированных интервалов. При помощи меню Создать можно также сгенерировать узел извлечения, позволяющий применять эти пороги к различным наборам данных.

Категориальное поле. В выпадающем списке выберите поле для просмотра. В показанных именах полей для ясности используется исходное имя поля.

Плитка. В выпадающем меню выберите для просмотра плитку (например, 10 или 100). Эта опция доступна, только если интервалы были сгенерированы при помощи плиточного метода (по равному числу или сумме).

Пороги интервалов разделения. Здесь выводятся значения порогов для каждого сгенерированного интервала наряду с числом записей, попадающих в каждый интервал. Только для метода оптимального разделения на интервалы: выводится число записей в каждом интервале в виде процента от всего количества. Имейте в виду, что при использовании метода разделения на интервалы с ранжированием пороги неприменимы.

Прочитать значения. Читает значения из набора данных. Учтите, что пороги будут также перезаписываться при обработке в потоке новых данных.

Генерирование узла извлечения

С помощью меню Создать можно создать узел извлечения на основе текущих порогов. Это полезно для применения установленных порогов интервалов к различным наборам данных. Более того, знание этих точек разбиения повышает эффективность (скорость выполнения) операции извлечения относительно операции разделения на интервалы при работе с большими наборами данных.

Узел анализа RFM

Узел анализа RFM (Recency, Frequency, Monetary - недавность, частота, деньги) позволяет определить количественно, какие покупатели вероятнее всего будут лучшими, изучив, как давно они последний раз делали у вас покупки (недавность), как часто они их совершали (частота) и сколько они потратили в итоге всех сделок (деньги).

Внутренняя логика анализа RFM предполагает, что клиенты, покупающие продукт или услугу один раз, скорее всего купят их снова. Категоризованные данные о покупателях разделяются на ряд интервалов при помощи критериев разделения на интервалы, корректируемых нужным вам образом. В каждом из этих интервалов покупателям назначается оценка; затем эти оценки объединяются для получения общей оценки RFM. Эта оценка является представлением принадлежности покупателя к интервалам, создаваемым для каждого из параметров RFM. Этих разделенных на интервалы данных может оказаться вполне достаточно для ваших потребностей (например, для идентификации наиболее частых и высокоценных покупателей); иначе эти данные можно передать в поток для дальнейшего моделирования и анализа.

Однако заметим, что хотя возможность анализа и ранжирования оценок RFM и является полезным инструментом, следует знать определенные факторы при его использовании. Можно поддаться искушению нацелить бизнес на покупателей с самыми высоким ранжированием, однако чрезмерное навязывание услуг может привести к раздражению и фактическому обвалу повторных сделок. Следует также помнить, что не стоит пренебрегать покупателями с низкими оценками; вместо этого, потратив внимание, можно повысить их оценки. И наоборот, отдельно взятые высокие оценки не обязательно отражают хорошую перспективу продаж, в зависимости от рынка. Например, покупатель в интервале 5 для недавности, означающем, что покупки совершались совсем недавно, фактически может не быть лучшим целевым покупателем для продающих дорогие продукты с продолжительным сроком службы, такие как автомобили или телевизионная техника.

Примечание: В зависимости от способа хранения данных может потребоваться предварить узел анализа RFM узлом агрегации RFM для преобразования данных в удобный формат для использования. Например, входные данные должны быть в формате покупателей (с одной строкой для каждого покупателя); если данные покупателей представлены в транзакционной форме, при помощи восходящего потока узла агрегации RFM извлеките поля для недавности, частоты и денег. Дополнительную информацию смотрите в разделе “Узел агрегации RFM” на стр. 78.

Узлы агрегации RFM и анализа RFM в IBM SPSS Modeler конфигурируются под использование независимого разделения на интервалы; то есть, они ранжируют данные и разделяют на интервалы для каждого измерения значения недавности, частоты или денег безотносительно к остальным двум их значениям.

Параметры узла анализа RFM

Недавность. При помощи средства выбора полей (кнопки справа от текстового поля) выберите поле недавности. Это может быть дата, отметка времени или одноразрядное число. Имейте в виду, что когда дата самой последней транзакции представлена датой или отметкой времени, самое высокое значение считается самым последним; где задано число, оно представляет время, истекшее с момент а самой последней транзакции, и самым последним считается самое низкое значение.

Примечание: Если перед узлом анализа RFM в потоке используется узел агрегации RFM, поля Недавность, Частота и Деньги, генерируемые узлом агрегации RFM, должны быть выбраны в качестве входных на узле анализ RFM.

Частота. При помощи средства выбора полей выберите поле частоты для использования.

Деньги. При помощи средства выбора полей выберите поле денег для использования,

Число интервалов. Для каждого из трех типов вывода число должно быть создано число интервалов. Значение по умолчанию - 5.

Примечание: Минимальное число интервалов - 2, а максимальное - 9.

Вес. По умолчанию, самая высокая важность при вычислении оценок присваивается данным недавности, после чего следуют частота, а затем деньги. При необходимости веса можно скорректировать, повлияв на одно или несколько из них, чтобы изменить поле присвоения самой высокой важности.

Оценка RFM вычисляется следующим образом: (оценка недавности x вес недавности) + (оценка частоты x вес частоты) + (оценка денег x вес денег).

Совпадающие наблюдения. Задайте способ разделения на интервалы одинаковых оценок. Возможны опции:

- **Добавлять в следующий.** Выберите эту опцию для переноса значений совпадающих наблюдений в следующий интервал разделения.
- **Сохранять в текущем.** Сохраняет значения совпадающих наблюдений в текущем (более низком) интервале разделения. Этот метод может привести к тому, что всего будет создано меньше интервалов. (Это - значение по умолчанию).

Пороги интервалов разделения. Укажите, вычислять ли оценки RFM и выделение интервалов повторно всегда при вызове узла или вычислять их только при необходимости (например, если были добавлены новые данные). Если выбрать опцию **Прочитать со вкладки Значения интервалов разделения, если она доступна**, можно будет отредактировать верхнюю и нижнюю точки отсечения для различных интервалов на вкладке Значения интервалов.

При вызове узла анализа RFM он разделяет на интервалы поля Недавность, Частота и Деньги и добавляет в набор данных следующие новые поля:

- Оценка недавности. Ранг (значение интервала) для поля Недавность.
- Оценка частоты. Ранг (значение интервала) для поля Частота.
- Денежная оценка. Ранг (значение интервала) для поля Деньги.
- Оценка RFM. Взвешенная сумма оценок недавности, частоты и денег.

Добавлять выбросы в последние интервалы разделения. Если включить этот переключатель,записи, находящиеся под нижним интервалом, будут добавляться в нижний интервал, а записи над верхним интервалом будут добавляться в верхний интервал; в противном случае этим записям будет присваиваться пустое значение. Этот переключатель доступен, только если выбрана опция **Прочитать со вкладки Значения интервалов разделения, если она доступна**.

Разделение на интервалы узла анализа RFM

На вкладке Значения интервалов можно просмотреть и в определенных случаях скорректировать пороги для сгенерированных интервалов.

Примечание: Скорректировать значения на этой вкладке можно, только если на вкладке Параметры выбрана опция **Прочитать со вкладки Значения интервалов разделения, если она доступна**.

Категориальное поле. В выпадающем списке выберите поле для разделения на интервалы. Доступны значения, выбранные на вкладке Параметры.

Таблица значений интервалов. Здесь выводятся значения порогов для каждого сгенерированного интервала. Если на вкладке Параметры выбрать опцию **Прочитать со вкладки Значения интервалов разделения, если она доступна**, можно будет скорректировать верхнюю и нижнюю точки отсечения для каждого интервала, щелкнув дважды кнопкой мыши по нужной ячейке.

Прочитать значения. Читает значения из набора данных и заполняет таблицу значений интервалов. Имейте в виду, что если на вкладке Параметры выбрать опцию **Всегда пересчитывать**, пороги интервалов будут перезаписываться при обработке в потоке новых данных.

Узел ансамбля

Узел ансамбля объединяет в сочетание нескольких слепков моделей, получающее более точные предсказания, чем любая их этих моделей по отдельности. Объединяя в сочетание предсказания из нескольких моделей, можно избежать ограничений в отдельных моделях, что приведет к более высокой общей точности. Модели, объединенные этим способом, обычно выполняются по меньшей мере не хуже, чем лучшие модели по отдельности, а зачастую и лучше.

Это объединение узлов происходит автоматически на узлах автоматизированного моделирования автоклассификации, автонумерации и автокластеризации.

После применения узла ансамбля можно, применив узел анализа или узел оценки, сравнить точность объединенных результатов с каждой из входных моделей. Для этого убедитесь, что опция **Отфильтровывать поля, генерируемые моделями ансамблей** не включена на вкладке Параметры узла ансамбля.

Выходные поля

Каждый узел ансамбля генерирует поле, содержащее объединенные оценки. Имя основано на заданном поле назначения и снабжено префиксом *$XF_*, *$XS_* или *$XR_*, в зависимости от его шкалы измерений: флаговом поле, номинальном поле (набора) или непрерывном поле (диапазона) соответственно. Например, если назначение является полем флага с именем *отклик*, выходное поле будет *$XF_отклик*.

Поля доверительной вероятности или склонности. Для флаговых и номинальных полей создаются дополнительные поля доверительной вероятности или склонности на основе метода ансамбля, что подробно описано в следующей таблице.

Таблица 27. Создание поля методом ансамбля.

Метод ансамбля	**Имя поля**
Голосование Голосование со взвешенными доверительными вероятностями Простое голосование со взвешенными склонностями Скорректированное голосование со взвешенными склонностями Интервалы с наибольшей доверительной вероятностью	*$XFC_<поле>*
Средняя простая склонность	*$XFRP_<поле>*

Таблица 27. Создание поля методом ансамбля (продолжение).

Метод ансамбля	Имя поля
Средняя скорректированная простая склонность	*$XFAP_<поле>*

Параметры узла ансамбля

Поле назначения для ансамбля. Выберите одно поле которое будет использоваться как назначение несколькими моделями восходящего потока. Модели восходящего потока могут использовать флаговые, номинальные или непрерывные назначения, но для объединения оценок хотя бы две модели должны совместно использовать одно и то же назначение.

Отфильтровывать поля, генерируемые моделями ансамблей. Удаляет из вывода все дополнительные поля, сгенерированные отдельными моделями, снабжающими данными поле ансамбля. Включите этот переключатель, если вас интересует только объединенная оценка всех входных моделей. Убедитесь, что этот переключатель выключен, если вы хотите, например, применить узел анализа или узел оценки для сравнения точности объединенной оценки с точностью каждой из входных моделей.

Доступные параметры зависят от шкалы измерения поля, выбранного в к качестве назначения.

Непрерывные назначения

Для непрерывных назначений оценки будут осредняться. Это - единственный доступный метод для объединения оценок.

При осреднении оценок или отметок узел ансамбля методом вычисления среднеквадратичной ошибки находит разницу между измеренными или оцененными значениями и значениями true и показывает степень согласованности этих оценок. Для новых моделей вычисляемые среднеквадратичные ошибки генерируются по умолчанию; однако для существующих моделей, если они будут генерироваться, этот переключатель можно выключить.

Категориальные цели

Для категориальных назначений поддерживается ряд методов, включая **метод голосования**, который подсчитывает, сколько раз выбиралось каждое возможное предсказанное значение, и выбирает значение с наивысшим итоговым количеством. Например, если три из пяти моделей предсказывают *да*, а остальные две - *нет*, предсказание *да* выигрывает по голосам со счетом 3:2. В ином варианте голоса могут быть **взвешены** на основе значения доверительной вероятности или склонности для каждого предсказания. Затем веса суммируются, и снова выбирается значение с наивысшим итоговым количеством. Доверительная вероятность для окончательного предсказания будет суммой весов для одержавшего победу значения, деленной на число моделей, включенных в ансамбль.

Все категориальные поля. Для флаговых и номинальных полей поддерживаются следующие методы:

- Голосование
- Голосование со взвешенными доверительными вероятностями
- Интервалы с наибольшей доверительной вероятностью

Только флаговые поля. Только для флаговых полей; доступен также ряд методов на основе склонностей.

- Простое голосование со взвешенными склонностями
- Скорректированное голосование со взвешенными склонностями
- Средняя простая склонность
- Средняя скорректированная склонность

Совпадающие наблюдения голосования. Для методов голосования можно указать, как разрешать совпадающие наблюдения.

- **Произвольный выбор.** Одно из совпадающих значений выбирается случайным образом.
- **Наибольшая доверительная вероятность.** Выигрывает совпадающее значение, предсказанное с наивысшей доверительной вероятностью. Имейте в виду, что она не обязательно должна совпадать наивысшей доверительной вероятностью всех предсказанных значений.
- **Простая или скорректированная склонность (только для флаговых полей).** Выигрывает совпадающее значение, предсказанное с наибольшей абсолютной склонностью, которая здесь вычисляется следующим образом:

```
abs(0,5 - склонность) *
2
```

Или (в случае скорректированной склонности) так:

```
abs(0,5 - скорректированная склонность) * 2
```

Узел раздела

Узлы раздела используются для генерирования поля раздела, разбивающего данные на отдельные поднаборы или выборки для этапов обучения, испытания и проверки при построении модели. Сгенерировав модель при помощи одной выборки и испытав ее при помощи другой, отдельной выборки, можно получить надежный показатель качества обобщения модели на более крупные наборы данных, подобные текущим данным данным.

Узел раздела генерирует номинальное поле с ролью, заданной как **Раздел**. Другой вариант - если в данных уже существует соответствующее поле, его можно назначить как поле раздела при помощи узла типа. В этом случае никакого отдельного узла раздела не требуется. В качестве поля раздела можно использовать любое инстанциированное поле с двумя или тремя значениями, но флаговые поля использовать нельзя. Дополнительную информацию смотрите в разделе “Задание роли поля” на стр. 139.

В потоке может быть определено несколько полей раздела, но тогда на вкладке Поля на каждом узле моделирования, где используется разделение, должно быть выбрано одно поле раздела. (Если представлен только один раздел, он будет использоваться автоматически при всяком включении разделения.)

Включение разделения. Для возможности использования раздела при анализе разделение должно быть включено на вкладке Опции модели соответствующего узла построения модели или анализа. Выключение этой опции делает возможным отключить разделение без удаления этого поля.

для создания поля раздела на основе какого-либо другого критерия, например, диапазона дат или положения, можно также использовать узел извлечения. Дополнительную информацию смотрите в разделе “Узел извлечения” на стр. 144.

Пример. При построении потока RFM для выявления недавних заказчиков, положительно ответивших на предыдущие маркетинговые кампании, маркетинговый отдел компании продаж использует узел раздела для разбиения данных на обучающий и контрольный разделы.

Опции узла раздела

Поле раздела. Задает имя поля, созданного узлом.

Разделение. Вы можете разделить данные на две выборки (обучающую и контрольную) или на три выборки (обучающую, контрольную и проверочную).

- **Обучение и контроль.** Разделяет данные на две выборки, позволяя обучать модель на одной выборке и тестировать на другой.
- **Обучение, контроль и проверка.** Разделяет данные на три выборки, позволяя обучать модель на одной выборке, тестировать и уточнять модель при помощи второй выборки и проверять полученные

результаты по третьей. Однако, эта опция уменьшает соответственно размер каждого раздела и может оказаться наиболее удобной при работе с очень большими наборами данных.

Размер раздела. Задает относительный размер каждого раздела. Сумма размеров раздела меньше 100% означает, что записи, не включенные в раздел, будут отброшены. Например, если у пользователя 10 миллионов записей и он задал размеры раздела для обучения 5%, а для тестирования 10%, после вызова узла будет примерно 500000 обучающих записей и один миллион контрольных записей с отброшенным остатком.

Значения. Задает значение, используемое для представления каждой выборки раздела в данных.

- **Использовать значения, определенные системой ("1", "2" и "3").** Использует для представления каждого раздела целое число; например, у всех записей, попадающих в обучающую выборку, для поля раздела будет значение 1. Это обеспечивает переносимость данных между локалями и гарантирует, что в случае повторной инстанциации поля раздела в другом месте (например, при считывании данных обратно в базу данных) порядок сортировки будет сохранен (так что 1 все равно будет представлять обучающий раздел). Однако значения требуют некоторой интерпретации.
- **Добавить метки к значениям, определенным системой.** Объединяет с меткой целое число; например, у записей обучающего раздела будет значение `1_Обучение`. Это делает возможным при просмотре данных отличить друг от друга отдельные значения и сохранить порядок сортировки. Однако значения будут специфичны конкретной локали.
- **Использовать метки как значения.** Использует метку без целочисленного значения, например: **Обучение**. Позволяет задавать значения посредством редактирования меток. Однако это делает данные зависящими от локали, и при повторной инстанциации столбца раздела значениям будет придан их естественный порядок сортировки, который может не соответствовать их "семантическому" порядку.

Значение. Доступно только при выборе опции **Повторное назначение разделов**. При выборке или разделении записей на основе случайного процента эта опция позволяет продублировать одни и те же результаты в другом сеансе. Задав начальное значение, используемое генератором случайных чисел, можно гарантировать назначение одних и тех же записей при каждом вызове узла. Введите нужное начальное значение рандомизации или нажмите кнопку **Сгенерировать**, чтобы автоматически сгенерировать случайное значение. Если эта опция не выбрана, при каждом вызове узла будет генерироваться отличающаяся выборка.

Примечание: При использовании опции **Задать начальное значение генератора псевдослучайных чисел** для записей, читаемых из базы данных, предварительно может потребоваться узел Сортировка для подготовки выборки, чтобы обеспечить одинаковый результат при каждом выполнении узла. Это связано с зависимостью начального значения рандомизации от порядка записей, который не гарантированно будет оставаться одним и тем же в реляционной базе данных. Дополнительную информацию смотрите в разделе “Узел сортировки” на стр. 79.

Использовать уникальное поле для назначения разделов. Доступно только при выборе опции **Повторное назначение разделов**. (Только для баз данных яруса 1) Включите этот переключатель, чтобы использовать обратный перенос SQL для назначения записей в разделы. В выпадающем списке выберите поле с уникальными значениями (например, поле ID), чтобы гарантировать назначение записей случайным, но повторяемым образом.

Ярусы базы данных объясняются в описании узла источника базы данных. Дополнительную информацию смотрите в разделе “Узел источника базы данных” на стр. 17.

Генерирование узлов выбора

При помощи меню Создать на узле раздела можно автоматически сгенерировать узел выбора для каждого раздела. Например, можно выбрать все записи в обучающем разделе, чтобы получить дополнительную оценку или анализ при помощи только этого раздела.

Узел Задать как флаг

Узел Задать как флаг используется для получения полей на основе категориальных значений для одного или нескольких полей. Например, набор данных может содержать номинальное поле *АД* (артериальное давление) со значениями *Высокое*, *Нормальное* и *Низкое*. Для упрощения работы с данными можно создать флаговое поле для высокого артериального давления, указывающее наличие или отсутствие высокого артериального давления у пациента.

Задание опций для узла Задать как флаг

Задать флаги. Возвращает список всех полей со шкалой измерений *Номинальная* (набора). Выберите нужное в списке для вывода значений в наборе. Из них можно выбрать значения для создания флагового поля. Имейте в виду, чтобы можно было увидеть доступные номинальные поля (и их значения), данные уже должны быть полностью инстанциированы при помощи источника восходящего потока или узла типа. Дополнительную информацию смотрите в разделе “Узел Тип” на стр. 128.

Расширение имени поля. Выберите эту опцию для включения элементов управления, позволяющих задать расширение, добавляемое к имени нового флагового поля как суффикс или префикс. По умолчанию имена новых полей автоматически создаются путем объединения имени исходного поля со значением поля и образования метки (такой как *Им-поля_значение-поля*).

Доступные значения набора. Здесь описаны значения, выбранные в наборе выше. Выберите одно или несколько значений, для которых вы хотите сгенерировать флаги. Например, если в поле, называемом *артериальное_давление*, используются значения *высокое*, *среднее* и *низкое*, можно выбрать *высокое* и добавить его в список справа. Тогда будет создано поле с флагом для записей со значением, указывающим высокое артериальное давление.

Создать поля флагов. Здесь выводятся вновь созданные флаговые поля. Можно задать опции именования новых полей при помощи элементов, управляющих расширением имен полей.

Значение True. Задайте значение true для использования узлом при задании флага. По умолчанию это значение - **T**.

Значение False. Задайте значение false для использования узлом при задании флага. По умолчанию это значение - **F**.

Ключи агрегации. Выберите эту опцию для объединения записей в группы на основе заданных выше полей ключей. При включенной опции **Ключи агрегации** все флаговые поля в группе будут "включены", если *любая* запись была задана как true. При помощи средства выбора полей укажите, какие поля ключей будут использоваться для агрегирования записей.

Узел реструктуризации

Узел реструктуризации модно использовать для генерирования нескольких полей на основе значения номинального или флагового поля. Вновь сгенерированные поля могут содержать значение из другого поля или числовые флаги (0 и 1). Функции этого узла аналогичны функциям узла Задать как флаг. Однако предлагаемая эффективность более высока. Этот узел позволяет создавать поля любого типа (включая числовые флаги) при помощи значений из другого поля. Затем можно выполнить операцию агрегации или другие операции с другими узлами вниз по потоку. (Узел Задать как флаг позволяет агрегировать поля одним шагом, что может оказаться удобным при создании флаговых полей.

Например, следующий набор данных содержит номинальное поле *Расчетный счет* со значениями *Savings* и *Draft*. Для каждого расчетного счета записываются начальный баланс и текущий баланс, и у некоторых клиентов есть несколько счетов каждого типа. Скажем, вы хотите знать, у каждого ли заказчика есть расчетный счет конкретного типа, а если есть, то сколько у него денег на счету каждого типа. При помощи

узла реструктуризации вы генерируете поля для каждого из значений поля *Расчетный счет* и в качестве значения выбираете *Текущий_баланс*. Каждое поле будет заполнено значение текущего баланса для данной записи.

Таблица 28. Данные выборки перед реструктуризацией.

CustID	Расчетный счет	Open_Bal	Current_Bal
12701	Draft	1000	1005,32
12702	Savings	100	144.51
12703	Savings	300	321,20
12703	Savings	150	204,51
12703	Draft	1200	586,32

Таблица 29. Данные выборки после реструктуризации.

CustID	Расчетный счет	Open_Bal	Current_Bal	Account_Draft_ Current_Bal	Account_Savings _Current_Bal
12701	Draft	1000	1005,32	1005,32	$null$
12702	Savings	100	144,51	$null$	144.51
12703	Savings	300	321,20	$null$	321,20
12703	Savings	150	204,51	$null$	204,51
12703	Draft	1200	586,32	586,32	$null$

Использование узла реструктуризации с узлом агрегации

Во многих случаях узел реструктуризации полезно объединить в пару с узлом агрегации. В предыдущем примере у одного клиента (с ID 12703) было три счета. Узел агрегации позволяет вычислить общий баланс для каждого типа расчетных счетов. Полем ключа будет *CustID*, а полями агрегации - новые реструктурированные поля *Account_Draft_Current_Bal* и *Account_Savings_Current_Bal*. Результаты показаны в следующей таблице.

Таблица 30. Данные выборки после реструктуризации и агрегации.

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

Задание опций для узла реструктуризации

Доступные поля. Возвращает список всех полей со шкалой измерений *Номинальная* (набора) или *Флаговая*. Выберите нужное в списке для вывода значений в наборе флагов, затем выберите из них значения, чтобы создать реструктурированные поля. Имейте в виду, чтобы можно было увидеть доступные поля (и их значения), данные уже должны быть полностью инстанциированы при помощи источника восходящего потока или узла типа. Дополнительную информацию смотрите в разделе “Узел Тип” на стр. 128.

Доступные значения. Здесь описаны значения, выбранные в наборе выше. Выберите одно или несколько значений, для которых вы хотите сгенерировать реструктурированные поля. Например, если в поле, называемом *Артериальное давление*, используются значения *высокое*, *среднее* и *низкое*, можно выбрать *высокое* и добавить его в список справа. Тогда для записей со значением *высокое* будет создано поле с заданным значением (смотрите ниже).

Создать реструктурированные поля. Здесь выводятся вновь созданные реструктурированные поля. По умолчанию имена новых полей автоматически создаются путем объединения имени исходного поля со значением поля и образования метки (такой как *Им-поля_значение-поля*).

Включить имя поля. Отключите эту опцию, чтобы убрать используемое как префикс имя исходного поля из имен новых полей.

Использовать значения из других полей. Задайте одно или несколько полей, значения которых будут использоваться для заполнения реструктурированных полей. Выбрать одно или несколько полей можно при помощи средства выбора полей. Для каждого выбранного поля создается одно новое поле. В конец имени реструктурированного поля добавляется имя поля значения, например: *АД_высокое_возраст* или *АД_низкое_возраст*. Каждое новое поле наследует тип исходного поля значения.

Создать флаги числовых значений. Выберите эту опцию, чтобы заполнить новые поля флагами числовых значений (0 для false и для true) вместо использования значения из другого поля.

Узел транспонирования

По умолчанию столбцы являются полями, а строки - записями или наблюдениями. Узел транспонирования позволяет при необходимости поменять данные в строках и столбцах так, чтобы поля стали записями, а записи - полями. Например, если в данных временных рядов каждый ряд представляет собой не столбец, а строку, эти данные можно транспонировать перед анализом.

Задание опций для узла транспонирования

Имена новых полей

Имена новых полей могут быть сгенерированы автоматически на основе заданного префикса или считаны из существующего поля в данных.

Использовать префикс. Эта опция генерирует имена полей автоматически на основе заданного префикса (*Поле1*, *Поле2* и так далее). Префикс можно настроить нужным вам образом. При выборе этой опции нужно задать число создаваемых полей, независимо от числа строк в исходных данных. Например, если задать **Число новых полей**, равное 100, все данные после первых 100 строк будут отброшены. Если исходные данные содержат менее 100 строк, некоторые поля будут пустыми. (Число полей можно увеличить нужным вам образом, но цель этого параметра - избежать транспонирования миллиона записей в миллион полей, что привело бы к генерированию не поддающегося управлению результату.)

Допустим, у вас есть данные с рядами в строках и отдельным полем (столбцом) для каждого месяца. Их можно транспонировать так, чтобы каждый ряд находится в отдельном поле со строкой для каждого месяца.

Читать из поля. Считывает имена полей из существующего поля. При выборе этой опции число новых полей определяется по данным, до заданного максимума. Каждое значение выбранного поля становится в выходных данных новым полем. У выбранного поля может быть любой тип хранения (целочисленный, строковый, дат и так далее), но во избежание повторяющихся имен полей число переменных должно совпадать с числом строк). Если встречаются дубликаты имен полей, выводится предупреждение.

- **Прочитать значения.** Если выбранное поле не было инстанциировано, выберите эту опцию, чтобы заполнить список имен новых полей. Если это поле уже было инстанциировано, этот шаг не требуется.
- **Максимальное количество значений для чтения.** При чтении имен полей из данных задается верхний предел, позволяющий избежать чрезмерно большого числа полей. (Как отмечено выше, транспонирование одного миллиона записей в один миллион полей может сгенерировать результат, не поддающийся управлению.

Например, если первый столбец в ваших данных задает имя для каждого ряда, эти значения можно использовать как имена полей в транспонированных данных.

Транспонировать. По умолчанию транспонируются только непрерывные поля (числового диапазона); либо с целочисленным типом хранения, либо с типом хранения действительных чисел. Дополнительно можно выбрать поднабор числовых полей либо транспонировать вместо них строковые поля. Однако у всех транспонируемых полей должен быть один и тот же тип хранения (числовой или строковый, но не оба), поскольку при смешивании входных полей могут быть сгенерированы в каждом выходном столбце смешанные значения, а это нарушает правило, что у всех значений поля должно быть одно и то же хранение. Другие типы хранения (дат, времени, отметок времени) транспонировать нельзя.

- **Все числовые.** Транспонирует все поля (с целочисленным типом хранения или типом хранения действительных чисел). Число строк в выводе совпадает с числом числовых полей в исходных данных.
- **Все строковые.** Транспонирует все строковые поля.
- **Пользовательские.** Позволяет выбрать поднабор числовых полей. Число строк в выводе совпадает с числом выбранных полей. *Примечание*: Эта опция доступна только для числовых полей.

Имя ID строки. Задает имя поля ID строки, созданного узлом. Значения этого поля определяются именами полей в исходных данных.

Совет: При транспонировании данных временных рядов из строк в столбцы, если данные содержат строку (такую как дата, месяц или год), помечающую период для каждого измерения, обязательно передайте эти метки в IBM SPSS Modeler в качестве имен полей (как в примерах выше, где в качестве имен полей в исходных данных показаны месяц или дата соответственно) вместо включения метки в первую строку данных. Это предотвратит смешивание меток и значений в каждом столбце (что привело бы к чтению чисел как строк, поскольку в столбце не могут быть смешанные типы хранения) .

Узел хронологии

Узлы хронологи часто используются для последовательных данных, таких как данные временных рядов. Они используются для создания новых полей, содержащих данные из полей в прежних записях. При использовании узла хронологии могут оказаться полезны данные, предварительно отсортированные по конкретному полю. Для получения таких данных можно использовать узел сортировки.

Задание опций для узла хронологии

Выбранные поля. При помощи средства выбора полей (кнопки справа от текстового поля) выберите поля, для которых вам нужна хронология. Каждое выбранное поле будет использоваться для создания новых полей для всех записей в наборе данных.

Смещение. Задайте самую последнюю запись перед текущей записью, из которой вы хотите извлечь хронологические значения поля. Например, если задано смещение 3, по мере прохождения каждой записи через этот узел значения поля для третьей записи предварительно будут включаться в текущую запись. При помощи параметров промежутка времени укажите период, за который будут извлекаться записи. Для настройки значения смещения используйте кнопки со стрелками.

Диапазон. Укажите число записей, для которых вы хотите извлечь значения. Например, если задать смещение 3 и промежуток времени 5, каждая запись, проходящая через узел, будет содержать пять полей, добавленных в нее для каждого поля, указанного в списке Выбранные поля. Это означает, что при обработке узлом записи 10 поля будут добавлены из записей с 7 по 3.

Где доступна хронология. Для обработки записей, не содержащих хронологических значений, выберите одну из приведенных ниже опций. Обычно это относится к первым нескольким записям в начале набора данных, для которых не существует предварительных записей для использования в качестве хронологических.

- **Отклонить записи.** Выберите эту опцию для отбрасывания записей, где ни одно хронологическое значение не доступно для выбранного поля.
- **Оставить хронология неопределенной.** Выберите эту опцию, чтобы сохранять записи, не содержащие доступных хронологических значений. Поле хронологии будет заполнено неопределенным значением, выводящимся как `$null$`.

- **Чем заполнять значения.** Задайте значение или строку, которые следует использовать для записей, где недоступно хронологическое значение. Значение замены по умолчанию - *undef* (системное пустое значение). Пустые значения выводятся строкой `$null$`.

При выборе значения замены следует помнить следующие правила правильного заполнения:

- У выбранных полей должен быть один и тот же тип хранения.
- Если у всех выбранных полей будет числовой тип хранения, значение замены должно быть синтаксически проанализировано как целое число.
- Если у всех выбранных полей будет тип хранения действительных чисел, значение замены должно быть синтаксически проанализировано как действительное число.
- Если у всех выбранных полей будет символический тип хранения, значение замены должно быть синтаксически проанализировано как строка.
- Если у всех выбранных полей будет тип хранения даты-времени, значение замены должно быть синтаксически проанализировано как поле даты-времени.

Если не будет соблюдаться хотя бы одно из вышеуказанных условий, при вызове узла хронологии вы получите сообщение об ошибке.

Узел переупорядочения полей

Узел переупорядочения полей позволяет определить естественный порядок, используемый для вывода полей вниз по потоку. Этот порядок влияет на вывод полей в самых разнообразных местах, таких как таблицы, списки и средство выбора полей. Эта операция, полезная, например, при работе с наборами данных, позволяет сделать поля более наглядными.

Задание опций переупорядочения полей

Существует два способа переупорядочения полей: пользовательское упорядочение и автоматическая сортировка.

Пользовательское упорядочение

Выберите опцию **Пользовательский порядок**, чтобы включить поддержку таблицы имен полей, где можно просмотреть все поля и при помощи кнопок со стрелками создать пользовательский порядок.

Чтобы переупорядочить поля:

1. Выберите поле в таблице. Для выбора нескольких полей используйте клавишу Ctrl+щелчок мыши.
2. Для перемещения полей на одну позицию вверх или вниз используйте кнопки с простыми стрелками.
3. Для перемещения полей в самый низ или верх списка используйте кнопки с линейными стрелками.
4. Задайте порядок не включенных сюда полей, переместив вверх или вниз строку-разделитель с индикатором `[другие поля]`.

Дополнительная информация об элементе [другие поля]

Другие поля. Цель строки-разделителя **[другие поля]** - разделение таблицы на две половины.

- Поля, выводящиеся выше этого разделителя, будут упорядочены (в соответствии с их порядком в таблице) в самом верху списка всех естественных порядков, используемых для вывода нисходящего потока полей этого узла.
- Поля, выводящиеся ниже этого разделителя, будут упорядочены (в соответствии с их порядком в таблице) в самом низу списка всех естественных порядков, используемых для вывода нисходящего потока полей этого узла.

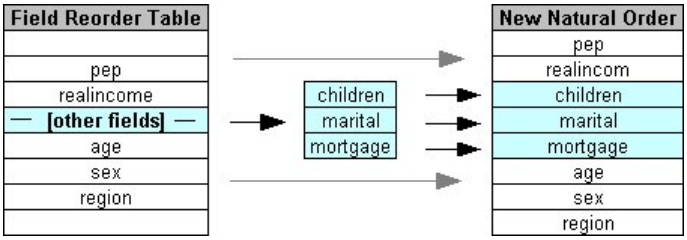


Рисунок 6. Диаграмма, показывающая, как “другие поля” вписываются в новый порядок полей

- Все остальные поля, не выводящиеся в таблице переупорядочения полей, будут выводиться между этими “верхними” и “нижними” полями, указываемыми размещением строки-разделителя.

В число дополнительных опций пользовательской сортировки входят:

- Сортировка полей по возрастанию или убыванию посредством нажатия кнопок со стрелками над каждым заголовком столбца (**Тип**, **Имя** и **Хранение**). При сортировке по столбцу не специфицированные здесь поля (указываемые строкой `[другие поля]`) сортируются последними, по своему естественному порядку.
- Нажмите кнопку **Очистить неиспользуемые**, чтобы удалить все неиспользуемые поля с узла переупорядочения полей. Неиспользуемые поля выводятся в таблице шрифтом красного цвета. Цвет указывает, что поле было удалено в операциях восходящего потока.
- Задайте упорядочение для всех новых полей (выводимых со значком молнии для указания нового или неспецифицированного поля). При нажатии кнопки **OK** или **Применить** этот значок исчезает.

Примечание: В случае добавления вверх по потоку новых полей после применения пользовательского порядка они будут добавляться в самый низ пользовательского списка.

Автоматическая сортировка

Выберите опцию **Автоматическая сортировка**, чтобы задать параметр для сортировки. Состав опций этого диалогового окна изменяется динамически, предоставляя опции для автоматической сортировки.

Сортировать по. Выберите один из трех способов сортировки полей, передаваемых на узел переупорядочения. Кнопки со стрелками указывают, будет ли порядок возрастающим или убывающим. Выберите что-то одно, чтобы внести изменения.

- Имя
- Тип
- Хранение

Поля, добавляемые в восходящий поток узла переупорядочения полей после применения автоматической сортировки будут автоматически размещаться на правильных для них позициях на основе выбранного типа сортировки.

Узел интервалов времени

Исходный узел Интервалы времени в SPSS Modeler несовместим с Analytic Server (AS) и был объявлен устаревшим в SPSS Modeler выпуск 18.0.

Заменяющий узел Интервалы времени (новый в SPSS Modeler выпуска 17.0) содержит подмножество функций исходного узла Интервалы времени, которые можно использовать с Analytic Server.

Узел Интервалы времени используется для указания интервалов и получения нового поля времени для оценки или прогнозирования. Поддерживается весь диапазон интервалов времени от секунд до лет.

Используйте этот узел для вывода нового поля времени; у этого нового поля будет такой же тип хранения, как у выбранного входного поля времени. Этот узел генерирует следующие элементы:

- Поле, заданное на вкладке Поля как **Поле времени**, с выбранным префиксом/суффиксом. По умолчанию префикс - `$TI_`.
- Поля, заданные на вкладке Поля как **Поля измерений**.
- Поля, заданные на вкладке Поля как **Поля агрегирования**.

Может быть также сгенерирован ряд дополнительных полей, в зависимости от выбранного интервала или периода (такого как минута или секунда, на которые попадает измерение).

Интервал времени - опции полей

Используйте вкладку Поля на узле Интервалы времени для выбора данных, из которых будет получен новый интервал.

Поля Выводит все входные поля для узла со значками их типов измерений. У всех полей времени тип измерения - 'количественный'. Выберите поле, которое будет использоваться в качестве входного.

Поле Время Содержит входное поле, из которого будет получаться новый интервал времени; допускается только одно непрерывное поле. Это поле используется узлом Интервалы времени как ключ агрегации для преобразования интервала. У нового поля будет такой же тип хранения, как у выбранного входного поля времени. Если выбрано целочисленное поле, оно рассматривается как индекс времени.

Поля измерений Дополнительно здесь можно добавить поля для создания конкретного временного ряда, основанного на значениях этого поля. Например, при работе с геопространственными данными в качестве измерения можно использовать поле точек. В этом примере выходные данные с узла Интервалы времени сортируются во временные ряды для каждого значения точки в поле точек.

Поля для агрегации Выберите поля, которые будут агрегироваться как часть изменения периода поля времени. На вкладке Построить для таблицы **Пользовательские параметры для заданных полей** доступны только выбранные здесь поля. Все невыбранные поля отфильтровываются из данных, покидающих узел. Это означает, что все оставшиеся в списке **Поля** поля отфильтровываются из данных.

Интервал времени - опции построения

Используйте вкладку Построить, чтобы указать опции для изменения временного интервала и информацию о том, как будут агрегироваться данные, на основании их типов измерения.

При агрегировании данных все существующие поля дат или отметок времени эффективно заменяются генерируемыми полями и отбрасываются из вывода. Другие поля агрегируются на основе опций, задаваемых на этой вкладке.

Интервал времени Выберите интервал и периодичность построения для рядов. Дополнительную информацию смотрите в разделе Поддерживаемые интервалы.

Параметры по умолчанию Выберите, какое агрегирование будет применяться к данным разного типа по умолчанию. Функция по умолчанию применяется на основе шкалы измерения; например, непрерывные поля агрегируются при помощи суммирования, тогда как для номинальных полей используется мода. Можно задать функции по умолчанию для трех разных шкал измерения:

- **Количественная** Доступные функции для непрерывных полей включают в себя **Сумму**, **Среднее**, **Минимум**, **Максимум**, **Медиану**, **1-й квартиль** и **3-й квартиль**.
- **Номинальная** Доступные опции - **Мода**, **Минимум** и **Максимум**.
- **Флаг** Доступные опции **True, если любое - true** или **False, если любое - false**.

Пользовательские параметры для заданных полей Для отдельных полей можно задать исключения из параметров агрегации по умолчанию. Используйте значки справа, чтобы добавить или удалить поля из таблицы, или щелкните по нужному столбцу, чтобы изменить функцию агрегации для данного поля. Поля без типа исключаются из списка, и добавить их в таблицу нельзя.

Расширение имени нового поля Задайте **Префикс** или **Суффикс**, применимые ко всем полям, сгенерированным узлом.

Узел перепроектирования

При работе с геопространственными данными или с данными карт два наиболее общих способа, используемых для определения координат, - это системы проективных и географических координат. В IBM SPSS Modeler такие элементы как пространственные функции Построителя выражений, узел пространственно-временных предсказаний (Spatio-Temporal Prediction, STP) и узел визуализации карт используют проективную систему координат, и поэтому любые импортируемые данные, которые используют географическую систему координат, нужно перепроектировать. Если возможно, при использовании геопространственные поля (все поля с геопространственной шкалой измерения) перепроектируются автоматически (не в случаях, когда они импортируются). Когда какие-то поля невозможно перепроектировать автоматически, для изменения системы координат используется узел перепроектирования. Возможность такого перепроектирования означает, что вы можете исправить ситуацию, когда происходит ошибка из-за использования неправильной системы координат.

Примеры ситуаций, когда может потребоваться перепроектирование для изменения системы координат, показаны в следующем списке:

- *Присоединить* При попытке присоединить два набора данных с разными системами координат к геопространственному полю SPSS Modeler выводит следующее сообщение об ошибке: `Системы координат поля <Поле1> и поля <Поле2> несовместимы. Перепроектируйте одно или оба поля для использования одинаковой системы координат.`

  <Поле1> и *<Поле2>* - это имена геопространственных полей, из-за которых произошла ошибка.
- *Выражение if/else* Если вы используете выражение, содержащее оператор if/else, с геопространственными полями или возвращаете эти типы в обоих частях выражения, но с разными системами координат, SPSS Modeler выводит следующее сообщение об ошибке: `Условное выражение содержит несовместимые системы координат: <аргумент1> и <аргумент2>.`

  <аргумент1> и *<аргумент2>* это аргументы операторов then и else, которые возвращают данные геопространственного типа с разными системами координат.
- *Конструирование списка геопространственных полей* Чтобы создать поле списка, содержащее несколько геопространственных полей, все такие поля, передаваемые в выражение списка, должны использовать одинаковую систему координат. Если это не так, появится следующее сообщение об ошибке: `Системы координат поля <Поле1> и поля <Поле2> несовместимы. Перепроектируйте одно или оба поля для использования одинаковой системы координат.`

Дополнительную информацию о системе координат смотрите в теме Определение геопространственных опций раздела Работа с потоками в Руководстве пользователя SPSS Modeler.

Задание опций для узла Перепроектировать

Поля

Геопространственные поля

По умолчанию этот список пуст. Если вам надо, чтобы некоторые геопространственные поля не перепроектировались, их можно перенести в этот список из списка **Поля для перепроектирования**.

Поля для перепроектирования

По умолчанию этот список содержит все входные геопространственные поля этого узла. Все поля в этом списке перепроектируются к системе координат, заданной в области **Система координат**.

Система координат

Как в потоке по умолчанию
Выберите эту опцию для использования системы координат по умолчанию.

Задать Если выбрана эта опция, можно нажать кнопку **Изменить** для вывода диалогового окна Выбрать систему координат и выбрать систему координат, которая будет использоваться для перепроектирования.

Дополнительную информацию о системе координат смотрите в теме Определение геопространственных опций раздела Работа с потоками в Руководстве пользователя SPSS Modeler.

Глава 5. Узлы диаграмм

Общие возможности узлов диаграмм

В нескольких фазах процесса исследования данных используются графики и диаграммы для изучения переданных в IBM SPSS Modeler данных. Например, можно соединить узел Построение графиков или Распределение с источником данных, чтобы получить информацию о типах и распределениях данных. Затем можно выполнить преобразования записей и полей для подготовки данных к операциям нисходящего моделирования. Другое общее использование графиков - это проверка распределения и взаимосвязей между вновь полученными значениями полей.

Палитра Графики содержит следующие узлы:

Узел Панель выбора диаграмм предлагает много разных типов диаграмм на одном узле. Используя этот узел, можно выбрать поля данных, которые вы хотите изучать, а затем выбрать диаграмму из доступных для выбранных данных. Узел автоматически отфильтровывает все типы диаграмм, которые нельзя использовать для работы с выбранными полями.

Узел График показывает взаимосвязь между численными полями. Графики можно создавать, используя точки (диаграммы рассеяния) или линии.

Узел распределения показывает появление символических (категориальных) значений, таких как тип закладных или пол. Обычно узел распределения используется для показа разбалансировки данных, которую можно выправить при помощи узла балансировки до создания модели.

Узел Гистограмма показывает существующие значения для числовых полей. Он часто используется для изучения данных перед работой с ними и построением моделей. Аналогично узлу Распределение узел Гистограмма часто выявляет несбалансированность данных.

Узел Собрание показывает распределение значений для одного числового поля относительно значений другого. (При этом создаются диаграммы, похожие на гистограммы). Это полезно для иллюстрации переменной или поля, значения которых изменяется во времени. Используя 3D-представление, вы можете включить также символическую ось, показывающую распределения по категориям.

Узел нескольких графиков (Multiplot) создает график, выводящий несколько полей *Y* по отношению к одному полю *X*. Значения полей *Y* изображаются на графике как цветные линии, каждая из которых эквивалентна графику на узле График при заданном значении стиля **Линия** и режиме X **Сортировка**. Узел Несколько графиков полезен, когда нужно исследовать флуктуации нескольких переменных во времени.

Узел Web иллюстрирует силу взаимосвязи между значениями двух или более символических (категориальных) полей. На графике используются линии разной ширины для обозначения силы соединения. Например, вы можете использовать узел Web для изучения взаимосвязи между покупкой набора товаров на сайте интернет-магазина.

Узел Временной график выводит один или несколько наборов данных временных рядов. Обычно вы сначала используете узел Временные интервалы для создания поля *TimeLabel*, которое будет использовано для отметок по оси *x*.

Узел Оценка помогает оценить и сравнить прогнозирующие модели. Диаграмма оценки показывает, насколько хорошо модели предсказывают конкретные выходные данные. Он сортирует записи на основе предсказанного значения и доверительного интервала предсказания. Он разбивает записи на группы равного размера (**квантили**) и затем выводит значение бизнес-критерия для каждой квантили от самой высокой до самой низкой. Несколько моделей представляются разными линиями на графике.

Узел Визуализация карты может принять несколько входных соединений и вывести геопространственные данные на карту как несколько слоев. Каждый слой - это одно геопространственное поле; например, базовым слоем может быть карта страны, выше может накладываться один слой дорог, один слой рек и один слой городов.

После добавления узла графиков к потоку можно дважды щелкнуть по узлу, чтобы открыть диалоговое окно для задания опций. Большинство графиков содержит несколько уникальных опций, представленных на одной или нескольких вкладках. На этих вкладках содержатся также несколько опций, общих для всех графиков. Следующие и разделы содержат дополнительную информацию об этих общих опциях.

После конфигурирования опций для узла графиков его можно запустить из диалогового окна или как часть потока. В сгенерированном окне графиков можно создать узлы Получение (Набор и Флаг) и Выбор на основе выбора или региона данных, создавая тем самым эффективные подмножества данных. Например, эту мощную возможность можно использовать для обнаружения и исключения выбросов.

Эстетики, наложения, панели и анимация

Наложения и эстетики

Эстетики (и перекрытия) позволяют увеличивать размерность визуализаций. Эффект эстетики (группировка, кластеризация или стыкование) зависит от типа визуализации, типа полей/переменных, а также типа географических элементов и статистик. Например, категориальное поле/переменная для цвета можно использовать для группировки точек в диаграмме рассеяния или создания составных столбцов в столбчатой диаграмме со стыкованием. Количественные переменные для цвета можно использовать для обозначения значений, лежащих в диапазоне, для каждой точки на диаграмме рассеяния.

Следует пробовать применять разные варианты эстетик и перекрытий, чтобы найти те, которые будут удовлетворять вашим требованиям. Следующие описания помогут выбрать правильную эстетику и перекрытие.

Примечание: Не все эстетики и перекрытия доступны для всех типов визуализаций.

- **Цвет.** Когда цвет задан категориальной переменной/полем, визуализация разделяется на части на основании отдельных категорий, по одному цвету на каждую категорию. Когда цвет задан количественной переменной, цвет изменяется на основании значений количественной переменной/поля. Если графический элемент (например, столбец или ящик) представляет несколько записей/наблюдений и в качестве переменной цвета используется количественная переменная/поле, то цвет изменяется на основании *среднего значения* количественной переменной/поля.
- **Форма.** Форма задана категориальной переменной/полем, которое разделяет визуализацию на элементы различной формы, по одной форме на каждую категорию.
- **Прозрачность.** Если прозрачность задана при помощи категориального поля/переменной, визуализация разделяется на основе отдельных категорий - по одному уровню прозрачности для каждой категории. Если прозрачность задана количественным полем/переменной, прозрачность изменяется на основе

значений количественного поля/переменной. Если графический элемент (например, столбец или ящик) представляет несколько записей/наблюдений и для прозрачности используется количественное поле/переменная, цвет изменяется в зависимости от *среднего значения* количественного поля/переменной. При максимальном значении графические элементы полностью прозрачны. При минимальном значении они полностью непрозрачны.

- **Метка данных.** Метки данных заданы переменной (полем) любого типа, значения которой используются для задания меток элементов диаграммы.
- **Размер.** Когда размер задан категориальной переменной/полем, визуализация разделяется на части на основании отдельных категорий, по одному размеру на каждую категорию. Когда размер задан количественной переменной, цвет изменяется на основании значений количественной переменной/поля. Если графический элемент (например, столбец или ящик) представляет несколько записей/наблюдений и в качестве переменной размера используется количественная переменная/поле, цвет изменяется на основании *среднего значения* количественной переменной/поля.

Управление панелями и анимацией

Формирование панелей. При использовании панелей создаются таблицы диаграмм. Для каждой категории полей/переменных панели создается одна диаграмма, но все панели показываются одновременно. Панели полезны для проверки влияния полей/переменных панели на визуализацию. Например, можно создать гистограмм с панелями по полу, чтобы выяснить одинаковы ли частотные распределения для мужчин и для женщин. В частности, можно проверить, зависит ли зарплата от пола. Выберите категориальные поле/переменную для формирования панели.

Анимация. Анимация аналогична панелям, поскольку создает несколько диаграмм на основе значений поля/переменной анимации, однако эти диаграммы не показываются вместе. Вместо этого приходится использовать режим изучения диаграмм для анимации вывода и просмотра последовательности отдельных диаграмм. Кроме того, в отличие от панелей анимация не требует именно категориального поля/переменной. Можно указать непрерывное поле/переменную, значения которой будут разделяться на диапазоны автоматически. В режиме изучения можно изменять размеры диапазонов с помощью управляющих элементов анимации. Не все визуализации предполагают анимацию.

Использование вкладки Вывод

Для имен файлов и вывода генерируемых диаграмм для всех типов диаграмм можно задать следующие опции.

Примечание: У диаграмм узлов распределения есть дополнительные параметры.

Имя вывода. Задает имя диаграммы, генерируемой при вызове узла. Опция **Авто** выбирает имя на основе узла, генерирующего вывод. Дополнительно можно выбрать **Пользовательское**, чтобы задать другое имя.

Вывод на экран. Выберите эту опцию для генерирования и вывода диаграммы в новом окне.

Вывод в файл. Выберите эту опцию для сохранения вывода в виде файла.

- **График вывода.** Выберите эту опцию для генерирования вывода в формате диаграммы. Доступна только на узлах распределения.
- **Выходная таблица.** Выберите эту опцию для генерирования вывода в формате таблицы. Доступна только на узлах распределения.
- **Имя файла.** Задайте имя файла, используемое для сгенерированной диаграммы или таблицы. Чтобы задать конкретный файл и положение, нажмите кнопку просмотра **(...)**.
- **Тип файла.** Задайте тип файла в выпадающем списке. Для всех узлов диаграмм, кроме узла распределения с опцией **Выходная таблица**, доступны следующие типы файлов диаграмм:

  - Точечный рисунок (*.bmp*)

  - PNG (*.png*)

- Объект вывода (*.cou*)
- JPEG (*.jpg*)
- HTML (*.html*)
- Документ ViZml (*.xml*) для использования в других прикладных программах IBM SPSS Statistics.

Для опции **Выходная таблица** на узле распределения доступны следующие типы файлов переменных.

- Данные с разделителем-табулятором (*.tab*)
- Данные с разделителем-запятой (*.csv*)
- HTML (*.html*)
- Объект вывода (*.cou*)

Разбить вывод на страницы. Эта опция включается при сохранении вывода в виде HTML, упрощая управление размером каждой страницы HTML. (Применяется только для узла распределения.)

Строк на страницу. Эта опция включается, если выбрана опция **Разбить вывод на страницы**, позволяя определить длину каждой страницы HTML. Значение по умолчанию - 400 строк. (Применяется только для узла распределения.)

Использование вкладки Аннотации

Используемая для всех узлов эта вкладка предлагает опции для их переименования, задания пользовательских подсказок и хранения длинных аннотаций.

Трехмерные диаграммы

Для графиков и диаграмм собраний в IBM SPSS Modeler предусмотрена возможность вывода информации по третьей оси. Это обеспечивает дополнительную гибкость визуализации данных для выбора поднаборов или получения новых полей для моделирования.

После создания трехмерной диаграммы по ней можно щелкнуть и перетащить указателем мыши, чтобы повернуть и просмотреть под любым углом.

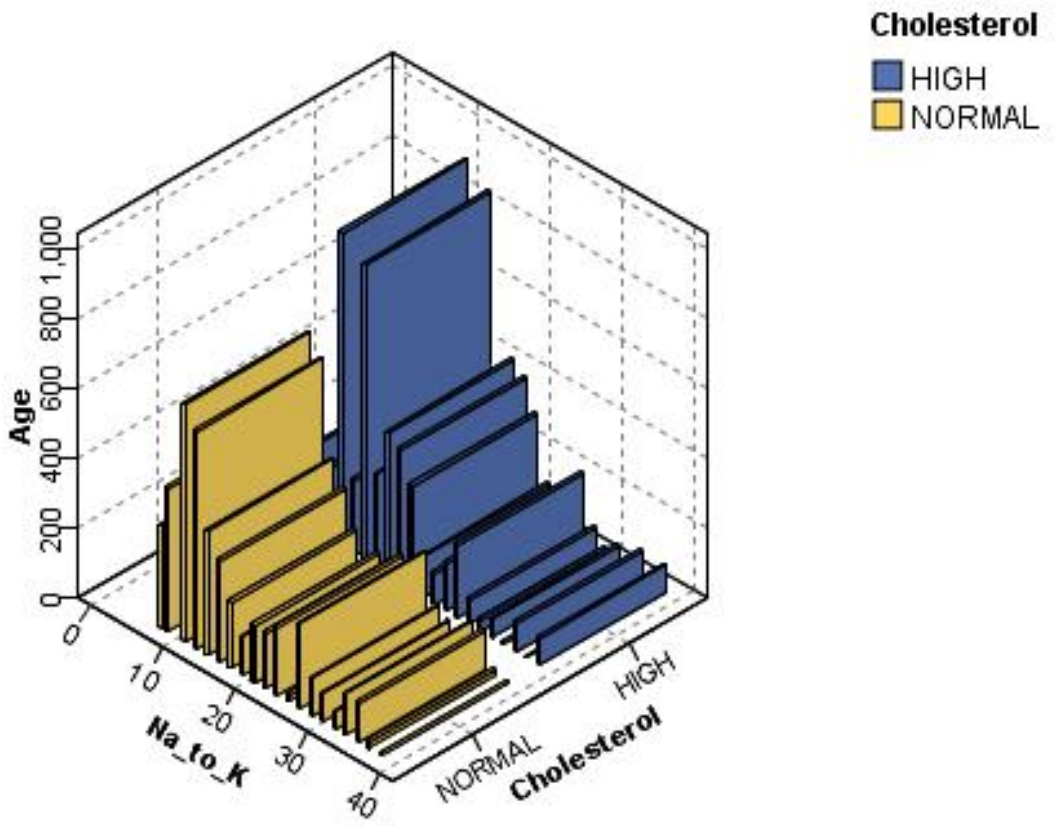


Рисунок 7. Диаграмма собрания с осями x, y и z

В IBM SPSS Modeler существует два способа построения трехмерных диаграмм: методом нанесения данных на третью ось (настоящие трехмерные диаграммы) и посредством вывода диаграмм с трехмерными эффектами. Оба метода доступны для графиков и диаграмм собраний.

Чтобы нанести данные на третью ось диаграммы

1. В диалоговом окне узла диаграмм перейдите на вкладку **График**.
2. Нажмите кнопку Трехмерная, чтобы включить поддержку опций для оси *z*.
3. С помощью кнопки Средство выбора полей выберите поле для оси *z*. В некоторых случаях здесь будут доступны только символические поля. Средство выбора полей выведет соответствующие поля.

Чтобы добавить трехмерные эффекты на диаграмму

1. Создав диаграмму, щелкните по вкладке **Диаграмма** в окне вывода.
2. Нажмите кнопку Трехмерная, чтобы переключить представление на трехмерную диаграмму.

Узел Панель выбора диаграмм

Узел Панель выбора диаграмм позволяет выбирать из множества вариантов вывода диаграмм (столбчатые диаграммы, круговые диаграммы, гистограммы, диаграммы рассеяния, heatmap и другие) в одном узле. Сначала на первой вкладке вы выбираете поля данных, которые хотите исследовать, а затем узел предоставляет возможность выбора типа диаграмм, которые будут использоваться для данных. Узел

автоматически отфильтровывает все типы диаграмм, которые нельзя использовать для работы с выбранными полями. На вкладке Подробности можно определить более подробные или дополнительные опции диаграмм.

Замечание: Узел Панель выбора диаграмм необходимо соединить с потоком данных, чтобы проводить изменения в узле или выбирать типы диаграмм.

Есть две кнопки, которые позволяют управлять доступностью шаблонов визуализации (и таблиц стилей, и карт):

Управление . Управление шаблонами визуализаций, таблицами стилей и картами на компьютере. На вашем локальном компьютере можно импортировать, экспортировать, переименовывать и удалять шаблоны визуализации , таблицы стилей и отображения. Дополнительную информацию смотрите в разделе “Управление шаблонами, таблицами стилей и файлами карт” на стр. 209.

Местоположение. Изменение папки, в которой хранятся шаблоны визуализаций, таблицы стилей и карты Текущее местоположение показывается справа от кнопки. Дополнительную информацию смотрите в разделе “Указание местоположения для хранения шаблонов, таблиц стилей и карт” на стр. 208.

Вкладка Основные Панели выбора диаграмм

Если вы не уверены в том, какой тип визуализации лучше всего подойдет для данных, воспользуйтесь вкладкой Базовая. При выборе переменных выводятся все типы визуализаций, которые подходят для выбранных данных. Примеры смотрите в разделе “Примеры Панели выбора диаграмм (Graphboard)” на стр. 198.

1. Выберите в списке одно или несколько полей (переменных). Используйте клавишу Ctrl+щелчок кнопкой мыши, чтобы выделить несколько полей.

   Имейте в виду, что от шкалы измерений поля зависит тип доступных визуализаций. Шкалу измерений можно изменить, щелкнув правой кнопкой мыши по полю в списке и выбрав параметр. Более подробную информацию о доступных типах уровней измерений смотрите в разделе “Типы полей (переменных)” на стр. 186.
2. Выберите тип визуализации. Описание доступных типов смотрите в разделе “Доступные встроенные типы визуализации Панели выбора диаграмм” на стр. 190.
3. Для некоторых визуализаций можно выбрать итожащую статистику. Набор доступных статистик, зависит от того, рассчитывается ли статистика для категориального поля/переменной или для количественного поля/переменной. Доступные статистики также зависят от самого шаблона. Ниже приведен полный список доступных статистик.
4. Если требуется настроить дополнительные параметры, например эстетику и поля/переменные панели, выберите вкладку **Детальная**. Дополнительную информацию смотрите в разделе “Вкладка Подробности Панели выбора диаграмм” на стр. 188.

Итожащие статистики, вычисляемые для количественного поля

- *Mean*. Мера центральной тенденции. Арифметическое среднее; сумма, деленная на число наблюдений.
- *Медиана*. Значение, выше и ниже которого попадает по половине наблюдений, иначе 50-й процентиль. Если число наблюдений четно, медиана есть арифметическое среднее двух находящихся в середине значений, если выборку упорядочить по убыванию или по возрастанию. Медиана представляет собой меру центральной тенденции, которая нечувствительна к выбросам, в отличие от среднего значения, которое могут исказить несколько экстремально больших или малых значений.
- *Мода*. Чаще всего встречающееся значение. Если таких значений несколько, каждое из них является модой.
- *Минимум*. Наименьшее значение числовой переменной.
- *Максимум*. Наибольшее значение числовой переменной.
- *Диапазон*. Разность между максимальным и минимальным значениями.

- *Середина диапазона*. Середина диапазона, то есть значение, расстояние до минимума от которого равно расстоянию до максимума.
- *Sum*. Сумма или итог для всех значений по всем наблюдениям, имеющим непропущенные значения.
- *Накопленная сумма*. Накопленная сумма значений. Каждый графический элемент показывает сумму для одной подгруппы плюс общую сумму для всех предшествующих групп.
- *Процент суммы*. Процент внутри каждой подгруппы, основанный на суммированных полях, по отношению к сумме по всем группам.
- *Накопленный процент суммы*. Накопленный процент внутри каждой подгруппы, основанный на суммированных полях, по отношению к сумме по всем группам. Каждый графический элемент показывает процент для одной подгруппы плюс общий процент для всех предшествующих групп.
- *Дисперсия*. Мера дисперсии относительно среднего значения. Равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньшее числа наблюдений. Дисперсия измеряется в единицах, которые равны квадратам единиц измерения самой переменной.
- *Стандартное отклонение*. Мера дисперсии вокруг среднего, выраженная в тех же единицах измерения, что и наблюдения. Равна корню квадратному из дисперсии. При нормальном распределении 68% наблюдений укладываются в одно стандартное отклонение от среднего, и 95% - в два стандартных отклонения. Если, например, средний возраст равен 45 годам со стандартным отклонением 10, то 95% наблюдений должны оказаться между 25 и 65 годами при нормальном распределении.
- *Стандартная ошибка*. Мера того, насколько значение статистики критерия меняется от выборки к выборке. Это стандартное отклонение выборочного распределения статистики. Например, стандартная ошибка среднего - это стандартное отклонение выборочных средних.
- *Эксцесс*. Мера сгруппированности наблюдений вокруг центральной точки. Для нормального распределения значение эксцесса равно 0. Положительный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы более плотно около центра и имеют более тонкие хвосты до экстремумов распределения, и более толстые хвосты в области экстремальных значений. Отрицательный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы менее плотно около центра и имеют более толстые хвосты до экстремумов распределения, и более тонкие хвосты в области экстремальных значений.
- *Асимметрия*. Мера асимметрии распределения. Нормальное распределение симметрично, и для него асимметрия равна 0. Распределение со значимой положительной асимметрией имеет длинный хвост справа. Распределение со значимой отрицательной асимметрией имеет длинный хвост слева. В качестве грубого правила можно сказать, что значение асимметрии, более чем вдвое превышающее ее стандартную ошибку, указывает на наличие асимметрии распределения.

Следующие статистики области могут приводить к более чем одному графическому элементу на каждую подгруппу. При использовании графических элементов с интервалами, областями или границами статистики области могут приводить к одному графическому элементу, показывающему диапазон. Все остальные графические элементы приводят к двум отдельным элементам, один из которых показывает начало диапазона, а другой - конец диапазона.

- **Область: диапазон**. Разность между максимальным и минимальным значениями.
- **Область: 95% доверительный интервал для среднего**. Диапазон значений, который в 95% случаях включает среднее значение для генеральной совокупности.
- **Область: 95% доверительный интервал для конкретного значения.** Диапазон значений, который в 95% случаях включает предсказанное значение для отдельного наблюдения.
- **Область: Среднеквадратичное отклонение 1 выше или ниже среднего**. Диапазон значений между 1 *стандартным отклонением* выше и ниже *среднего*.
- **Область: Среднеквадратичная ошибка 1 выше или ниже среднего**. Диапазон значений между 1 *стандартной ошибкой* выше и ниже *среднего*.

Итожащие статистики, основанные на количестве.

- **Количество**. Число строк/наблюдений.

- **Накопленное количество**. Накопленное число строк/наблюдений. Каждый графический элемент показывает количество для одной подгруппы плюс общее количество для всех предшествующих групп.
- **Процент количеств**. Процент строк/наблюдений в каждой подгруппе по отношению к общему числу строк/наблюдений.
- **Накопленный процент количества**. Накопленный процент строк/наблюдений в каждой подгруппе по отношению к общему числу строк/наблюдений. Каждый графический элемент показывает процент для одной подгруппы плюс общий процент для всех предшествующих групп.

Типы полей (переменных)

В списке полей рядом с каждым полем показывается значок, который указывают на тип поля и тип данных. Значки также обозначают наборы множественных ответов.

Таблица 31. Значки уровня измерения.

Шкала измерений	Числовой	Текстовое	Дата	Время
Количественная		(не задается)		
Порядковая				
Установить				

Таблица 32. Значки наборов множественных ответов.

Тип набора множественных ответов	Значок
Набор множественных ответов, множественные категории	
Набор множественных ответов, множественные дихотомии	

Шкала измерений

При создании визуализации важна шкала измерений поля. Ниже приведены описания шкал измерений. Шкалу (уровень) измерений можно временно изменить, щелкнув правой кнопкой мыши по полю в списке полей и выбрав подходящий вариант. В большинстве случаев необходимо учитывать только две самые распространенные классификации категориальных и непрерывных полей:

Категориальную. Это данные с ограниченным числом уникальных значений или категорий (например, пол или религия). Категориальные поля могут быть текстовыми или числовыми, в которых категории закодированы числовыми кодами (например, 0 = *Женский*, а 1 = *Мужской*). Также эти данные называются качественными данными. Номинальные поля, порядковые поля и флаги являются категориальными полями.

- *Набор*. Поле/переменная, значения которого представляют категории без естественного упорядочивания (например, подразделение компании, в котором работает сотрудник). Примеры номинальных переменных включают регион, почтовый индекс или религию. Также называется номинальной переменной.
- *Упорядоченный набор*. Поле/переменная, значения которых представляют категории с некоторым естественным для них упорядочением (например, уровни удовлетворенности обслуживанием от крайней неудовлетворенности до полной удовлетворенности). К примерам упорядоченных наборов относятся оценки, представляющие степень удовлетворенности или уверенности, или баллы, оценивающие предпочтение. Также называются порядковой переменной.

- *Флаг*. Поле или переменная с двумя отдельными значениями, например Да и Нет, 1 и 2. Известны также как дихотомические или двоичные переменные.

Непрерывные. Это данные, измеренные на интервальной шкале или на шкале отношений, для которых существует и порядок значений, и расстояния между значениями. {\f3 Например, зарплата 7195 рублей больше зарплаты 5398 рублей, а расстояние между этими зарплатами }-{\f3 1797 рублей. Такие данные также называют непрерывными или числовым диапазоном.}

Категориальные поля служат для задания категорий в визуализации, обычно для вывода отдельного графического элемента для каждой категории или для группировки графических элементов. Количественные переменные часто подытоживаются внутри категорий категориальных переменных. К примеру, на визуализации с переменной дохода по категориям пола будет представлено среднее значение дохода для мужчин и среднее значение дохода для женщин. Исходные значения количественных переменных могут быть представлены при помощи диаграммы рассеяния. Например, диаграмма рассеяния может для каждого наблюдения показывать зарплату в настоящее время и зарплату при приеме на работу. Категориальную переменную можно использовать для группировки наблюдений по полу.

Типы данных

Шкала (уровень) измерения не единственное свойство переменной, которое определяет ее тип. Переменная хранится как конкретный тип данных. Данные бывают текстовыми (нечисловые данные, например буквы), числовыми (действительными числами) и датами. В отличие от шкалы измерения, тип данных переменной невозможно изменить временно. Необходимо изменить способ хранения данных в исходном наборе данных.

Наборы переменных

Некоторые файлы данных поддерживают специальные переменные , которые называются **наборами множественных ответов**. Наборы множественных ответов не являются "переменными" в обычном смысле этого слова. В наборах множественных ответов для ввода ответов на вопросы, на которые можно дать больше одного ответа, используются несколько переменных. Наборы множественных ответов обрабатываются как категориальные переменные, и большинство операций, которые можно выполнять с категориальными переменными, можно делать также и с наборами множественных ответов.

Набором множественных ответов может быть дихотомический набор множественных ответов или категориальный набор множественных ответов.

Дихотомические наборы множественных ответов. Дихотомический набор множественных ответов обычно состоит из нескольких дихотомических полей, то есть полей только для двух возможных значений, таких как да/нет, присутствует/отсутствует, включено/выключено. Хотя переменные необязательно должны быть дихотомическими, все они кодируются одинаковым образом.

К примеру, в ходе исследования предлагается пять вариантов ответа на вопрос: Из каких источников вы получаете новости?. Каждый респондент может отметить несколько вариантов ответа. Пять возможных вариантов ответа представляются в файле данных в виде пяти переменных, имеющих коды 0 для ответа *Нет* (вариант не выбран) и 1 для ответа *Да* (вариант выбран).

Категориальные наборы множественных ответов. Категориальный набор множественных ответов состоит из нескольких переменных, которые кодируются одинаковым образом и часто имеют много возможных вариантов ответа. К примеру, респондентам может предлагаться вопрос: "Назовите не более трех национальностей, которые наилучшим образом описывают вашу этническую принадлежность". Естественно, что в ответах могут присутствовать сотни различных вариантов, но для кодировки используются только 40 наиболее часто встречающихся вариантов, а остальные варианты относятся к категории другое. В файле данных трем возможным вариантам ответов соответствуют три переменные, каждая из которых имеет 41 категорию (40 кодируемых национальностей и один вариант "другое").

Вкладка Подробности Панели выбора диаграмм

Вкладку Детальная следует использовать, когда известно, какой тип визуализации требуется создать, или когда требуется добавить в визуализацию дополнительную эстетику, панели и/или анимацию. Примеры смотрите в разделе “Примеры Панели выбора диаграмм (Graphboard)” на стр. 198.

1. Если тип визуализации был выбран на вкладке Базовая, он будет показан. В противном случае выберите тип в выпадающем списке. Информацию о типах визуализации смотрите в разделе “Доступные встроенные типы визуализации Панели выбора диаграмм” на стр. 190.
2. Непосредственно справа от эскиза визуализации находятся элементы управления, с помощью которых можно задавать переменные/поля, необходимые для данного типа визуализации. Необходимо указать все эти поля/переменные.
3. Для некоторых визуализаций можно выбрать итожащую статистику. В некоторых случаях (например, при выборе столбчатой диаграммы) можно использовать один из вариантов подытоживания для задания прозрачности. Описание сводной статистики смотрите в разделе “Вкладка Основные Панели выбора диаграмм” на стр. 184.
4. Можно выбрать одну или несколько дополнительных эстетик. Эти эстетики могут увеличивать количество показанных измерений, позволяя добавлять в визуализацию дополнительные поля/переменные. Например, можно использовать поле/переменную для изменения размеров точек в диаграмме рассеяния. Дополнительную информацию о возможной эстетике смотрите в разделе “Эстетики, наложения, панели и анимация” на стр. 180. Имейте в виду, что эстетику прозрачности невозможно настроить с помощью сценариев.
5. Если создается визуализация карты, то группа **Файлы карт** показывает файл или файлы карт, которые будут использованы. Если имеется файл карты, заданный по умолчанию, то выводится этот файл. Чтобы изменить файл карты, щелкните по **Выбрать файл карты** для вывода диалогового окна Выбрать карты. В этом диалоговом окне можно также задать файл карты, используемый по умолчанию. Дополнительную информацию смотрите в разделе “Выбор файлов карт для визуализации карт”.
6. Можно выбрать одну или несколько функций управления панелями или анимацией. Подробную информацию об управлении панелями и анимацией можно найти в разделе “Эстетики, наложения, панели и анимация” на стр. 180.

Выбор файлов карт для визуализации карт

Если выбирается шаблон для визуализации карты, то потребуется файл карты, который предоставляет географическую информацию, необходимую для изображения карты. Если имеется файл карты, заданный по умолчанию, то он будет использоваться для визуализации карты. Чтобы выбрать другой файл карты, щелкните по **Выбрать файл карты** на вкладке Детальная для вывода диалогового окна Выбрать карты.

Диалоговое окно Выбрать карты позволяет выбрать файл основной карты и файл опорной карты. Файлы карт предоставляют географическую информацию, необходимую для изображения карты. Ваше программное приложение устанавливается с набором стандартных файлов карт. Если имеются другие шейп-файлы ESRI, которые вы хотите использовать, то сначала нужно конвертировать эти шейп-файлы в файлы SMZ. Дополнительную информацию смотрите в разделе “Конвертирование и распространение шейп-файлов карт.” на стр. 210. После преобразования карты нажмите кнопку **Управление...** в диалоговом окне Выбор шаблона, чтобы импортировать карту в систему управления, где она будет доступна в диалоговом окне Выбрать карты.

Ниже представлены некоторые моменты, которые необходимо учесть при задании файлов карт:

- Для всех шаблонов карт необходим, по крайней мере, один файл карты.
- Файл карты, как правило, связывает атрибут ключа карты с ключом данных.
- Если для шаблона не требуется ключ карты, который связывает с ключом данных, то для него требуется файл опорной карты и поля, задающие координаты (такие как долгота и широта), для изображения элементов на опорной карте.
- Для шаблонов перекрытия карт требуется две карты: файл первичной карты и файл опорной карты. Опорная карта изображается первой, так чтобы она была позади файла основной карты.

Информацию о терминологии карт, их атрибутах и возможностях смотрите в разделе “Ключевые понятия для карт” на стр. 211.

Файл карты. Можно выбрать любой файл карты, имеющийся в системе управления. Эти файлы включают предустановленные файлы карт и файлы карт, которые были импортированы. Более подробную информацию об управлении файлами карт смотрите в разделе “Управление шаблонами, таблицами стилей и файлами карт” на стр. 209.

Ключ карты. Задайте атрибут, который нужно использовать в качестве ключа, связывающего файл карты и ключ данных.

Сохранить этот файл карты и установки в качестве используемых по умолчанию. Поставьте этот переключатель, если вы хотите, чтобы выбранный файл карты использовался по умолчанию. Если имеется файл карты, заданный по умолчания, то нет необходимости задавать файл карты каждый раз при создании визуализации карты.

Ключ данных. Этот элемент управления перечисляет те же значения, которые появляются на вкладке Детальная диалогового окна Панель выбора диаграмм. Это делается здесь для удобства, если возникла необходимость изменить ключ из-за выбора конкретного файла карты.

При визуализации вывести все элементы карты. Если стоит этот переключатель, то при визуализации изображаются все элементы карты, даже если отсутствует соответствующее значение ключа данных. Выключите этот переключатель, если хотите видеть только те возможности, для которых у вас есть данные. Элементы, идентифицируемые ключами карты, показанными в списке **Не сопоставленные ключи карты**, не будут изображаться при визуализации.

Сравнить значения карты и данных. Для создания визуализации карты ключ карты и ключ данных связываются друг с другом. Ключ карты и ключ данных должны быть извлечены из того же самого домена (например, страны и регионы). Щелкните по **Сравнить**, чтобы проверить, есть ли соответствие между значениями ключа данных и ключа карты. Выведенный значок информирует о состоянии сравнения. Эти значки описаны ниже. Если сравнение выполнено и имеются значения ключей данных, которым не соответствуют значения ключей карты, то такие значения ключей данных появятся в списке **Не сопоставленные ключи**. В списке **Не сопоставленные ключи карты** можно также увидеть, какие значения ключей карты не имеют соответствующих им значений ключей данных. Если выключен переключатель **Выводить все свойства карт при визуализации**, возможности, определенные этими значениями ключей карт, не будут обрабатываться.

Таблица 33. Значки сравнения.

Значок	Описание
	Сравнение не выполнено. Это состояние по умолчанию, перед тем как будет нажата кнопка **Сравнить**. Продолжая, следует проявлять осторожность, так как неизвестно, соответствуют ли друг другу значения ключа данных и ключа карты.
	Сравнение выполнено, и значения ключа данных и ключа карты полностью соответствуют. Для каждого значения ключа данных имеется соответствующий ему элемент, идентифицируемый ключом карты.
	Сравнение выполнено, и некоторые значения ключа данных и ключа карты не соответствуют друг другу. Для некоторых значений ключа данных отсутствуют соответствующие им элементы, идентифицируемые ключом карты. Продолжая, следует проявлять осторожность. Если продолжить, то визуализация карты не будет включать все значения данных.

Таблица 33. Значки сравнения (продолжение).

Значок	Описание
	Сравнение выполнено, и нет соответствующих друг другу значений ключа данных и ключа карты. Следует выбрать другой ключ данных или другой ключ карты, так как, если продолжить, то карта изображена не будет.

Доступные встроенные типы визуализации Панели выбора диаграмм

Вы можете создать несколько разных типов визуализаций. Все приведенные ниже встроенные типы диаграмм доступны на вкладках Базовая и Детальная. Некоторые из описаний шаблонов (особенно шаблонов карт) идентифицируют поля (переменные), заданные на вкладке Подробности с использованием **специального текста**.

Таблица 34. Доступные типы диаграмм.

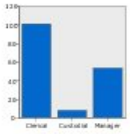
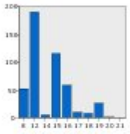
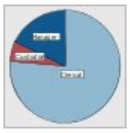
Значок диаграммы	Описание	Значок диаграммы	Описание
	**Столбцы** Рассчитывает значения итожащей статистики для количественного поля/переменной и показывает результаты в виде столбцов для каждой категории категориального поля/переменной. *Требуется*: Категориальное поле и количественное поле.		**Столбцы частот** Служит для просмотра доли наблюдений/строк в каждой категории категориального поля/переменной в виде столбцов. Эту диаграмму также можно создать с помощью узла Distribution. Этот узел обеспечивает несколько дополнительных возможностей. Дополнительную информацию смотрите в разделе “Узел Распределение” на стр. 230. *Требуется*: Одно категориальное поле.
	**Круг** Служит для вычисления суммы количественного поля/переменной и просмотра доли этой суммы в каждой категории категориального поля/переменной в виде сектора круга. *Требуется*: Категориальное поле и количественное поле.		**Круги частот** Служит для просмотра доли наблюдений/строк в каждой категории категориального поля/переменной в виде секторов круга. *Требуется*: Одно категориальное поле.

Таблица 34. Доступные типы диаграмм (продолжение).

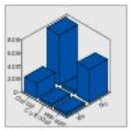
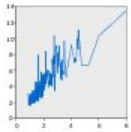
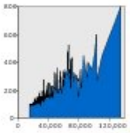
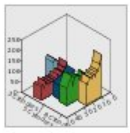
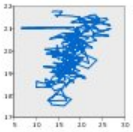
Значок диаграммы	Описание	Значок диаграммы	Описание
	**Трехмерные столбцы** Служит для расчета значений итожащей статистики для количественного поля/переменной и просмотра результатов для всех сочетаний значений двух категориальных полей/переменных. *Требуется*: Два категориальных поля и количественное поле.		**Трехмерные круги** Это то же самое, что и Круг, только с дополнительным трехмерным эффектом. *Требуется*: Категориальное поле и количественное поле.
	**Line** Рассчитывает значения итожащей статистики для поля/переменной для каждого значения другого поля/переменной и показывает результаты в виде линии, соединяющей значения. Диаграмму с линиями можно построить с помощью узла Plot. Этот узел обеспечивает несколько дополнительных возможностей. Дополнительную информацию смотрите в разделе “Узел График” на стр. 217. *Требуется*: Два поля любого типа.		**Область** Рассчитывает значения итожащей статистики для поля/переменной для каждого значения другого поля/переменной и показывает результаты в виде области, соединяющей значения. Разница между линией и областью минимальна и состоит в том, что область - это линия с закрашенным пространством под ней. Однако, если используется цвета, линия разделяется, а область состыковывается. *Требуется*: Два поля любого типа.
	**Трехмерные области** Выводит значения одного поля против значений другого поля, разбитых по категориальному полю/переменной. Для каждой категории строится элемент области. *Требуется*: Категориальное поле и два поля любого типа.		**Путь** Служит для просмотра значений одного поля/переменной, выведенных по значениям другого поля/переменной, с линией, соединяющей значения в том порядке, в котором они содержатся в исходном наборе данных. Порядок является основным различием между путем и линией. *Требуется*: Два поля любого типа.

Таблица 34. Доступные типы диаграмм (продолжение).

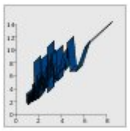
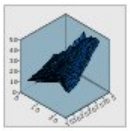
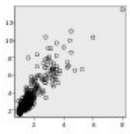
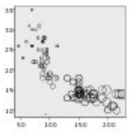
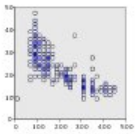
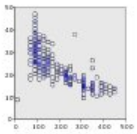
Значок диаграммы	Описание	Значок диаграммы	Описание
	**Лента** Рассчитывает значения итожащей статистики для поля/переменной для каждого значения другого поля/переменной и показывает результаты в виде ленты, соединяющей значения. Лента, по сути, является линией с трехмерным эффектом. Это не настоящая трехмерная диаграмма. *Требуется*: Два поля любого типа.		**Поверхность** Служит для просмотра значений трех полей/переменных и поверхности, соединяющей значения. *Требуется*: Три поля любого типа.
	**Диаграмма рассеяния** Служит для просмотра значений двух полей/переменных. Эта диаграмма позволяет выявить связь между полями/переменными (если таковая существует). Диаграмму рассеяния также можно создать с помощью узла Plot. Этот узел обеспечивает несколько дополнительных возможностей. Дополнительную информацию смотрите в разделе “Узел График” на стр. 217. *Требуется*: Два поля любого типа.		**Диаграмма с пузырями** Как и обычная диаграмме рассеяния, показывает значения двух полей/переменных. Разница заключается в том, что значения третьего поля/переменной задают размер каждой точки. *Требуется*: Три поля любого типа.
	**Диаграмма рассеяния с группировкой** Как и обычная диаграмме рассеяния, показывает значения двух полей/переменных. Разница заключается в том, что близкие значения делятся на группы, и эстетика цвета или размера используется для обозначения количества наблюдений в каждой группе. *Требуется*: Два количественных поля.		**Диаграмма рассеяния с группировкой шестиугольниками** Смотрите описание Диаграмм рассеяния с интервалами. Разница состоит в том, что группы имеют форму шестиугольников, а не кругов. Диаграмма рассеяния с группировкой шестиугольниками похожа на диаграмму рассеяния с группировкой. Однако количество значений в каждом интервале в этих диаграммах будет разным из-за формы групп. *Требуется*: Два количественных поля.

Таблица 34. Доступные типы диаграмм (продолжение).

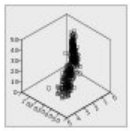
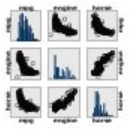
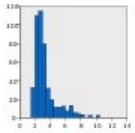
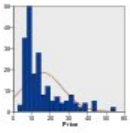
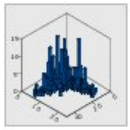
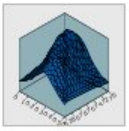
Значок диаграммы	Описание	Значок диаграммы	Описание
	**Трехмерная диаграмма рассеяния** Служит для просмотра значений трех полей/переменных. Эта диаграмма позволяет выявить связь между полями/переменными (если таковая существует). Трехмерную диаграмму рассеяния также можно создать с помощью узла Plot. Этот узел обеспечивает несколько дополнительных возможностей. Дополнительную информацию смотрите в разделе “Узел График” на стр. 217. *Требуется*: Три поля любого типа.		**Матрица диаграмм рассеяния** Выводит значения одного поля против значений другого поля для каждого из полей. Матрицу диаграмм рассеяния можно назвать таблицей диаграмм рассеяния. SPLOM также включает гистограмму каждого поля/переменной. *Требуется*: Не менее двух количественных полей.
	**Гистограмма** Служит для просмотра частотного распределения поля/переменной. С помощью гистограммы можно понять тип распределения и выявить асимметричное распределение. Эту диаграмму также можно создать с помощью узла Histogram. Этот узел обеспечивает несколько дополнительных возможностей. Дополнительную информацию смотрите в разделе “Вкладка График гистограммы” на стр. 234. *Требуется*: Одно поле любого типа.		**Гистограмма с нормальной кривой** Служит для просмотра частотного распределения количественного поля/переменной с наложенной кривой нормального распределения. *Требуется*: Одно количественное поле.
	**Трехмерная гистограмма** Служит для просмотра частотного распределения пары количественных полей/переменных. *Требуется*: Два количественных поля.		**Трехмерная диаграмма плотности** Служит для просмотра частотного распределения пары количественных полей/переменных. Похожа на трехмерную гистограмму, только вместо столбцов для вывода распределения используется поверхность. *Требуется*: Два количественных поля.

Таблица 34. Доступные типы диаграмм (продолжение).

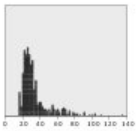
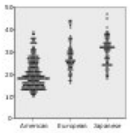
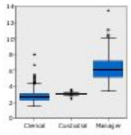
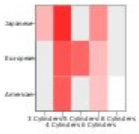
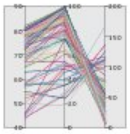
Значок диаграммы	Описание	Значок диаграммы	Описание
	**Точечный график** Показывает отдельных наблюдения/строки и состыковывает их в точках на оси *x*. Эта диаграмма похожа на гистограмму тем, что на ней показывается распределение данных, однако она показывает каждое наблюдение/строку вместо агрегированной суммы для определенной группы (диапазона) значений. *Требуется*: Одно поле любого типа.		**Двумерные точки** Служит для просмотра отдельных наблюдений/строк и состыковывания их в точках на оси *y* для каждой категории категориального поля/переменной. *Требуется*: Категориальное поле и количественное поле.
	**Коробчатая диаграмма** Рассчитывает пять статистик (минимум, первый квартиль, медиану, третий квартиль и максимум) для количественного поля/переменной для каждой категории категориального поля/переменной. Результаты показываются в виде элементов ящичной диаграммы с усами/схемы. С помощью ящичных диаграмм с усами можно понять, как различается распределение количественных данных для разных категорий. *Требуется*: Категориальное поле и количественное поле.		**Карта интенсивности** Рассчитывает среднее значение для количественной переменной для всех сочетаний категорий двух категориальных полей. *Требуется*: Два категориальных поля и количественное поле.
	**Параллельная** Создает параллельные оси для всех полей и рисует линию через значения поля для каждой строки/наблюдения в данных. *Требуется*: Не менее двух количественных полей.		**Хороплет частот** Вычисляет количество для каждой категории категориального поля (**Ключ данных**) и изображает карту, в которой используется насыщение цветом, чтобы представить количества на элементах карты, которые соответствуют категориям. *Требуется*: Одно категориальное поле. Файл карты, ключ которой соответствует категориям поля **Ключ данных**.

Таблица 34. Доступные типы диаграмм (продолжение).

Значок диаграммы	Описание	Значок диаграммы	Описание
	**Хороплет средних/медиан/сумм** Вычисляет среднее, медиану или сумму значений количественного поля (**Цвет**) для каждой категории категориального поля (**Ключ данных**) и изображает карту, в которой используется насыщение цветом, чтобы представить вычисленные статистики на элементах карты, которые соответствуют категориям. *Требуется*: Категориальное поле и количественное поле. Файл карты, ключ которой соответствует категориям поля **Ключ данных**.		**Хороплет значений** Изображает карту, в которой используется цвет, чтобы представить значения категориального поля (**Цвет**) для элементов карты, которые соответствуют значениям, заданным другим категориальным полем (**Data Key**). Если имеются несколько категориальных значений поля Цвет для каждого элемента, то используется модальное значение. *Требуется*: Два количественных поля. Файл карты, ключ которой соответствует категориям поля **Ключ данных**.
	**Координаты на хороплете частот** Это похоже на Хороплет количеств, за исключением того, что имеются два дополнительных количественных поля (**Долгота** и **Широта**), которые задают координаты для изображения точек на карте хороплета. *Требуется*: Категориальное поле и два количественных поля. Файл карты, ключ которой соответствует категориям поля **Ключ данных**.		**Координаты на хороплете средних/медиан/сумм** Это похоже на Хороплет средних/медиан/сумм, за исключением того, что имеются два дополнительных количественных поля (**Долгота** и **Широта**), которые задают координаты для изображения точек на карте хороплета. *Требуется*: Категориальное поле и три количественных поля. Файл карты, ключ которой соответствует категориям поля **Ключ данных**.
	**Координаты на хороплете значений** Это похоже на Хороплет значений, за исключением того, что имеются два дополнительных количественных поля (**Долгота** и **Широта**), которые задают координаты для изображения точек на карте хороплета. *Требуется*: Два категориальных поля и два количественных поля. Файл карты, ключ которой соответствует категориям поля **Ключ данных**.		**Столбцы частот на карте** Вычисляет долю строк/наблюдений в каждой категории категориального поля (**Категории**) для каждого элемента карты (**Ключ данных**), а также изображает карту и столбчатые диаграммы в центре каждого элемента карты. *Требуется*: Два категориальных поля. Файл карты, ключ которой соответствует категориям поля **Ключ данных**.

Таблица 34. Доступные типы диаграмм (продолжение).

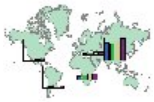
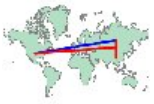
Значок диаграммы	Описание	Значок диаграммы	Описание
	**Столбцы на карте** Вычисляет итожащую статистику для количественного поля (**Значения**) и выводит результаты для каждой категории категориального поля (**Категории**) для каждого элемента карты **Ключ данных**) в виде столбчатых диаграмм, расположенных в центре каждого элемента карты. *Требуется*: Два категориальных поля и количественное поле. Файл карты, ключ которой соответствует категориям поля **Ключ данных**.		**Круговая диаграмма частот на карте** Выводит долю строк/наблюдений в каждой категории категориального поля (**Категории**) для каждого элемента карты (**Ключ данных**), а также изображает карту и доли в виде секторов круговой диаграммы в центре каждого элемента карты. *Требуется*: Два категориальных поля. Файл карты, ключ которой соответствует категориям поля **Ключ данных**.
	**Круговая диаграмма на карте** Вычисляет сумму значений количественного поля (**Значения**) в каждой категории категориального поля (**Категории**) для каждого элемента карты (**Ключ данных**), а также изображает карту и суммы в виде секторов круговой диаграммы в центре каждого элемента карты. *Требуется*: Два категориальных поля и количественное поле. Файл карты, ключ которой соответствует категориям поля **Ключ данных**.		**Диаграмма с линиями на карте** Вычисляет итожащую статистику для количественного поля (**Y**) для каждого значения другого поля (**X**) для каждого элемента карты (**Ключ данных**), а также изображает карту и линейные диаграммы, соединяющими значения, в центре каждого элемента карты. *Требуется*: Категориальное поле и два поля любого типа. Файл карты, ключ которой соответствует категориям поля **Ключ данных**.
	**Координаты на опорной карте** Изображает карту и точки, используя количественные поля (**Долгота** и **Широта**), которые задают координаты точек. *Требуется*: Два поля диапазона. Файл карты.		**Стрелки на опорной карте** Изображает карту и стрелки, используя количественные поля, задающие начальные точки (**Нач. долг.** и **Нач. шир.**), а также конечные точки (**Кон. долг.** и **Кон. шир.**) для каждой стрелки. Для каждой записи/наблюдения в данных создается стрелка на карте. *Требуется*: Четыре количественных поля. Файл карты.

Таблица 34. Доступные типы диаграмм (продолжение).

Значок диаграммы	Описание	Значок диаграммы	Описание
	**Карта с наложением точек** Изображает опорную карту и располагает поверх нее другую точечную карту, с точечными элементами, раскрашенными с помощью категориального поля (**Цвет**). *Требуется*: Два категориальных поля. Файл точечной карты, ключ которой соответствует категориям поля **Ключ данных**. Файл опорной карты.		**Карта с наложением полигонов** Изображает опорную карту и располагает поверх нее другую карту с полигонами, с элементами полигонов, раскрашенными с помощью категориального поля (**Цвет**). *Требуется*: Два категориальных поля. Файл карты с полигонами, ключ которой соответствует категориям поля **Ключ данных**. Файл опорной карты.
	**Карта с наложением линий** Изображает опорную карту и располагает поверх нее другую карту с линиями, с элементами линий, раскрашенными с помощью категориального поля (**Цвет**). *Требуется*: Два категориальных поля. Файл карты с линиями, ключ которой соответствует категориям поля **Ключ данных**. Файл опорной карты.		

Создание визуализаций карт

Для многих визуализаций нужно сделать только два выбора: нужные поля (переменные) и шаблон для визуализации этих полей. Кроме этого ничего выбирать или выполнять какие-либо действия не нужно. Для визуализаций карт требуется по крайней мере еще одно дополнительное действие: выбрать файл карт, определяющий географическую информацию для визуализации карты.

Основными шагами создания простой карты являются:

1. Выберите на вкладке Базовая представляющие интерес поля. Информацию о типе и количестве полей, нужных для визуализации различных карт, смотрите в разделе “Доступные встроенные типы визуализации Панели выбора диаграмм” на стр. 190.
2. Выберите шаблон карты.
3. Перейдите на вкладку Подробности.
4. Проверьте, что в **Ключ данных** и других требуемых раскрывающихся списках правильно выбраны поля.
5. В группе Файлы карт щелкните по **Выбрать файл карты**.
6. Используйте диалоговое окно Выбрать карты, чтобы выбрать файл карты и ключ карты. Значения ключа карты должны соответствовать значениям поля, которое задает **Ключ данных**. Чтобы сравнить эти значения, можно воспользоваться кнопкой **Сравнить**. Если выбирается шаблон карты с наложением, то также необходимо выбрать опорную карту. Опорная карта ссылок не привязана по ключу к данным. Она используется в качестве фона для главной карты. Дополнительную информацию о диалоговом окне Выбрать карты смотрите в разделе “Выбор файлов карт для визуализации карт” на стр. 188.
7. Щелкните **OK**, чтобы закрыть диалоговое окно Выбрать карты.

8. В диалоговом окне Панель выбора диаграмм щелкните по **Выполнить**, чтобы создать визуализацию карты.

Примеры Панели выбора диаграмм (Graphboard)

В этом разделе приводятся несколько примеров, демонстрирующих доступные возможности. Примеры также позволяют понять, как нужно интерпретировать получающиеся визуализации.

В этих примерах используется стрим *graphboard.str*, в котором содержатся ссылки на файлы данных *employee_data.sav* и *customer_subset.sav* и *worldsales.sav*. Эти файлы находятся в папке *Demos*, внутри папки, где установлен IBM SPSS Modeler Client. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *graphboard.str* находится в папке *streams*.

Знакомится с примерами рекомендуется в том порядке, в котором они представлены. Последующие примеры основаны на предыдущих.

Пример: Столбчатая диаграмма с итожащей статистикой

Мы построим столбчатую диаграмму, которая подытоживает непрерывное числовое поле/переменную для каждой категории набора/категориальной переменной. В частности, мы построим столбчатую диаграмму, на которой будет показана средняя зарплата для мужчин и женщин.

В этом и нескольких следующих примерах используется файл *Employee data*, который является гипотетическим набором данных, содержащим информацию о сотрудниках компании.

1. Добавьте узел источников Statistics File, который указывает на *employee_data.sav*.
2. Добавьте узел Graphboard (Панель выбора диаграмм) и откройте его для редактирования.
3. На вкладке Основные выберите *Пол* и *Текущий заработок* . (Используйте клавишу Ctrl, чтобы выделить несколько полей/переменных).
4. Выберите **Bar** (столбцы).
5. В раскрывающемся списке Summary выберите пункт **Mean** (среднее).
6. Нажмите кнопку **Выполнить**.
7. На экране результатов нажмите кнопку панели инструментов Показать поля и метки значения (вторая в группе двух в центре панели инструментов).

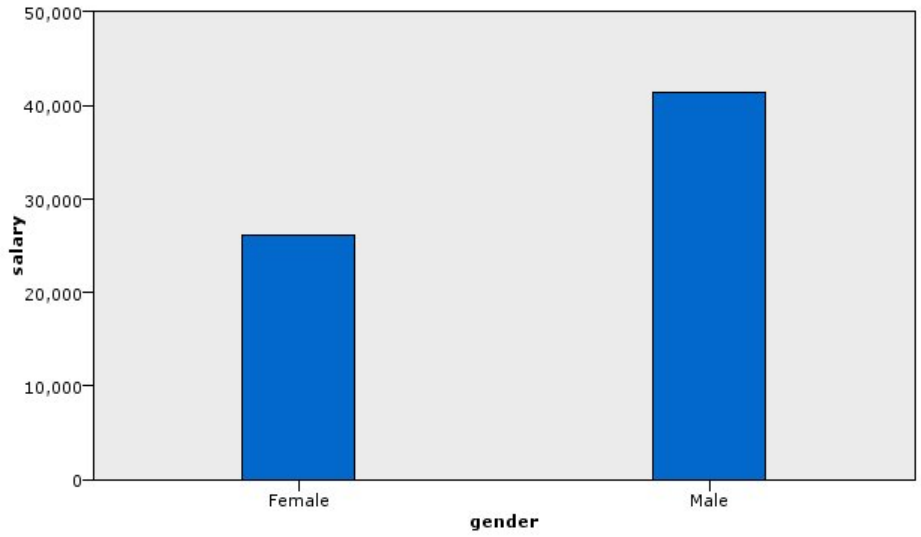


Рисунок 8. Столбчатая диаграмма с итожащей статистикой

Наблюдаются следующие результаты:

- Столбцы говорят нам о том, что средняя зарплата мужчин больше средней зарплаты женщин.

Пример: Составная столбчатая диаграмма со сводной статистикой

Теперь мы построим составную столбчатую диаграмму, чтобы понять, зависит ли разница средней зарплаты мужчин и женщин от типа работы. Возможно, для некоторых типов работы, женщины в среднем зарабатывают больше мужчин.

Примечание: В этом примере используются *Данные о наемных работниках*.

1. Добавьте узел Graphboard (Панель выбора диаграмм) и откройте его для редактирования.
2. На вкладке Основные выберите *Категория занятости* и *Текущий заработок*. (Используйте клавишу Ctrl, чтобы выделить несколько полей/переменных).
3. Выберите **Bar** (столбцы).
4. В списке Сводка выберите пункт **Mean** (среднее).
5. Перейдите на вкладку Подробности. Обратите внимание на то, что здесь дублируются настройки, выполненные на предыдущей вкладке.
6. В группе Optional Aesthetics в раскрывающемся списке Color (цвет) выберите *gender* (пол).
7. Нажмите кнопку **Выполнить**.

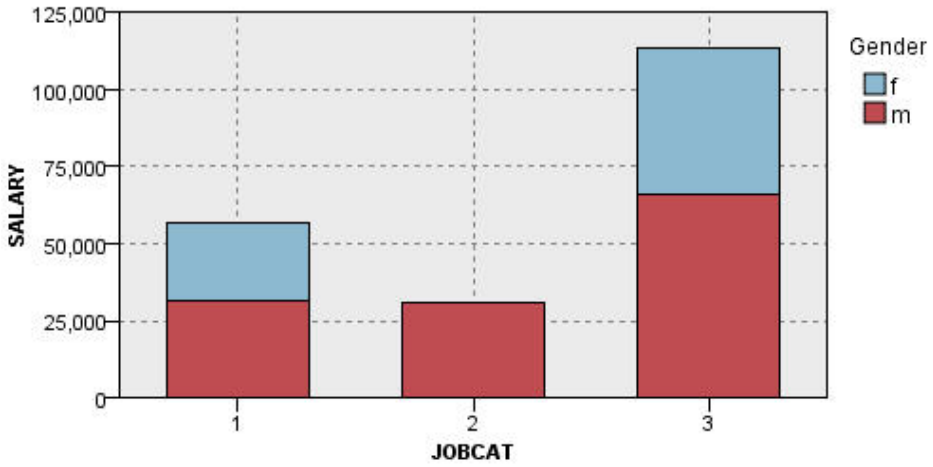


Рисунок 9. Составная столбчатая диаграмма

Наблюдаются следующие результаты:

- Разница средней зарплаты для каждого типа работы не такая большая, как в столбчатой диаграмме, в которой сравнивалась средняя зарплата всех мужчин и женщин. Возможно, в каждой группе разное количество мужчин и женщин. Это можно проверить, создав столбчатую диаграмму количеств.
- Вне зависимости от типа работы средняя зарплата мужчин всегда больше средней зарплаты женщин.

Пример: панельная гистограмма

Мы построим гистограмму с разделением на панели по полу, чтобы сравнить частотное распределение зарплаты для мужчин и женщин. Частотное распределение показывает, как много наблюдений/строк лежат в определенных диапазонах зарплат. С помощью гистограммы с панелями мы можем понять, как дальше анализировать разницу между зарплатами разных полов.

Примечание: В этом примере используются *Данные о наемных работниках*.

1. Добавьте узел Graphboard (Панель выбора диаграмм) и откройте его для редактирования.
2. На вкладке Основные выберите *Текущий заработок* .
3. Выберите **Histogram** (гистограмма).

4. Перейдите на вкладку Подробности.
5. В группе Panels and Animation, в раскрывающемся списке Panel Across выберите *gender* (пол).
6. Щелкните по кнопке **Выполнить**.

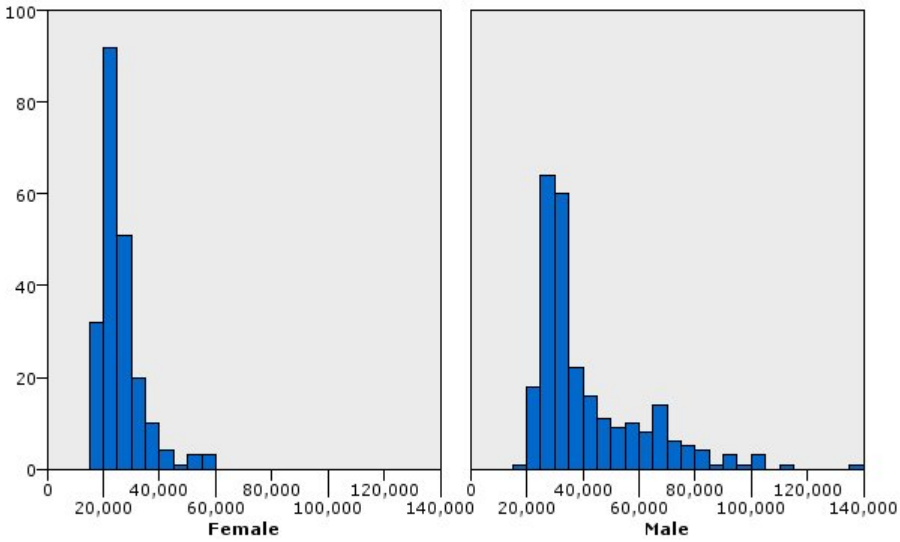


Рисунок 10. Гистограмма с панелями

Наблюдаются следующие результаты:

- Ни одно частотное распределение не является нормальным. Другими словами, гистограммы не похожи на колоколообразные кривые, на которые они были бы похожи, если бы дынные были бы распределены нормально.
- Более высокие столбцы находятся в левой части каждой диаграммы. Поэтому как среди мужчин, так и среди женщин, тех, у кого зарплаты невелики, больше, чем тех, у кого высокие зарплаты.
- Частотные распределения зарплаты среди мужчин и женщин неодинаковы. Обратите внимание на форму гистограмм. Мужчин с высокими зарплатами больше, чем женщин с высокими зарплатами.

Пример: панельная диаграмма с точками

Как и на гистограмме, на диаграмме с точками показывается распределение непрерывного числового диапазона. В отличие от гистограммы, которая отражает количества для диапазонов данных, разбитых на интервалы, диаграмма с точками показывает каждую строку/наблюдение в данных. Поэтому диаграмма с точками обеспечивает большую степень детализации по сравнению с гистограммой. Диаграмма с точками может быть хорошим началом анализа распределений частот.

Примечание: В этом примере используются *Данные о наемных работниках*.

1. Добавьте узел Graphboard (Панель выбора диаграмм) и откройте его для редактирования.
2. На вкладке Основные выберите *Текущий заработок* .
3. Выберите **Dot Plot** (Диаграмма с точками).
4. Перейдите на вкладку Подробности.
5. В группе Panels and Animation, в раскрывающемся списке Panel Across выберите *gender* (пол).
6. Щелкните по кнопке **Выполнить**.
7. Чтобы лучше видеть диаграмму разверните окно вывода.

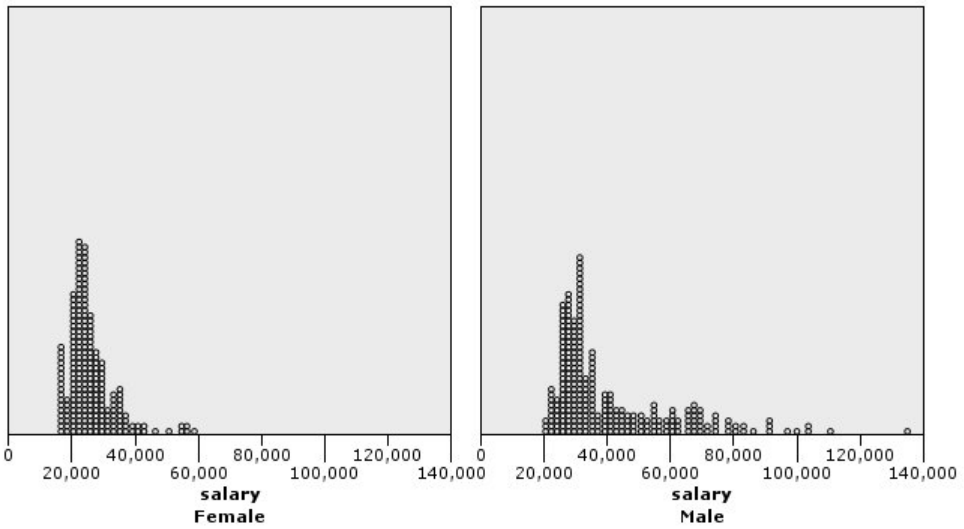


Рисунок 11. Панельная диаграмма с точками

По сравнению с гистограммой (смотрите раздел “Пример: панельная гистограмма” на стр. 199), можно наблюдать следующее:

- Пиковое значение 20 000, которое показано на гистограмме для женщин, существенно ниже на диаграмме с точками. Имеется множество наблюдений/строк, сконцентрированных вокруг этого значения, однако большинство значений находятся ближе к 25 000. Этот уровень детализации не виден на гистограмме.
- Хотя на гистограмме для мужчин видно, что средняя зарплата мужчин постепенно снижается после 40 000, диаграмма с точками говорит о том, что распределение довольно равномерное от этого значения до 80 000. Для любого значения зарплаты в этом диапазоне существует трое или больше мужчин, которые получают такую зарплату.

Пример: ящичная диаграмма с усами

Ящичная диаграмма с усами является еще одним полезным средством визуализации распределения данных. На ящичной диаграмме с усами показываются несколько статистических мер, которые мы рассмотрим после построения диаграммы.

Примечание: В этом примере используются *Данные о наемных работниках*.

1. Добавьте узел Graphboard (Панель выбора диаграмм) и откройте его для редактирования.
2. На вкладке Основные выберите *Пол* и *Текущий заработок* . (Используйте клавишу Ctrl, чтобы выделить несколько полей/переменных).
3. Выберите **Boxplot** (Ящики).
4. Щелкните по кнопке **Выполнить**.

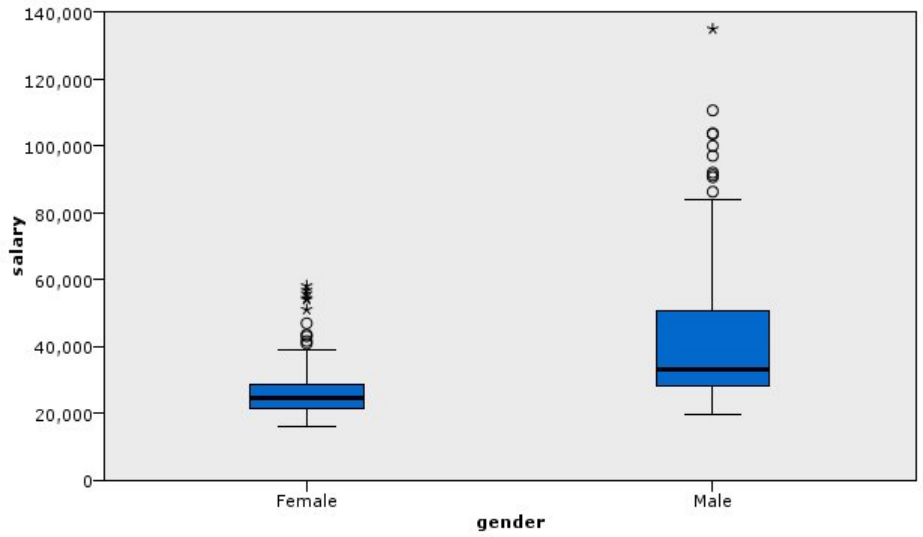


Рисунок 12. Ящики с усами

Ниже приводится описание разных частей ящичной диаграммы с усами:

- Темная линия посередине ящиков - это медиана переменной *salary* (зарплата). Половина наблюдений/строк имеют значение выше этой медианы, а половина - ниже ее. Как и среднее значение, медиана является мерой положения центра распределения. В отличие от среднего значения, наблюдения/строки с экстремальными значениями оказывают на нее меньшее влияние. В этом примере медиана меньше среднего (сравните с разделом “Пример: Столбчатая диаграмма с итожащей статистикой” на стр. 198). Разница между средним значением и медианой указывает на то, что существует несколько наблюдений/строк с экстремальными значениями, которые увеличивают среднее значение. То есть на то, что несколько сотрудников получают большие зарплаты.
- Нижняя граница ящика соответствует 25-й процентили. Двадцать пять процентов наблюдений/строк имеют значения ниже 25-й процентили. Верхняя граница ящика соответствует 75-й процентили. Двадцать пять процентов наблюдений/строк имеют значения выше 75-й процентили. Это значит, что 50% наблюдений/строк лежат в пределах ящика. Ящик значительно короче для женщин, чем для мужчин. Это говорит о том, что зарплаты у женщин различаются между собой меньше, чем у мужчин. Верхнюю и нижнюю границы ящика часто называют **сгибами** .
- Т-образные столбцы, выходящие за пределы ящиков, называются **внутренними ограничителями** или **усами** . Их длина больше высоты ящика в 1,5 раза или, если в этом диапазоне нет ни одного значения наблюдения/строки, их длина будет соответствовать минимальному и максимальному значениям. При нормальном распределении данных в диапазоне "усов" должно лежать примерно 95% данных. В этом примере "усы" для женщин меньше, чем для мужчин, что является еще одним свидетельством того, что зарплаты у женщин различаются между собой меньше, чем у мужчин.
- Точки - это **выбросы**. Выбросы - это значения, которые лежат за пределами усов. Выбросы - это экстремальные значения. Звездочки - это **экстремальные выбросы**. Они представляют наблюдения/строки, которые имеют значения, превышающие высоту ящиков, больше чем в три раза. Имеется несколько выбросов как для женщин, так и для мужчин. Помните, что среднее значение больше медианы. Причиной этого являются выбросы.

Пример: круговая диаграмма

Теперь мы воспользуемся другим набором данных для изучения некоторых других типов визуализации. Набор данных называется *customer_subset*, это гипотетический файл данных, в котором содержится информация о клиентах.

Сначала мы построим круговую диаграмму, чтобы увидеть какая доля клиентов приходится на каждый географический регион.

1. Добавьте узел источников Statistics File, который указывает на *customer_subset.sav*.
2. Добавьте узел Graphboard (Панель выбора диаграмм) и откройте его для редактирования.
3. На вкладке Основные выберите *Географический индикатор* .
4. Выберите **Круговая диаграмма количеств**.
5. Щелкните по кнопке **Выполнить**.

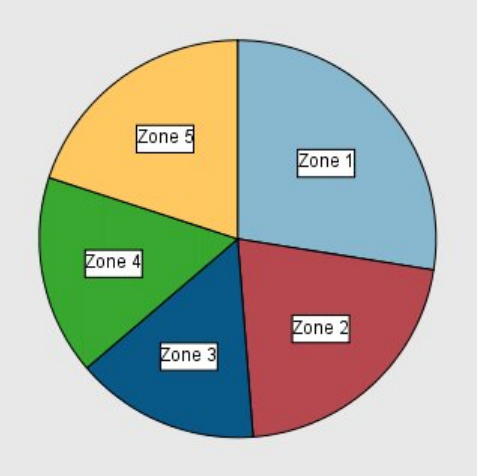


Рисунок 13. Круговая диаграмма

Наблюдаются следующие результаты:

- В регионе Zone 1 больше клиентов, чем в любом другом регионе.
- Клиенты распределены равномерно по другим регионам.

Пример: Тепловая карта

Мы создадим категориальную тепловую карту, чтобы понять различия в средних доходах клиентов между различными регионами и возрастными группами.

Примечание: В этом примере используется файл *customer_subset*.

1. Добавьте узел Graphboard (Панель выбора диаграмм) и откройте его для редактирования.
2. На вкладке Основные выберите *Географический индикатор*, *Возрастная категория* и *Доход домовладения в тысячах* , в таком порядке. (Используйте клавишу Ctrl, чтобы выделить несколько полей/переменных).
3. Выберите **Тепловая карта**.
4. Щелкните по кнопке **Выполнить**.
5. В окне вывода нажмите кнопку панели инструментов Показать поля и метки значения (правая из двух в центре панели инструментов).

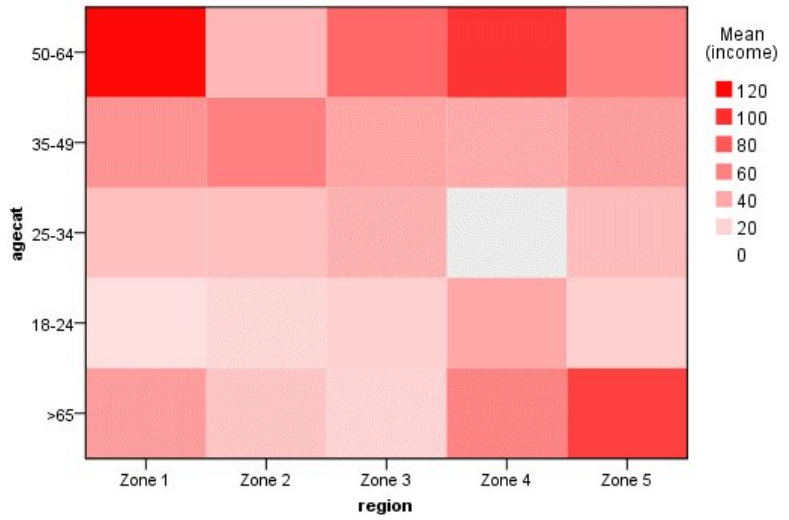


Рисунок 14. Категориальная тепловая карта

Наблюдаются следующие результаты:

- Тепловая карта похожа на таблицу, в которой значения в ячейках представляются при помощи цветов, а не цифр. Яркий, насыщенный красный цвет означает максимальное значение, тогда как серый - минимальное. Значение в каждой ячейки является средним значением непрерывного поля/переменной для каждой пары категорий.
- За исключением региона Zone 2 и региона Zone 5 у клиентов в возрасте от 50 до 64 лет средний доход домохозяйства выше, чем в других возрастных группах.
- В регионе Zone 4 нет клиентов в возрасте от 25 до 34 лет.

Пример: Матрица диаграмм рассеяния (SPLOM)

Мы построим матрицу диаграмм рассеяния для нескольких переменных с целью понять, существуют ли какие-либо взаимосвязи между переменными в наборе данных.

Примечание: В этом примере используется файл *customer_subset*.

1. Добавьте узел Graphboard (Панель выбора диаграмм) и откройте его для редактирования.
2. На вкладке Основные выберите *Возраст в годах*, *Доход домовладения в тысячах* и *Долг по кредитной карте в тысячах* . (Используйте клавишу Ctrl, чтобы выделить несколько полей/переменных).
3. Выберите **SPLOM** .
4. Щелкните по кнопке **Выполнить**.
5. Чтобы лучше видеть матрицу разверните окно вывода.

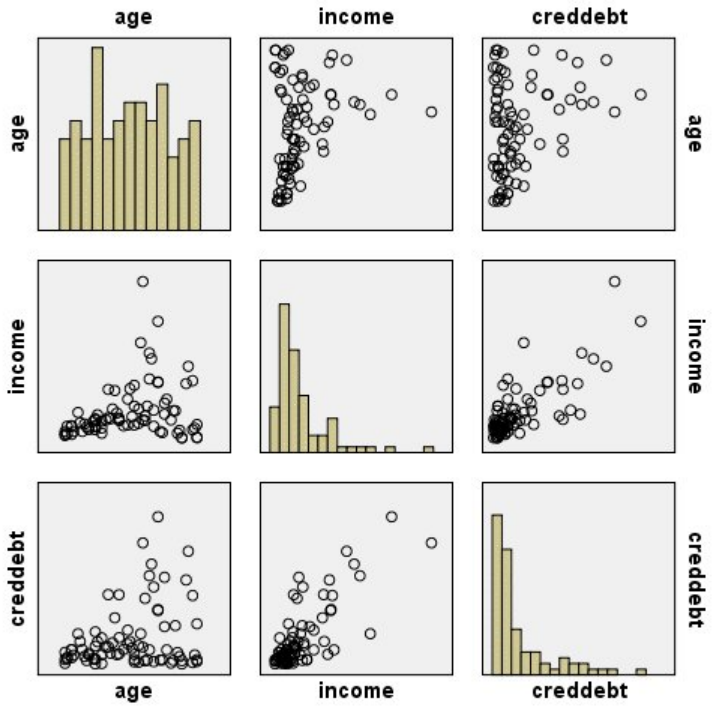


Рисунок 15. Матрица диаграммы рассеяния (SPLOM)

Наблюдаются следующие результаты:

- Гистограммы, расположенные по диагонали, показывают распределение каждой переменной в SPLOM. Гистограмма для *age* показывается в левой верхней ячейке, гистограмма для *income* - в центральной, а гистограмма для *creddebt* - в правой нижней. Ни одна из переменных не имеет нормального распределения. Другими словами, ни одна из гистограмм не напоминает колоколообразную кривую. Также следует обратить внимание на то, что у гистограмм для *income* и *creddebt* асимметричны.
- Между переменной *age* и всеми остальными переменными не наблюдается связи.
- Между *income* и *creddebt* наблюдается линейная зависимость. Другими словами, *creddebt* увеличивается при увеличении *income*. Возможно, стоит построить отдельные диаграммы рассеяния для этих переменных и других связанных с ними переменных для дополнительного изучения связей.

Пример: Хороплет (цветовая карта) сумм

Теперь создадим визуализацию карты. Затем в следующем примере создадим вариант этой визуализации. Набором данных эля этого служит *worldsales*, который представляет собой гипотетический файл данных, содержащий товарооборот по континентам и товарам.

1. Добавьте узел Graphboard (Панель выбора диаграмм) и откройте его для редактирования.
2. На вкладке Основные выберите *Континент* и *Доход*. (Используйте клавишу Ctrl, чтобы выделить несколько полей/переменных).
3. Выберите **Хороплет сумм**.
4. Перейдите на вкладку Подробности.
5. В группе Необязательная эстетика в раскрывающемся списке Метка данных выберите *Континент*.
6. В группе Файлы карт щелкните по **Выбрать файл карты**.
7. В диалоговом окне Выбрать карты для **Карта** выберите *Континенты*, а для **Ключ карты** выберите *КОНТИНЕНТ*.
8. В группе Сравнить значения карты и данных, щелкните по **Сравнить**, чтобы убедиться в том, что ключи карты соответствуют ключам данных. В этом примере все значения ключей данных имеют соответствующие ключи карты и элементы. Можно также видеть, что нет данных для Океании.
9. В диалоговом окне Выбрать карты щелкните по **OK**.

10. Щелкните по кнопке **Выполнить**.

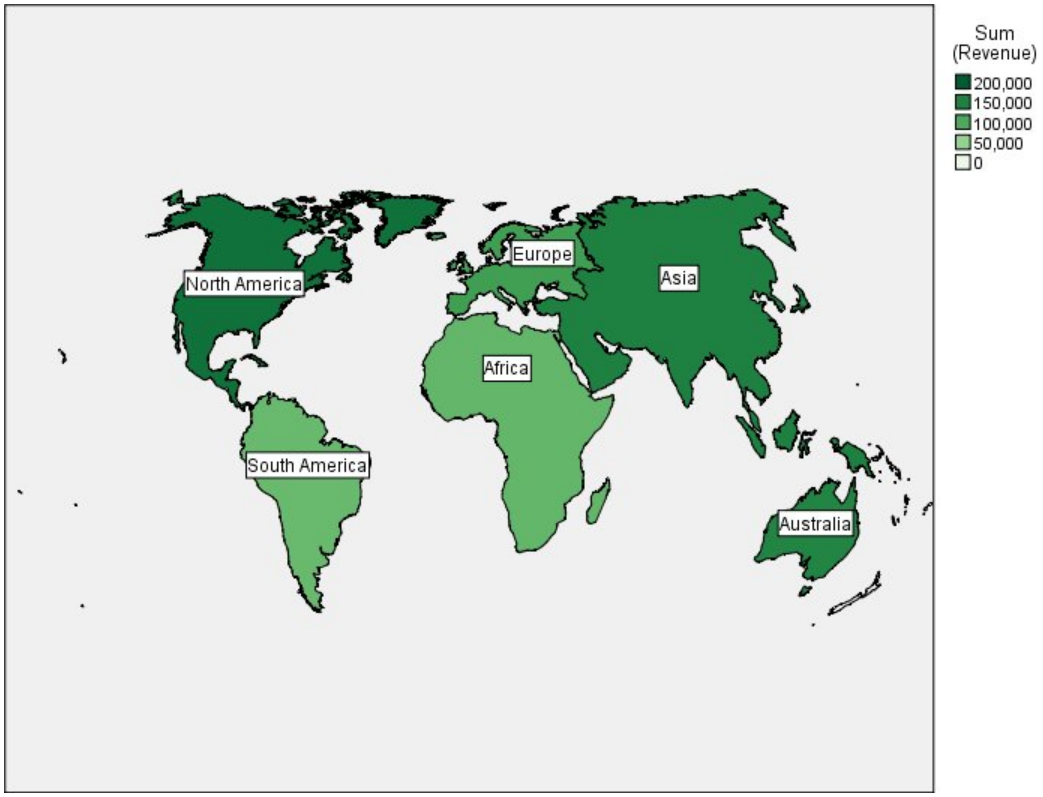


Рисунок 16. Хороплет сумм

Эта визуализация карты позволяет легко увидеть, что доход является максимальным в Северной Америке и минимальным в Южной Америке и Африке. Каждый континент имеет метку, так как было использовано *Континент* в качестве эстетики меток данных.

Пример: Столбчатая диаграмма на карте

Этот пример показывает, как доход разбивается по товарам для каждого континента.

Примечание: В данном примере используется *worldsales*.

1. Добавьте узел Graphboard (Панель выбора диаграмм) и откройте его для редактирования.
2. На вкладке Основные выберите *Континент*, *Товар* и *Доход*. (Используйте клавишу Ctrl, чтобы выделить несколько полей/переменных).
3. Выберите **Столбцы на карте**.
4. Перейдите на вкладку Подробности.

   При использовании более чем одного поля заданного типа важно проверить, что каждому полю правильно назначена позиция.
5. В раскрывающемся списке Категории выберите пункт *Товар*.
6. В раскрывающемся списке Значения выберите пункт *Доход*.
7. В раскрывающемся списке Ключ данных выберите *Континент*
8. В раскрывающемся списке Итог выберите пункт *Сумма*.
9. В группе Файлы карт щелкните по **Выбрать файл карты**.
10. В диалоговом окне Выбрать карты для **Карта** выберите *Континенты*, а для **Ключ карты** выберите *КОНТИНЕНТ*.

11. В группе Сравнить значения карты и данных, щелкните по **Сравнить**, чтобы убедиться в том, что ключи карты соответствуют ключам данных. В этом примере все значения ключей данных имеют соответствующие ключи карты и элементы. Можно также видеть, что нет данных для Океании.
12. В диалоговом окне Выбрать карты щелкните по **OK**.
13. Щелкните по кнопке **Выполнить**.
14. Чтобы лучше видеть дисплей разверните окно вывода.

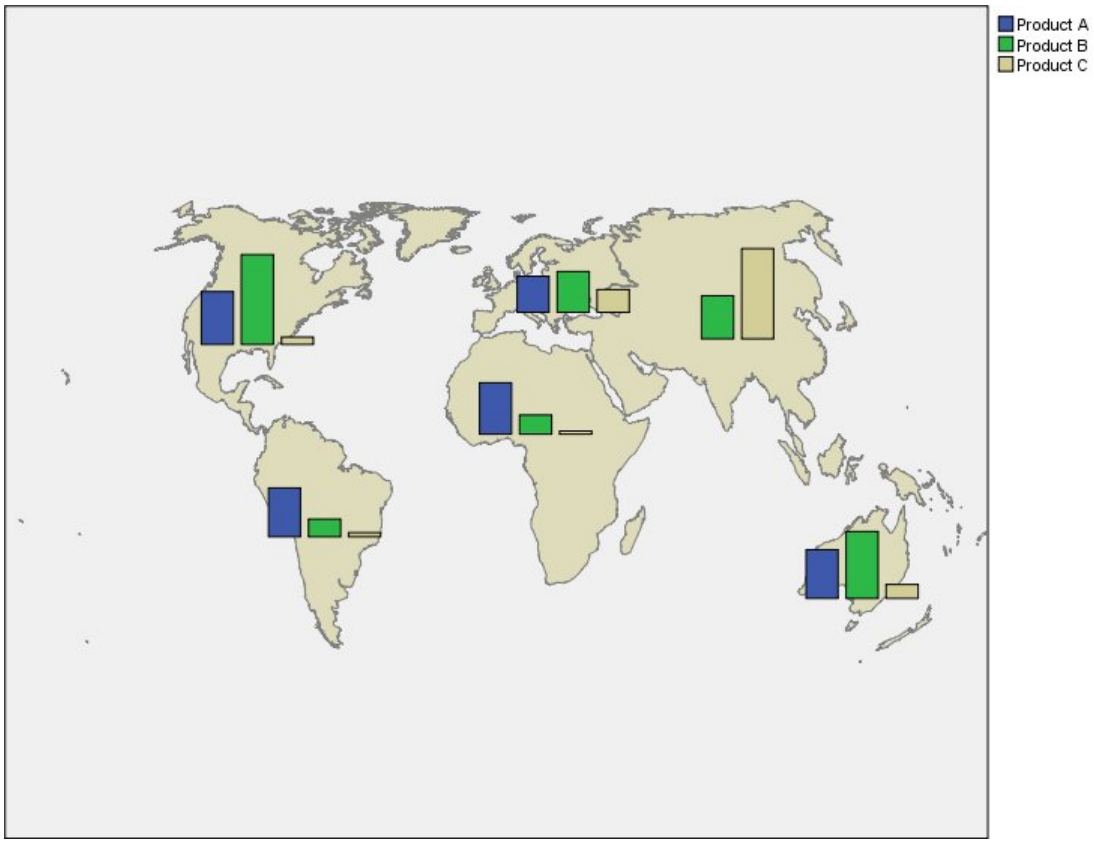


Рисунок 17. Столбчатая диаграмма на карте

Наблюдаются следующие результаты:

- Распределение суммарного дохода по товарам очень похоже в Северной Америке и Африке.
- *Товар С* дает наименьший доход всюду, за исключением Азии.
- Доход отсутствует или минимален от *Продукт А* в Азии.

Вкладка Вид панели выбора диаграмм

Перед созданием графиков можно задать опции внешнего вида.

Общие опции вида

Заголовок. Введите текст для использования в качестве заголовка диаграммы.

Подзаголовок. Введите текст для использования в качестве подзаголовка диаграммы.

Подпись. Введите текст для использования в качестве подписи диаграммы.

Выборка. Задайте способ выборки для больших наборов данных. Можно задать максимальный размер набора данных или использовать число записей по умолчанию. Для больших наборов данных производительность повышается, если выбрать опцию **Выборка**. Другой вариант - построение графика с использованием всех точек, задав опцию **Использовать все данные**, но нужно иметь в виду, что это может сильно снизить производительность программы.

Опции вида таблицы стилей

Есть две кнопки, которые позволяют управлять доступностью шаблонов визуализации (и таблиц стилей, и карт):

Управление . Управление шаблонами визуализаций, таблицами стилей и картами на компьютере. На вашем локальном компьютере можно импортировать, экспортировать, переименовывать и удалять шаблоны визуализации , таблицы стилей и отображения. Дополнительную информацию смотрите в разделе “Управление шаблонами, таблицами стилей и файлами карт” на стр. 209.

Местоположение. Изменение папки, в которой хранятся шаблоны визуализаций, таблицы стилей и карты Текущее местоположение показывается справа от кнопки. Дополнительную информацию смотрите в разделе “Указание местоположения для хранения шаблонов, таблиц стилей и карт”.

Следующий пример показывает, где на графике размещаются опции внешнего вида. (*Примечание*: Эти опции используются не для всех графиков.)

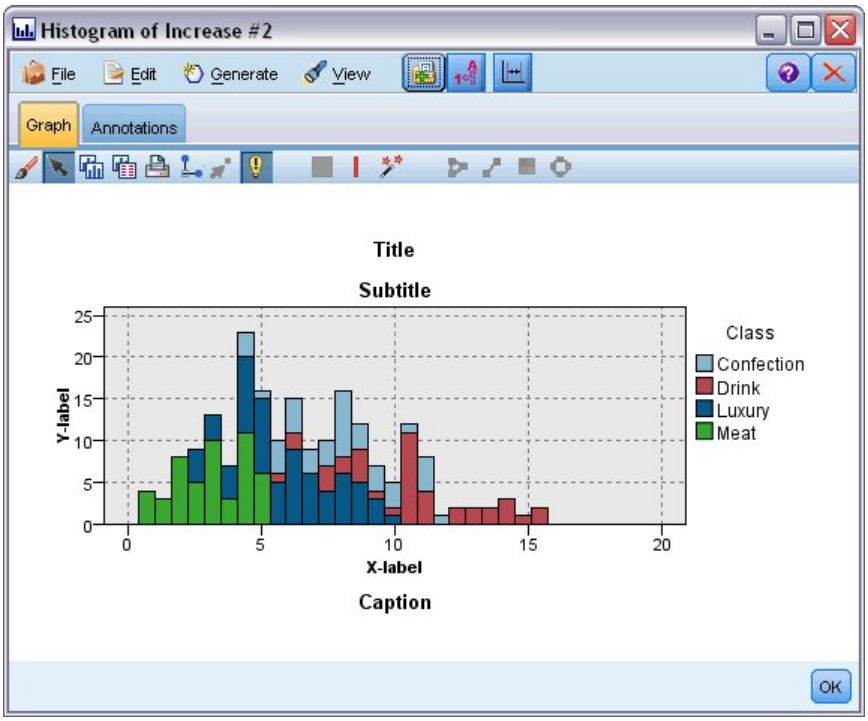


Рисунок 18. Размещение различных опций внешнего вида диаграммы

Указание местоположения для хранения шаблонов, таблиц стилей и карт

Шаблоны визуализации, таблицы стилей и файлы карт хранятся в определенной локальной папке или в IBM SPSS Collaboration and Deployment Services Repository. При выборе шаблонов, таблиц стилей и карт выводятся только встроенные для данной папки. Единое место хранения всех шаблонов, таблиц стилей и файлов карт облегчает работу с ними для приложений IBM SPSS. Дополнительную информацию о

добавлении шаблонов, таблиц стилей и файлов карт в это местоположение смотрите в разделе “Управление шаблонами, таблицами стилей и файлами карт”.

Указание местоположения для хранения шаблонов, таблиц стилей и файлов карт

1. В диалоговом окне шаблона или таблицы стилей нажмите **Местоположение...**, чтобы открыть диалоговое окно Шаблоны, таблицы стилей и файлы карт.
2. Выберите местоположение по умолчанию для хранения шаблонов, таблиц стилей и файлов карт:

   Локальный компьютер . Шаблоны, таблицы стилей и файлы карт хранятся в определенной папке на локальном компьютере. В системе Windows XP этой папкой будет *C:\Documents and Settings\ <user >\Application Data\SPSSInc\Graphboard* . Папку невозможно изменить.

   IBM SPSS Collaboration and Deployment Services Repository. Шаблоны, таблицы стилей и файлы карт хранятся в определенной пользователем папке в IBM SPSS Collaboration and Deployment Services Repository. Чтобы выбрать папку, нажмите **Папка** . Дополнительную информацию смотрите в разделе “Использование IBM SPSS Collaboration and Deployment Services Repository в качестве местоположения для хранения шаблонов, таблиц стилей и файлов карт”.
3. Щелкните по **OK**.

Использование IBM SPSS Collaboration and Deployment Services Repository в качестве местоположения для хранения шаблонов, таблиц стилей и файлов карт

Шаблоны визуализации и таблицы стилей можно хранить в IBM SPSS Collaboration and Deployment Services Repository. Этим местоположением является определенная папка в IBM SPSS Collaboration and Deployment Services Repository. Если это местоположение выбрано в качестве местоположения по умолчанию, то можно выбрать любые находящиеся там шаблоны, таблицы стилей и файлы карт.

Указание папки в IBM SPSS Collaboration and Deployment Services Repository для хранения шаблонов, таблиц стилей и файлов карт

1. В диалоговом окне с кнопкой Положение щелкните **Положение...**.
2. Выберите IBM SPSS Collaboration and Deployment Services Repository
3. Нажмите **Папка** .

   Примечание: Если вы еще не соединены с IBM SPSS Collaboration and Deployment Services Repository, появится запрос на информацию о соединении.
4. В диалоговом окне Выберите папку выберите папку, в которой будут храниться шаблоны, таблицы стилей и файлы карт.
5. Дополнительно вы можете выбрать метку из **Получить метку**. После этого будут отображаться только шаблоны, таблицы стилей и файлы карт с этой меткой.
6. Если требуется папка, в которой содержится определенный шаблон или таблица стилей, рекомендуется выполнить поиск этого шаблона, таблицы стилей или файла карты на вкладке Поиск. В диалоговом окне Выбрать папку автоматически выбирается папка, в которой располагается найденный шаблон, таблица стилей или файл карты.
7. Нажмите **Выбрать папку**.

Управление шаблонами, таблицами стилей и файлами карт

Используя диалоговое окно Управление шаблонами, таблицами стилей и картами можно управлять шаблонами, таблицами стилей и файлами карт, расположенными на вашем локальном компьютере. Это диалоговое окно позволяет импортировать, экспортировать, переименовывать и удалять шаблоны визуализаций, таблицы стилей и файлы карт.

Щелкните по **Управление...** в одном из диалоговых окон, где можно выбрать шаблоны, таблицы стилей или карты.

Диалоговое окно Управление шаблонами, таблицами стилей и картами

На вкладке Шаблон перечислены все шаблоны, хранящиеся на локальном компьютере. На вкладке Таблица стилей перечислены все таблицы стилей, хранящиеся на локальном компьютере, а также примеры визуализаций, созданные на основе гипотетических данных. Можно выбрать одну из таблиц стилей, чтобы применить ее стили к примеру визуализации. Дополнительную информацию смотрите в разделе “Применение таблиц стилей” на стр. 282. На вкладке Карта перечислены все локальные файлы карт. В этой вкладке также выводятся ключи карты, включая значения для примера, комментарий, если он был создан вместе с картой, и изображение предварительного просмотра карты.

Следующие кнопки доступны в любой из вкладок, которая открыта в настоящий момент.

Импорт . Импорт шаблона визуализаций, таблицы стилей или файла карты из файловой системы. Импорт шаблона, таблицы стилей или файла карты делает их доступными в приложении IBM SPSS. Если шаблон, таблицу стилей или файл карты вам передал другой пользователь, то прежде чем использовать, его (ее) необходимо импортировать.

Экспорт . Экспорт шаблона визуализаций, таблицы стилей или файла карты в файловую систему. Если шаблон, таблицу стилей или файл карты необходимо отправить ее другому пользователю, то сначала его (ее) нужно экспортировать.

Переименовать . Переименование выбранного шаблона визуализаций, таблицы стилей или файла карты. Невозможно изменить имя на уже используемое.

Экспортировать ключ карты. Экспортировать ключи карты в виде файла (CSV) значений, разделенных запятыми. Эта кнопка доступна только на вкладке Карта.

Удалить . Удаление выбранных шаблонов визуализаций, таблиц стилей или файлов карт. Вы можете выбрать несколько шаблонов, таблиц стилей или файл карт с помощью Ctrl-click Операцию удаления невозможно отменить, поэтому соблюдайте осторожность.

Конвертирование и распространение шейп-файлов карт.

Панель выбора диаграмм дает возможность визуализировать карту, используя комбинацию шаблона визуализации и файла SMZ. Файлы SMZ сходны с шейп-файлами ESRI (файлы формата SHP) в том, что они содержат географическую информация для изображения карты (например, границы стран), но они оптимизированы для визуализации карт. Панель выбора диаграмм предустановлена с выбранным числом файлов SMZ. Если у вас имеется шейп-файл ESRI, который вы хотите использовать для визуализации карт, то сначала необходимо конвертировать этот шейп-файл в файл SMZ, используя утилиту преобразования карт. Утилита преобразования карт поддерживает шейп-файлы ESRI с точками, полилиниями и полигонами (шейп-типы 1, 3 и 5), содержащие единственный слой.

Кроме конвертирования шейп-файлов ESRI утилита преобразования карт позволяет изменить уровень детализации карты, поменять метки элементов, соединить элементы, переместить элементы в числе многих других возможных изменений. Утилиту преобразования карт также можно использовать для модификации существующих файлов SMZ (включая предустановленные).

Редактирование предустановленных файлов SMZ

1. Экспорт файла SMZ из системы управления. Дополнительную информацию смотрите в разделе “Управление шаблонами, таблицами стилей и файлами карт” на стр. 209.
2. Чтобы открыть и редактировать экспортированный файл SMZ, воспользуйтесь утилитой преобразования карт. Рекомендуется сохранить файл под другим именем. Дополнительную информацию смотрите в разделе “Применение утилиты преобразования карт” на стр. 211.
3. Импорт измененного файла SMZ в систему управления. Дополнительную информацию смотрите в разделе “Управление шаблонами, таблицами стилей и файлами карт” на стр. 209.

Дополнительные ресурсы для файлов карт

Гео-пространственные данные в файловом формате SHP, которые можно использовать с целью построения карт, можно получить из многих частных и общественных источников. Проверьте местные государственные web-сайты, если вы ищите бесплатные данные. Многие из шаблонов в этом программном продукте основаны на общедоступных данных, полученных в GeoCommons ( ) и U.S. Census Bureau (http://www.census.gov).

ВАЖНОЕ ПРИМЕЧАНИЕ: Информация, относящаяся к продуктам сторонних разработчиков (не IBM), получена от поставщиков этих продуктов, из опубликованных ими сведений и из других публичных источников. Компания IBM не тестировала эти продукты и не может подтвердить правильность их работы, совместимость и другие утверждения, касающиеся продуктов, не принадлежащих компании IBM. Вопросы о возможностях этих продуктов следует направлять их поставщикам. Любые приводимые здесь ссылки на web-сайты, не относящиеся к компании IBM, даются исключительно для удобства и ни в коей мере не служат целям поддержки или рекламы этих web-сайтов. Материалы на этих web-сайтах не являются частью материалов для данного программного продукта компании IBM, если это не указано в сопровождающем его файле уведомлений, и материалами таких сайтов можно пользоваться только на свой страх и риск.

Ключевые понятия для карт

Понимание некоторых ключевых понятий, относящихся к шей-файлам, поможет эффективно применять утилиту преобразования карт.

Шейп-файл предоставляет географическую информацию для изображения карты. Существуют три типа шейп-файлов, которые поддерживает утилита преобразования карт:

- **Точка.** Шейп-файл определяет положение точек, таких как города.
- **Полилиния.** Шейп-файл определяет пути, такие как реки, и их положение.
- **Полигон.** Шейп-файл определяет ограниченные территории, такие как страны, и их положение.

Чаще всего приходится иметь дело с шейп-файлами с полигонами. Карты хороплетов создаются из шейп-файлов с полигонами. Карты хороплетов используют цвет, чтобы представить значения внутри отдельных полигонов (территорий). Шейп-файлы с точками и полилиниями обычно накладываются на шейп-файл с полигонами. Примером может служить шейп-файл с точками городов США, накладывающийся на шейп-файл с полигонами штатов США.

Шейп-файл состоит из **элементов**. Элементы представляют собой отдельные географические объекты. Например, элементами могут быть страны, штаты, города и т.д. Шейп-файл также содержит данные, относящиеся к элементам. Эти данные хранятся в **атрибутах**. Атрибуты имеют сходство с полям и переменными в файле данных. Для конкретного элемента имеется, по крайней мере, один атрибут, то есть **ключ карты**. Ключом карты может быть метка, такая как имя страны или штата. Ключ карты - это то, что привязывается к переменной/полю в файле данных, чтобы сформировать визуализацию карты.

Обратите внимание на то, что в файле SMZ можно сохранять только ключевой атрибут или атрибуты. Утилита преобразования карт не поддерживает сохранение дополнительных атрибутов. Это означает, что для объединения на различных уровнях необходимо создать несколько файлов SMZ. Например, если вы хотите объединить штаты и территории США, необходимо разделить файлы SMZ: одни из них с ключом для идентификации штатов, а другие - территорий.

Применение утилиты преобразования карт

Как запустить утилиту преобразования карт

Выберите в меню:

Инструменты > Утилита преобразования карт

Имеются четыре основных экрана (шага) для утилиты преобразования карт. Один из шагов также включает дополнительные шаги для более детального управления редактированием файла карты.

Шаг 1 - Выбрать файл назначения и исходный файл

Сначала необходимо выбрать исходный файл карты и местоположение конвертированного файла карты. Для шейп-файла вам потребуются оба файла *.shp* и *.dbf*.

Выберите для конвертирования файл .shp (ESRI) или .smz. Перейдите к существующему файлу карты на вашем компьютере. Это тот файл, который вы конвертируете и сохраните как файл SMZ. Файл *.dbf* для шейп-файла *должен* храниться в том же каталоге и с тем же именем, что и файл *.shp*. Файл *.dbf* необходим, поскольку он содержит информацию об атрибутах для файла *.shp*.

Задайте местоположение и имя для конвертированного файла карты. Введите путь и имя для файла SMZ, который будет создан из исходного источника карты.

- **Импорт в средство выбора шаблона.** В дополнение к сохранению файла в файловой системе можно добавить карту в список Управление на Панели выбора диаграмм. Если вы выберите эту возможность, то данная карта автоматически будет доступна на панели выбора диаграмм для программных продуктов IBM SPSS, установленных на вашем компьютере. Если не выполнить импорт на панель выбора диаграмм сейчас, то будет необходимо импортировать вручную позже. Более подробную информацию об импорте карт в систему управления выбором шаблонов смотрите в разделе “Управление шаблонами, таблицами стилей и файлами карт” на стр. 209.

Шаг 2 - Выбрать ключ карты

Теперь вы выберите, какие ключи карты включить в файл SMZ. Затем можно изменить некоторые параметры, которые влияют на изображение карты. Последующие шаги работы с утилитой преобразования карт включают предварительный просмотр карты. Выбранные параметры изображения будут использованы для предварительного просмотра карты.

Выберите исходный ключ карты. Выберите атрибут, то есть исходный ключ для идентификации и маркирования элементов на карте. Например, исходным ключом для карты мира может быть атрибут, задающий названия стран. Исходный ключ также свяжет ваши данные с элементами карты, поэтому убедитесь в том, что значения (метки) выбираемого атрибута будут соответствовать значениям ваших данных. При выборе атрибута выводятся примеры меток. Если нужно изменить эти метки, то такая возможность представится на одном из следующих шагов.

Выбрать для включения дополнительные ключи. В дополнение к исходному ключу карты пометьте любые другие атрибуты ключей, которые нужно включить в формируемый файл SMZ. Например, некоторые атрибуты могут содержать переведенные метки. Если ожидаются данные на других языках, то можно сохранить эти атрибуты. Обратите внимание на то, что можно выбрать только те дополнительные ключи, которые представляют те же элементы, что и исходный ключ. Например, если бы исходный ключ представлял собой полные названия штатов США, то можно было бы выбрать только такие альтернативные ключи, которые представляют штаты США, например, сокращения имен штатов.

Автоматически сгладить карту. Шейп-файлы с полигонами обычно содержат слишком много точек данных и слишком много деталей для визуализации карт со статистикой. Излишние детали могут рассеивать внимание и отрицательно влиять на восприятие. Вы можете снизить уровень детализации карты путем сглаживания. В результате карта будет выглядеть более "сжатой" и будет быстрее выводиться. Когда карта сглаживается автоматически, максимальный угол составляет 15 градусов, а сохраняемый процент - 99. Информацию об этих параметрах смотрите в разделе “Сгладить карту” на стр. 213. Обратите внимание на то, что имеется возможность дополнительного сглаживания на последующем шаге.

Удалить границы между соприкасающимися полигонами в одном и том же элементе. Некоторые элементы могут включать подэлементы, имеющие границы, которые являются внутренними по отношению к основному элементу. Например, карта мира с континентами может включать внутренние границы стран, расположенных на каждом из континентов. Если вы выберите эту возможность, то внутренние границы не будут изображаться на карте. В примере с картой мира выбор данной возможности приведет к удалению границ стран с сохранением границ континентов.

Шаг 3 - Редактировать карту

Теперь, когда заданы основные параметры для карты, можно отредактировать более специальные параметры. Такую модификацию делать необязательно. Данный шаг работы с утилитой преобразования карт проведет вас через последовательность взаимосвязанных задач и предварительный просмотр карты, который позволит проверить сделанные изменения. Некоторые задачи могут оказаться недоступными в зависимости от типа шейп-файла (с точками, полилиниями или полигонами) и системы координат.

Все задачи имеют следующие общие элементы управления на левой стороне диалогового окна утилиты преобразования карт.

Вывести метки на карте. По умолчанию метки и элементов не показываются при предварительном просмотре. Вы можете задать вывод меток. Хотя метки могут помочь идентифицировать элементы, они могут повлиять на прямой выбор на карте при предварительном просмотре. Включите эту возможность, когда это необходимо, например, при редактировании меток элементов.

Раскрасить карту для предварительного просмотра. По умолчанию при предварительном просмотре карты области окрашиваются с использованием чистого цвета. Все элементы имеют один и тот же цвет. Вы можете задать раскрашивание отдельных элементов карты разными цветами. Эта возможность может помочь различить различные элементы на карте. Это будет особенно полезно при соединении элементов, чтобы увидеть, как новые элементы будут изображаться при предварительном просмотре.

Все задачи также имеют следующий общий элемент управления на правой стороне диалогового окна утилиты преобразования карт.

Отменить. Нажмите кнопку **Отменить**, чтобы вернуться в последнее состояние. Можно отменить не более 100 изменений.

Сгладить карту: Шейп-файлы с полигонами обычно содержат слишком много точек данных и слишком много деталей для визуализации карт со статистикой. Излишние детали могут рассеивать внимание и отрицательно влиять на восприятие. Вы можете снизить уровень детализации карты путем сглаживания. В результате карта будет выглядеть более "сжатой" и будет быстрее выводиться. Эта возможность недоступна для карт с точками и полилиниями.

Максимальный угол. Максимальный угол, который должен иметь значение между 1 и 20, задает допуск для сглаживания наборов точек, которые расположены почти на одной прямой. Большее значение увеличивает допуск для линейного сглаживания и впоследствии приводит к отбрасыванию большего числа точек, приводя к менее детальной карте. Чтобы применить линейное сглаживание, утилита преобразования карт проверяет внутренний угол, сформированный каждым набором из трех точек на карте. Если 180 минус данный угол меньше заданного значения, то утилита преобразования карт отбрасывает среднюю точку. Иными словами, утилита преобразования карт проверяет, является ли линия, формируемая тремя рассматриваемыми точками, почти прямой. Если это так, то утилита преобразования карт рассматривает данную линию как прямую линию между оконечными точками и отбрасывает среднюю точку.

Сохраняемый процент. Сохраняемый процент, который должен быть значением между 90 и 100, задает сохраняемый размер участка земли при сглаживании карты. Эта возможность действует только для элементов, имеющих несколько полигонов, как это имеет место для элементов, включающих острова. Если общая площадь элемента минус полигон превосходит заданный процент исходной площади, то утилита преобразования карт удалит полигон с карты. Утилита преобразования карт никогда не удалит все полигоны для данного элемента. То есть, всегда сохранится, по крайней мере, один полигон для элемента, независимо от применяемой степени сглаживания.

Выбрав максимальный угол и сохраняемый процент, щелкните по **Применить**. Изображение предварительного просмотра будет обновлено в соответствии с изменениями сглаживания. Если вам снова нужно сгладить карту, повторите это действие, пока не будет достигнут нужный уровень сглаживания. Обратите внимание на то, что на сглаживание накладывается ограничение. При повторном сглаживании можно достичь момента, когда станет невозможно продолжать сглаживать данную карту.

Отредактировать метки элементов: Метки элементов можно при необходимости редактировать (возможно, чтобы привести в соответствие с ожидаемыми данными), а также, чтобы изменить положение меток на карте. Даже если вы не думаете, что метки нужно изменять, их следует просмотреть, прежде чем создать визуализацию карты. Так как по умолчанию метки не выводятся при предварительном просмотре, то для их вывода поставьте переключатель **Вывести метки на карте**.

Ключи. Выберите ключ, содержащий метки элементов, которые требуется просмотреть и/или отредактировать.

Элементы. Этот список выводит метки элементов, которые содержит выбранный ключ. Чтобы отредактировать метку, дважды щелкните по ней в списке. Если метки показаны на карте, можно также дважды щелкнуть по метке элемента прямо на изображении предварительного просмотра карты. Если вы хотите сравнить данные метки с метками в имеющемся файле данных, щелкните по **Сравнить**.

X/Y. В этих текстовых полях представлены текущие центральные точки метки выбранной возможности на карте. Значениями являются координаты на карте. Они могут быть локальными декартовыми координатами (например, State Plane Coordinate System в США) или географическими координатами (где **X** - долгота, а **Y** - широта). Введите координаты нового положения метки. Если метки выведены, то также можно щелкнуть по метке на карте и перетащить ее. В текстовых полях положение будет обновлено.

Сравнить. Если имеется файл данных, который содержит значения данных, предназначенные для сопоставления меток элементов для конкретного ключа, щелкните по **Сравнить**, чтобы вывести диалоговое окно Сравнить с внешним источником данных. В этом диалоговом окне можно открыть файл данных и сравнить его значения напрямую со значениям меток элементов на карте для ключа.

Диалоговое окно Сравнить с внешним источником данных: Диалоговое окно Сравнить с внешним источником данных позволяет открыть файл значений, разделенных табуляцией (с расширением *.txt*), или файл значений, разделенных запятой (с расширением *.csv*), или файл данных, отформатированный для IBM SPSS Statistics (с расширением *.sav*). Когда такой файл открыт, можно выбрать поле в этом файле данных, чтобы сравнить метки элементов для конкретного ключа карты. Затем можно откорректировать любые несоответствия в файле карты.

Поля в файле данных. Выберите поле, значения которого нужно сравнить с метками элементов. Если первая строка в файле *.txt* или *.csv* содержит описательные метки для каждого поля, поставьте переключатель **Использовать первую строку как метки столбцов**. В противном случае каждое поле будет идентифицироваться по его положению в файле данных (например "Столбец 1", "Столбец 2" и т.д.).

Ключ для сравнения. Выберите ключ карты, для которого нужно сравнить метки элементов со значениями поля файла данных.

Сравнение. Щелкните, когда все готово для сравнения значений.

Результаты сравнения. По умолчанию в таблице результатов сравнения выводятся только те значения поля в файле данных, для которых нет соответствия. Программа пытается найти соответствующую метку элемента обычно путем выявления вставленных или пропущенных пробелов. Щелкните по раскрывающемуся списку в столбце *Метка карты*, чтобы сопоставить метку элемента в файле карты с выведенным значением поля. Если в вашем файле карты отсутствует соответствующая метка элемента, выберите *Оставить без сопоставления*. Если нужно увидеть все значения в полях, даже те, которые уже соответствуют метке возможности, выключите переключатель **Выводить только несогласованные случаи**. Вы можете это сделать, чтобы переопределить одно или несколько сопоставлений.

Каждый элемент можно использовать только один раз для его сопоставления со значением поля. Если необходимо сопоставить несколько элементов одному значению поля, то можно соединить элементы и затем сопоставить значению поля новый соединенный элемент. Более подробную информацию о возможностях объединения смотрите в разделе “Соединить элементы” на стр. 215.

Соединить элементы: Соединение элементов используется для создания более крупных территорий на карте. Например, при конвертировании карты штатов, можно было бы объединить штаты (элементы в данном примере) в более крупные Северный, Южный, Восточный и Западный районы.

Ключи. Выберите ключ карты, содержащий метки элементов, что поможет идентифицировать элементы, которые требуется соединить.

Элементы. Щелкните по первому элементу, который нужно соединить. Удерживая нажатой клавишу Ctrl, щелкните по другим элементам, которые нужно соединить. Обратите внимание на то, что эти элементы также будут выделены на изображении предварительного просмотра карты. Вы можете выделить элементы прямо на изображении предварительного просмотра карты, щелкая по ним, удерживая, когда это необходимо, нажатой клавишу Ctrl .

Выбрав элементы для соединения, щелкните по **Соединить**, чтобы вывести диалоговое окно Метка объединенного элемента, где можно присвоить метку новому элементу. После соединения элементов можно поставить переключатель **Раскрасить карту для предварительного просмотра**, чтобы убедиться в том, что получились результаты, которые вы ожидали.

После соединения элементов можно переместить метку нового элемента. Это можно сделать, выполняя задачу *Отредактировать метки элементов*. Дополнительную информацию смотрите в разделе “Отредактировать метки элементов” на стр. 214.

Диалоговое окно Метка объединенного элемента: Диалоговое окно Метка объединенного элемента позволяет назначить метки новому объединенному элементу.

Таблица Метки выводит информацию для каждого ключа в файле карты и позволяет назначить метку для каждого ключа.

Новая метка. Введите новую метку для объединенного элемента, чтобы назначить ее конкретному ключу карты.

Ключ. Ключ, для которого назначается новая метка.

Старые метки. Метки элементов, которые будут соединены в новый элемент.

Удалить границы между соприкасающимися полигонами. Поставьте этот переключатель, чтобы удалить границы элементов, которые были соединены вместе. Например, если штаты объединены в географические районы, то эта возможность позволит удалить границы отдельных штатов.

Переместить элементы: Элементы на карте можно перемещать. Это может оказаться полезным, когда нужно соединить элементы вместе, как для случая материка и удаленных островов.

Ключи. Выберите ключ карты, содержащий метки элементов, что поможет идентифицировать элементы, которые требуется переместить.

Элементы. Щелкните по первому элементу, который нужно переместить. Обратите внимание на то, что этот элемент также будут выделен на изображении предварительного просмотра карты. Можно также щелкнуть по элементу прямо на изображении предварительного просмотра карты.

X/Y. В этих текстовых полях представлены текущие центральные точки возможности на карте. Значениями являются координаты на карте. Они могут быть локальными декартовыми координатами (например, State Plane Coordinate System в США) или географическими координатами (где **X** - долгота, а **Y** - широта). Введите координаты нового положения элемента. Можно также щелкнуть по элементу на карте и перетащить его. В текстовых полях положение будет обновлено.

Удалить элементы.: Ненужные элементы на карте можно удалить. Это может оказаться полезным, когда нужно избавиться от нагромождения элементов, удаляя те из них, которые не представляют интереса при визуализации карты.

Ключи. Выберите ключ карты, содержащий метки элементов, что поможет идентифицировать элементы, которые требуется удалить.

Элементы. Щелкните по первому элементу, который нужно удалить. Если требуется удалить несколько элементов одновременно, щелкните по дополнительным элементам, удерживая нажатой клавишу Ctrl. Обратите внимание на то, что эти элементы также будут выделены на изображении предварительного просмотра карты. Вы можете выделить элементы прямо на изображении предварительного просмотра карты, щелкая по ним, удерживая, когда это необходимо, нажатой клавишу Ctrl .

Удалить отдельные компоненты: Кроме удаления элементов целиком можно удалить некоторые из отдельных компонентов, из которых состоят элементы, такие как озера и небольшие острова. Эта возможность недоступна для карт с точками.

Компоненты. Щелкните по компонентам, который нужно удалить. Если требуется удалить несколько компонентов одновременно, щелкните по дополнительным компонентам, удерживая нажатой клавишу Ctrl. Обратите внимание на то, что эти компоненты также будут выделены на изображении предварительного просмотра карты. Вы можете выделить компоненты прямо на изображении предварительного просмотра карты, щелкая по ним, удерживая, когда это необходимо, нажатой клавишу Ctrl . Так как список имен элементов не описательный (каждому элементу назначается номер в этой возможности), необходимо проверить при предварительном просмотре карты, что выбраны правильные элементы.

Задать проекцию:

Проекция карты задает способ, которым трехмерная Земля представляется в двух измерениях. Все проекции приводят к искажениям. Однако некоторые проекции оказываются более подходящими, в зависимости от того, рассматривается ли глобальная карта или более локальная. Кроме того, некоторые проекции сохраняют форму исходных элементов. Проекции, которые сохраняют форму, являются конформными проекциями. Такая возможность доступна только для карт с географическими координатами (долгота и широта).

В отличие от других возможностей, предоставляемых утилитой преобразования карт, проекцию можно изменить после того, как создана визуализация карты.

Проекция. Выберите проекцию карты. При создании всемирной или полусферической карты используйте проекции *Локальная*, *Меркатора* или *Трипель Уинкеля*. Для территорий меньшего размера используйте проекции *Локальная*, *Конформная коническая Ламберта* или *Поперечная Меркатора*. Все проекции используют эллипсоид WGS83 в качестве базы.

- **Локальная** проекция используется всегда, когда карта была создана с использованием локальной системы координат, такой как State Plane Coordinate System в США. Эти системы координат задаются декартовыми координатами, а не географическими координатами (долгота и широта). В локальной проекции горизонтальные и вертикальные линии имеют одинаковый масштаб в декартовой системе координат. Локальная проекция не является конформной.
- Проекция **Меркатора** является конформной проекцией для всемирных карт. Горизонтальная и вертикальная линии являются прямыми и всегда перпендикулярны друг другу. Обратите внимание на то, что проекция Меркатора стремится к бесконечности по мере приближения к Северному и Южному полюсам, поэтому ее нельзя использовать, если карта включает Северный или Южный полюс. Искажение увеличивается по мере приближения к этим границам.
- Проекция **Трипель Уинкеля** является неконформной проекцией для всемирных карт. Хотя она не является конформной, она обеспечивает удачный баланс между формой и размером. Исключая экватор и нулевой меридиан, все линии являются изогнутыми. Если всемирная карта включает Северный или Южный полюс, то эта проекция является хорошим выбором.

- Как ее имя и предполагает, **Конформная коническая Ламберта** проекция является конформной и используется для карт участков земли, размера континентов или меньшего, которые больше в направлении на Восток и Запад, чем в направлении на Север и Юг.
- **Поперечная Меркатора** является еще одной конформной проекцией для карт участков земли, размера континентов или меньшего. Используйте эту проекцию для участков земли, которые больше в направлении на Север и Юг, чем в направлении на Восток и Запад.

Шаг 4 - Завершить работу

На этом этапе можно добавить комментарий для описания файла карты, а также создать файл данных примера на основе ключей карты.

Ключи карты. Если в файле карты имеются несколько ключей, выберите ключ карты, для которого нужно вывести метки элементов при предварительном просмотре. При создании файла данных на основе карты, эти метки будут использованы для значений данных.

Комментарий. Введите комментарий, описывающий карту или предоставляющий дополнительную информацию, которая может быть важной для пользователей, например, источники исходных шейп-файлов. Этот комментарий появится в системе управления панели выбора диаграмм.

Создать набор данных на основе меток элементов. Установите этот переключатель, если необходимо создать текстовый файл данных на основе выведенных меток элементов. Нажав кнопку **Просмотр...**, вы сможете задать положение и имя файла. Если добавить расширение *.txt*, то файл будет сохранен как файл значений, разделенных символами табуляции. Если добавить расширение *.csv*, то файл будет сохранен как файл значений, разделенных запятыми. Если добавить расширение *.sav*, файл будет сохранен в формате IBM SPSS Statistics. SAV является значением по умолчанию, когда расширение не задается.

Распространение файлов карты

На первом шаге работы с утилитой преобразования карт выбирается каталог для сохранения конвертированного файла SMZ. Также выбирается, добавить ли карту в систему управления для панели выбора диаграмм. Если выбрано сохранение файла в системе управления, то данная карта будет доступна в любом программном продукте IBM SPSS, который запускается на том же компьютере.

Чтобы передать данную карту для использования другим пользователям, необходимо переслать им файл SMZ. Эти пользователи тогда смогут использовать систему управления, чтобы импортировать карту. Можно просто послать файл, расположение которого задано на шаге 1. Если необходимо послать файл, который имеется в системе управления, то сначала его нужно экспортировать:

1. На панели выбора диаграмм щелкните по **Управление...**
2. Выберите вкладку Карта
3. Выберите карту, которую нужно передать другим пользователям.
4. Щелкните по **Экспорт...** и выберите каталог, в котором нужно сохранить файл.

Теперь можно можно послать физический файл карты другим пользователям. Пользователям потребуется выполнить данную операцию в обратном порядке и импортировать карту в систему управления.

Узел График

Узлы График показывают взаимосвязь между значениями численных полей. Графики можно строить с помощью точек (такой график называют диаграммой рассеяния, scatterplot) или линий. Задав режим X в диалоговом окне, можно создать три типа линейных графиков.

Режим X = Sort

Если задать для режима X значение **Sort**, данные будут сортироваться по значениям для поля, откладываемым по оси *x*. При этом график представляет собой одну линию, идущую слева направо.

Использование номинального поля для перекрытия приводит к изображению на графике нескольких линий разного оттенка, идущих слева направо.

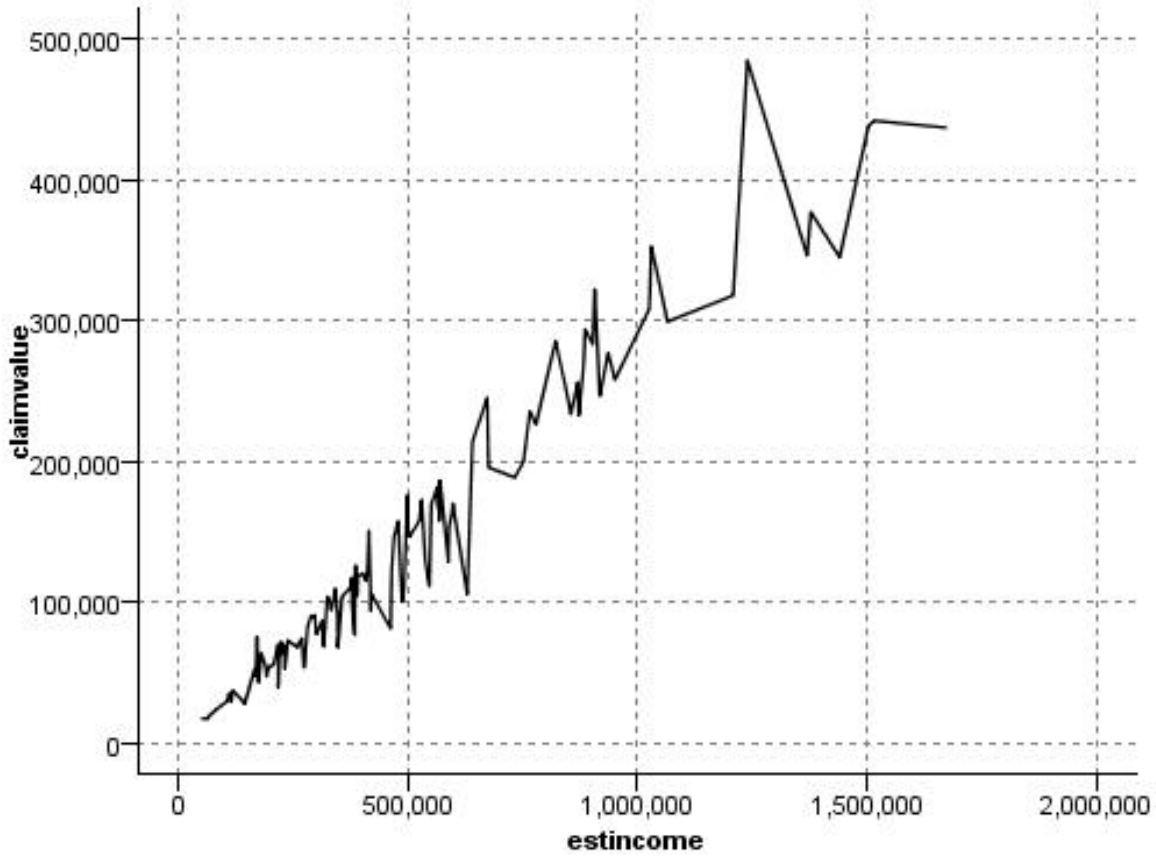


Рисунок 19. Линейный график при заданном для режима X значении Sort

Режим X = Overlay

При задании для режима X значения **Overlay** один график состоит из нескольких линейных графиков. Для графика с перекрытием данные не сортируются; пока значения по оси *x* увеличиваются, данные будут наноситься на график в виде одной линии. Если значения уменьшаются, возникнет новая линия. Например, при возрастании *x* от 0 до 100, значения *y* будут откладываться по одной линии. Когда значение *x* окажется меньше 100, в дополнение к первой линии начнет рисоваться новая. На конечном графике может оказаться множество графиков, полезных для сравнения нескольких рядов значений *y*. Такой тип графиков полезен для данных с периодической временной компонентой, как в случае потребления электроэнергии в последовательные 24-часовые промежутки времени.

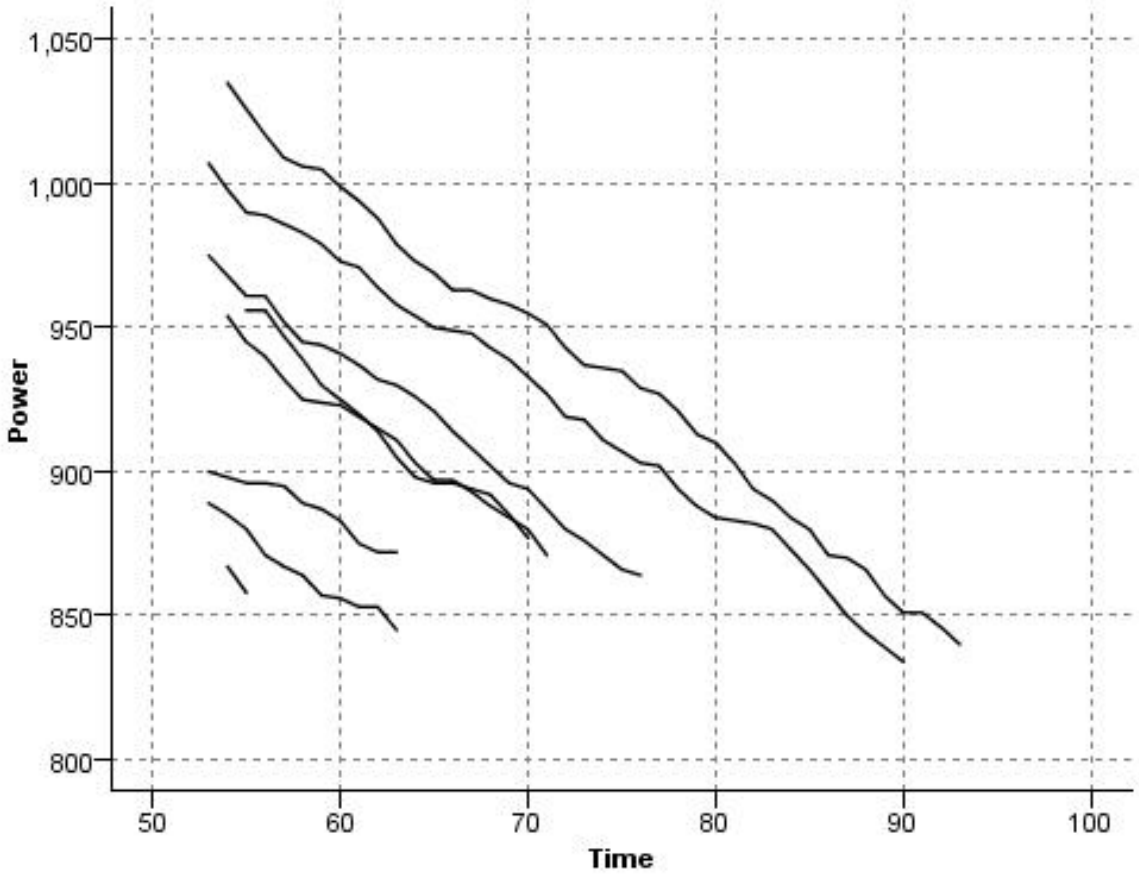


Рисунок 20. Линейный график при заданном для режима X значении Overlay

Режим X = As Read

При задании для режима X значения **As Read** значения *x* и *y* откладываются на графике, как они считываются из источника данных. Эта опция полезна для данных с компонентой временных рядов, когда вы интересуетесь тенденциями или структурами, зависящими от порядка данных. Перед созданием этого типа графика может потребоваться отсортировать данные. Может быть полезно также сравнить два аналогичных графика, у которых для режима X заданы значения **Sort** и **As Read**, чтобы определить, насколько структура зависит от сортировки.

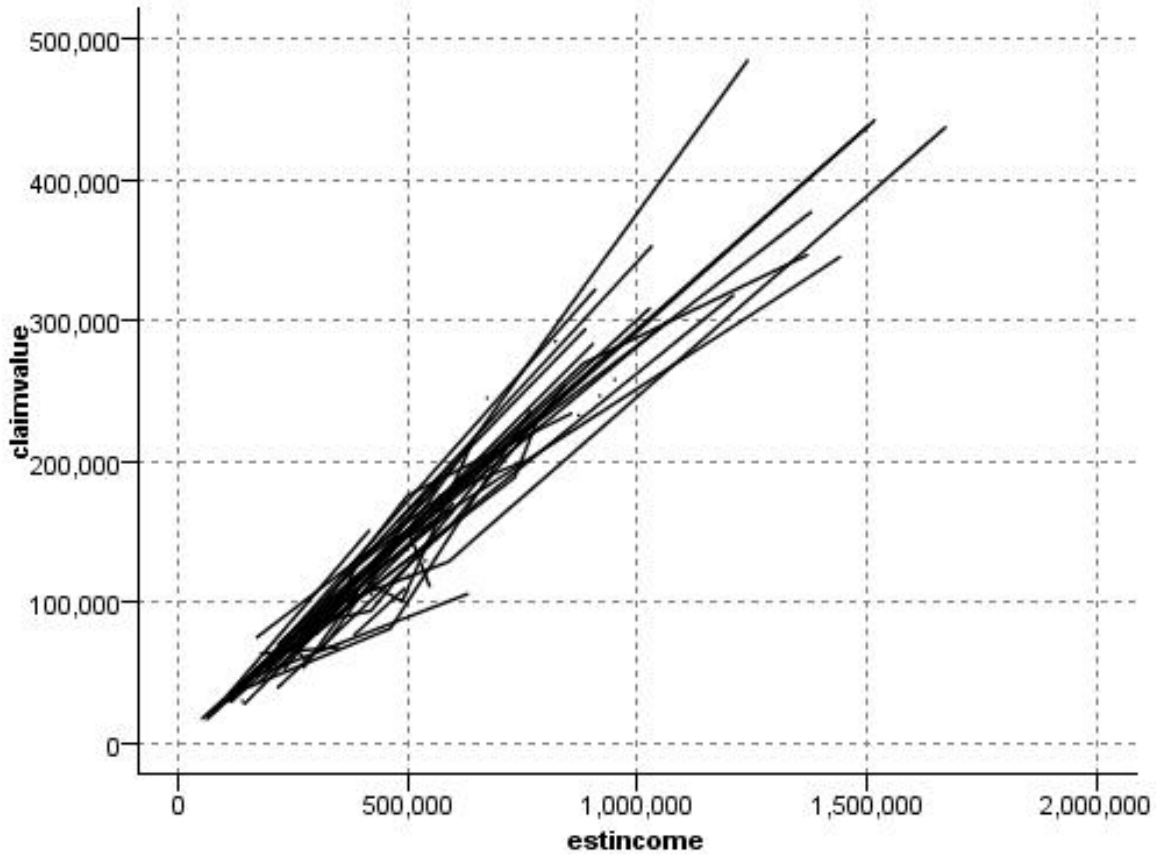


Рисунок 21. Линейный график, ранее выведенный при значении Sort для режима X, построенный заново при значении As Read

Для построения диаграмм рассеяния и линейных графиков можно использовать также узел Graphboard (панель выбора диаграмм). Однако из этого узла можно выбрать больше опций. Дополнительную информацию смотрите в разделе “Доступные встроенные типы визуализации Панели выбора диаграмм” на стр. 190.

Вкладка Узел графика

На графике выводятся значения поля *Y* относительно значений поля *X*. Часто эти поля соответствуют соответственно зависимой переменной и независимой переменной.

Поле X. Выберите в списке поле для вывода по горизонтальной оси *x*.

Поле Y. Выберите в списке поле для вывода по вертикальной оси *y*.

Поле Z. После нажатия кнопки Трехмерная диаграмма можно будет выбрать поле в списке для вывода по оси *z*.

Наложение. Показать категории для значений данных можно несколькими способами. Например, можно использовать *maincrop* (ведущая культура) как наложение цвета для указания значений *estincome* (оцененная прибыль) и *claimvalue* (заявленная полезность) для ведущей культуры, выращенной соискателями. Дополнительную информацию смотрите в разделе “Эстетики, наложения, панели и анимация” на стр. 180.

Тип наложения. Задает, будет ли выводиться функция наложения или сглаживатель. Функции сглаживателя и наложения всегда вычисляются как функция y.

- **Нет.** Наложение не взводится.
- **Сглаживатель.** Выводит сглаженную линию совпадений, вычисленную при помощи итерационного устойчивого выражения локально взвешенных наименьших квадратов. Этот метод эффективно вычисляет ряд регрессий, каждая из которых направлена на малую область графика. Он генерирует серию "локальных" линий регрессий, которые затем объединяются для создания сглаженной кривой.

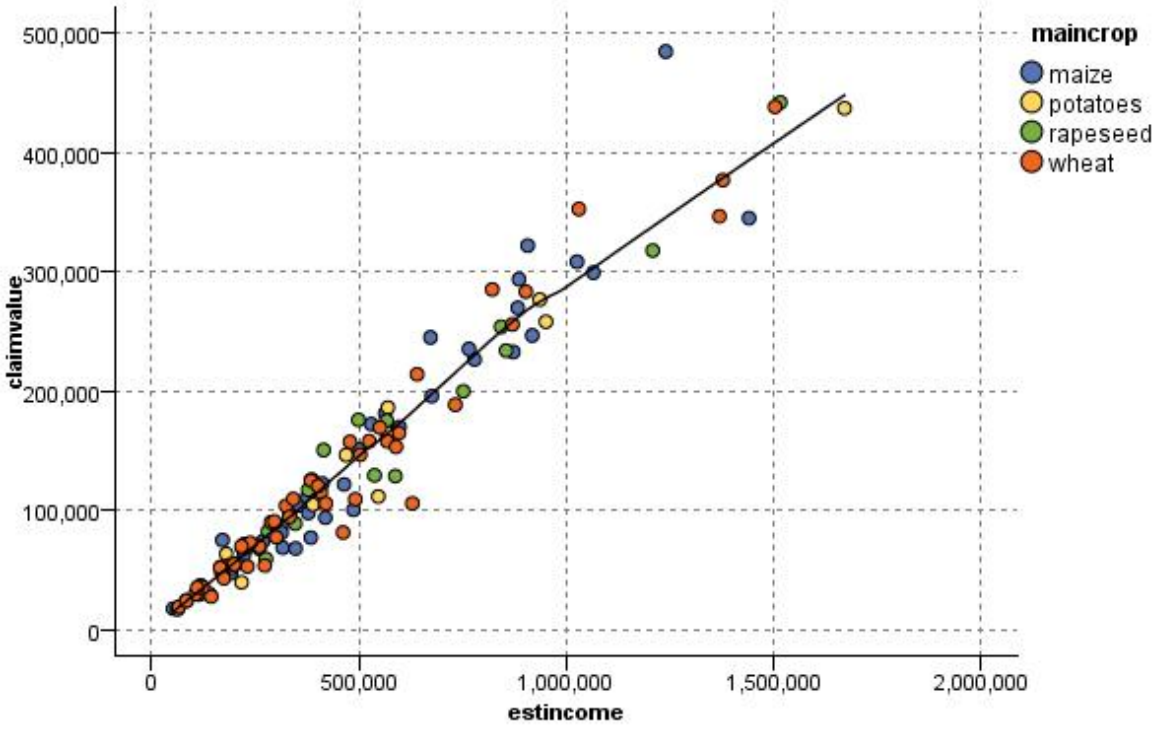


Рисунок 22. График с наложением сглаживателя LOESS

- **Функция.** Выберите эту опцию, чтобы задать известную функцию для сравнения фактических значений. Например, чтобы сравнить фактические значения с предсказанными, на график можно нанести в качестве наложения функцию y = x. Задайте в текстовом окне функцию для **y =**. По умолчанию используется функция y = x, но можно функцию любого другого типа, например, квадратичную функцию или произвольное выражение в значениях x.

  Примечание: Функции наложения для панельной и анимационной диаграммы недоступны.

После задания опция для графика его можно вызвать непосредственно из этого диалогового окна, нажав кнопку **Выполнить**. Однако на вкладке Опции можно задать дополнительные спецификации, такие как разделение на интервалы, режим X и стиль.

Вкладка Опции графика

Стиль. Выберите в качестве стиля графика **Точечный** или **Линейный**. Выбор опции **Линейный** активирует элемент управления **Режим X**. При выборе **Точечный** в качестве формы точки по умолчанию будет использоваться знак плюс (+). После создания диаграммы форму и размер точки можно изменить.

Режим X. Для линейных графиков, выбрав Режим X, нужно определить стиль лилейного графика. Выберите **Сортировка**, **Наложение** или **Как читается**. Для опции **Наложение** или **Как читается** нужно задать

максимальный размер набора данных, используемый для выборки первых *n* записей. В противном случае будет использоваться размер по умолчанию 2000 записей.

Автоматический диапазон по X. Выберите эту опцию, чтобы использовать весь диапазон значений данных по этой оси. Выключите этот переключатель, чтобы использовать явный поднабор значений на основе заданных вами значений **Мин** и **Макс**. Введите значение или задайте его кнопками со стрелками. По умолчанию выбраны автоматические диапазоны для быстрого построения диаграмм.

Автоматический диапазон по Y. Выберите эту опцию, чтобы использовать весь диапазон значений данных по этой оси. Выключите этот переключатель, чтобы использовать явный поднабор значений на основе заданных вами значений **Мин** и **Макс**. Введите значение или задайте его кнопками со стрелками. По умолчанию выбраны автоматические диапазоны для быстрого построения диаграмм.

Автоматический диапазон по Z. Только, если на вкладке График задана трехмерная диаграмма. Выберите эту опцию, чтобы использовать весь диапазон значений данных по этой оси. Выключите этот переключатель, чтобы использовать явный поднабор значений на основе заданных вами значений **Мин** и **Макс**. Введите значение или задайте его кнопками со стрелками. По умолчанию выбраны автоматические диапазоны для быстрого построения диаграмм.

Дрожание. Называемое также **возбуждением**, дрожание полезно, для точечных графиков наборов данных, где многие значения повторяются. Чтобы увидеть более четкое распределение значений, можно применить дрожание, чтобы распределить точки случайным образом вокруг фактического значения.

Примечание для пользователей более ранних версий IBM SPSS Modeler: В качестве значения дрожания, используемого на графике, в этом выпуске IBM SPSS Modeler используется другой показатель. В прежних версиях это значение было фактическим числом, но сейчас оно представляет собой долю от размера кадра. Это означает, что значения возбуждения в старых потоках, скорее всего, будут слишком большими. Для этого выпуска ненулевые значения возбуждения будут преобразованы в значение 0,2.

Максимальное число записей к графику. Задайте метод нанесения на диаграмму больших наборов данных. Можно задать максимальный размер набора данных или использовать размер по умолчанию 2000 записей. Для больших наборов данных производительность повысится, если выбрать опцию **Интервал** или **Выборка**. Другой вариант - построение графика с использованием всех точек, задав опцию **Использовать все данные**, но нужно иметь в виду, что это может сильно снизить производительность программы.

Примечание: Если задан Режим X **Наложение** или **Как читается**, эти опции отключаются, и используются только первые *n* записей.

- **Ящик.** Выберите эту опцию, чтобы включать разделение на интервалы, когда набор данных содержит больше записей, чем заданное число. При разделении на интервалы диаграмма перед фактическим построением графика разбивается на мелкую сетку, и подсчитывается число точек, которые будут выводиться в каждой из ячеек сетки. На окончательной диаграмме будет построена одна точка для каждой ячейки в центроиде интервала (среднем от всех положений в интервале). Размер символических элементов на диаграмме соответствует числу точек в данном регионе (если только размер не используется как наложение). Использование центроидов и размеров для представления числа точек делает график, разделяемый на интервалы, преимущественным способом представления больших наборов данных, поскольку предотвращает перекрытие в регионах большой плотности (с недифференцируемыми градациями цвета) и сокращает число артефактов символических элементов (синтезированных шаблонов плотности). Артефакты символических элементов проявляются там, где определенные символические элементы (в частности знак плюс [+]) сталкиваются способом, приводящим к образованию областей большой плотности, не представленных в необработанных данных.
- **Выборка.** Выберите эту опцию для случайной выборки данных в соответствии с числом записей, введенных в текстовом поле. Значение по умолчанию - 2000.

Вкладка Внешний вид графика

Перед созданием графиков можно задать опции внешнего вида.

Заголовок. Введите текст для использования в качестве заголовка диаграммы.

Подзаголовок. Введите текст для использования в качестве подзаголовка диаграммы.

Подпись. Введите текст для использования в качестве подписи диаграммы.

Метка X. Либо примите метку оси *x* (горизонтальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Метка Y. Либо примите метку оси *y* (вертикальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Метка Z. Доступно только для трехмерных диаграмм. Либо примите метку оси *z*, сгенерированную автоматически, либо выберите **Пользовательская**, чтобы задать пользовательскую метку.

Показать сетку. Эта опция, включаемая по умолчанию, выводит позади графика сетку, упрощающую определение точек отсечения для регионов и полос. Линии сетки всегда выводятся белым цветом, если только не используется белый фон диаграммы; в этом случае они выводятся серым цветом.

Использование диаграммы графика

Простые графики и графики множественных зависимостей - это по существу зависимости, построенные по координатам *X* и *Y*. Например, если вы изучаете возможные мошенничества в заявках на получение сельскохозяйственных субсидий, возможно, вы захотите построить график заявленной прибыли относительно оценки прибыли при помощи нейросети. Применение наложения (например, для категории культуры), покажет, есть ли взаимосвязь между заявленными свойствами (полезностью или количеством) и категорией культуры.

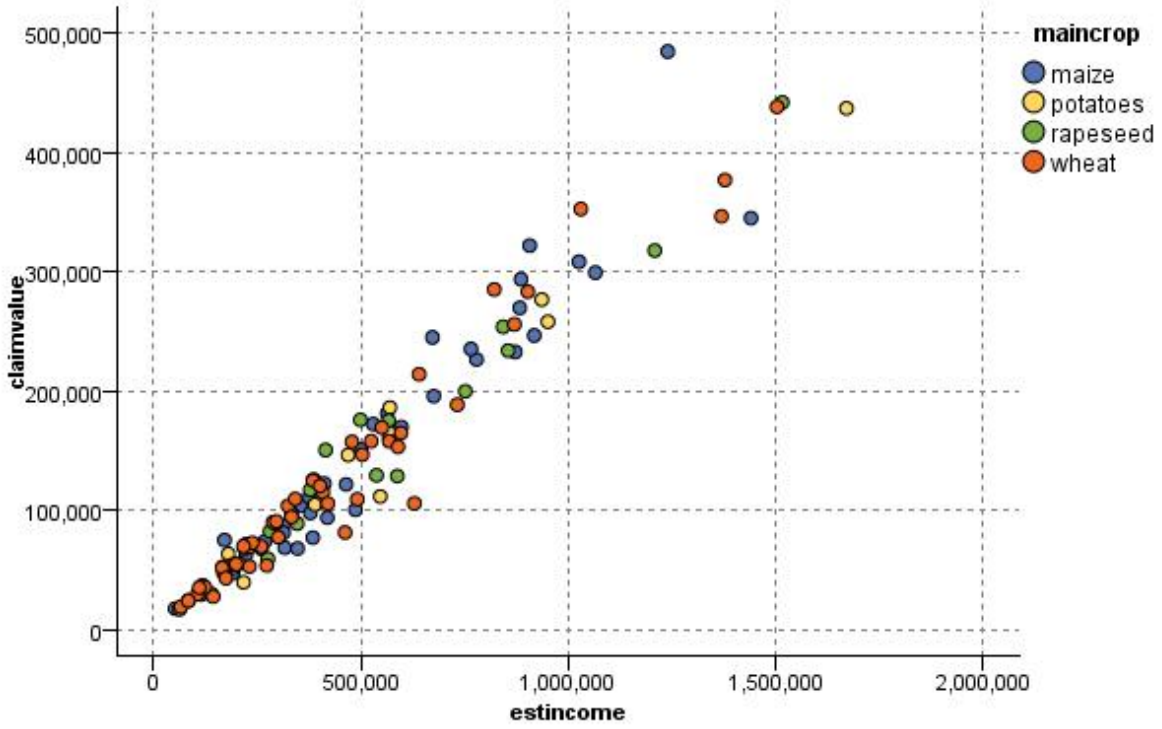


Рисунок 23. График взаимосвязи между оцененной прибылью и заявленной полезностью с категорией ведущей культуры в качестве наложения

Поскольку простые графики, графики нескольких зависимостей и графики оценок - это двумерные представления, выводимые по осям *Y* и *X*, при работе с ними можно легко определить регионы, пометить элементы или даже нарисовать полосы. Можно также сгенерировать узлы для данных, представленных этими регионами, полосами или элементами. Дополнительную информацию смотрите в разделе “Исследование графиков” на стр. 260.

Узел Несколько графиков

Несколько графиков (Multiplot) - это специальный тип графика, на котором выводятся значения нескольких полей *Y* относительно значений одного поля *X*. Значения полей *Y* изображаются на графике как цветные линии, каждая из которых эквивалентна графику на узле График при заданном значении стиля **Линия** и режиме X **Сортировка**. Узел Несколько графиков полезен, когда у вас есть данные временных рядов, и нужно исследовать флуктуации нескольких переменных во времени.

Вкладка График множественных зависимостей

Поле X. Выберите в списке поле для вывода по горизонтальной оси *x*.

Поля Y. Выберите одно или несколько полей в списке для вывода по диапазону значений полей *X*. Для выбора нескольких полей используйте кнопку Средство выбора полей. Для удаления полей из списка нажмите кнопку Удалить.

Наложение. Показать категории для значений данных можно несколькими способами. Например, при помощи анимационного наложения можно вывести несколько графиков для значений в данных. Эта опция полезна для наборов полей, содержащих свыше 10 категорий. При использовании той опции для наборов

данных с более чем 15 категориями можно заметить снижение производительности. Дополнительную информацию смотрите в разделе “Эстетики, наложения, панели и анимация” на стр. 180.

Нормализовать. Выберите эту опцию для масштабирования всех значений *Y* с переводом в диапазон 0–1 для вывода на диаграмме. Нормализация помогает изучить взаимосвязь между линиями, которые в противном случае были бы скрыты из-за различий в диапазонах значений для каждого ряда, и рекомендуется при нанесении на одну и ту же диаграмму нескольких зависимостей или при сравнении графиков на примыкающих друг к другу панелях. (Если все значения данных попадают в один и тот же диапазон, нормализация не требуется.)

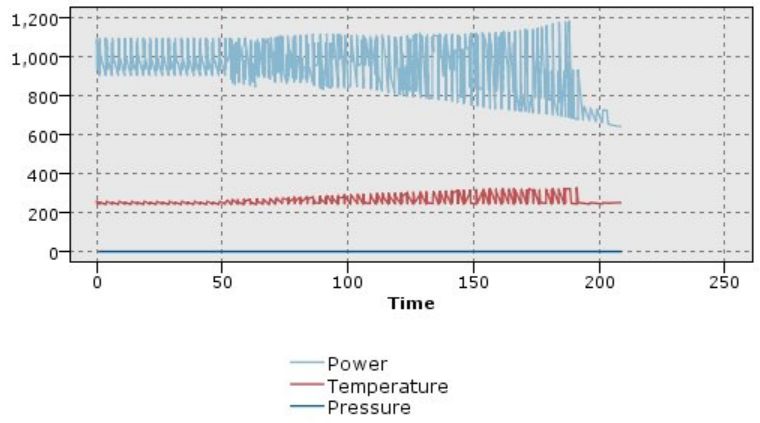


Рисунок 24. Стандартный график нескольких зависимостей, показывающий неравномерность работы силовой установки по времени (без нормализации; график для давления невиден)

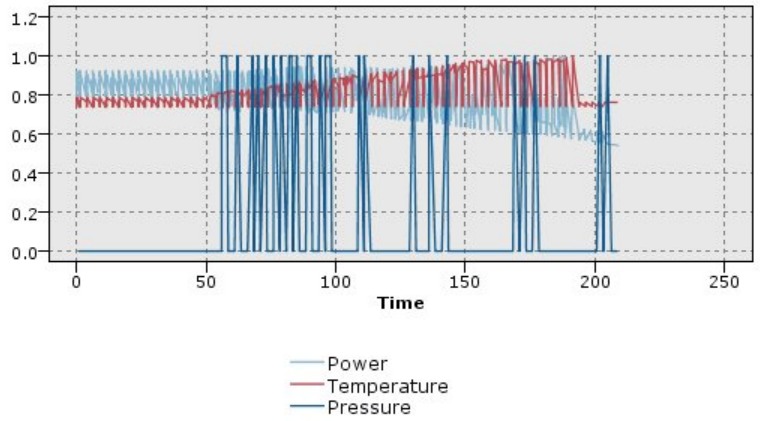


Рисунок 25. Нормализованный график нескольких зависимостей, показывающий зависимость для давления

Функция наложения. Выберите эту опцию, чтобы задать известную функцию для сравнения фактических значений. Например, чтобы сравнить фактические значения с предсказанными, на график можно нанести в качестве наложения функцию `y = x`. Задайте в текстовом окне функцию для **`y =`**. По умолчанию используется функция `y = x`, но можно функцию любого другого типа, например, квадратичную функцию или произвольное выражение в значениях `x`.

Примечание: Функции наложения для панельной и анимационной диаграммы недоступны.

Когда число записей больше. Задайте метод нанесения на диаграмму больших наборов данных. Можно задать максимальный размер набора данных или использовать размер по умолчанию 2000 точек. Для больших наборов данных производительность повысится, если выбрать опцию **Интервал** или **Выборка**. Другой вариант - построение графика с использованием всех точек, задав опцию **Использовать все данные**, но нужно иметь в виду, что это может сильно снизить производительность программы.

Примечание: Если задан Режим X **Наложение** или **Как читается**, эти опции отключаются, и используются только первые *n* записей.

- **Ящик.** Выберите эту опцию, чтобы включать разделение на интервалы, когда набор данных содержит больше записей, чем заданное число. При разделении на интервалы диаграмма перед фактическим построением графика разбивается на мелкую сетку, и подсчитывается число соединений, которые будут выводиться в каждой из ячеек сетки. На окончательной диаграмме будет использоваться по одному соединению для каждой ячейки в центроиде интервала (среднем от всех точек соединений в интервале).
- **Пример.** Выберите эту опцию для случайной выборки данных в соответствии с заданным числом записей.

Вкладка Внешний вид графика множественных зависимостей

Перед созданием графиков можно задать опции внешнего вида.

Заголовок. Введите текст для использования в качестве заголовка диаграммы.

Подзаголовок. Введите текст для использования в качестве подзаголовка диаграммы.

Подпись. Введите текст для использования в качестве подписи диаграммы.

Метка X. Либо примите метку оси *x* (горизонтальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Метка Y. Либо примите метку оси *y* (вертикальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Показать сетку. Эта опция, включаемая по умолчанию, выводит позади графика сетку, упрощающую определение точек отсечения для регионов и полос. Линии сетки всегда выводятся белым цветом, если только не используется белый фон диаграммы; в этом случае они выводятся серым цветом.

Использование диаграммы нескольких зависимостей

Простые графики и графики множественных зависимостей - это по существу зависимости, построенные по координатам *X* и *Y*. Например, если вы изучаете возможные мошенничества в заявках на получение сельскохозяйственных субсидий, возможно, вы захотите построить график заявленной прибыли относительно оценки прибыли при помощи нейросети. Применение наложения (например, для категории культуры), покажет, есть ли взаимосвязь между заявленными свойствами (полезностью или количеством) и категорией культуры.

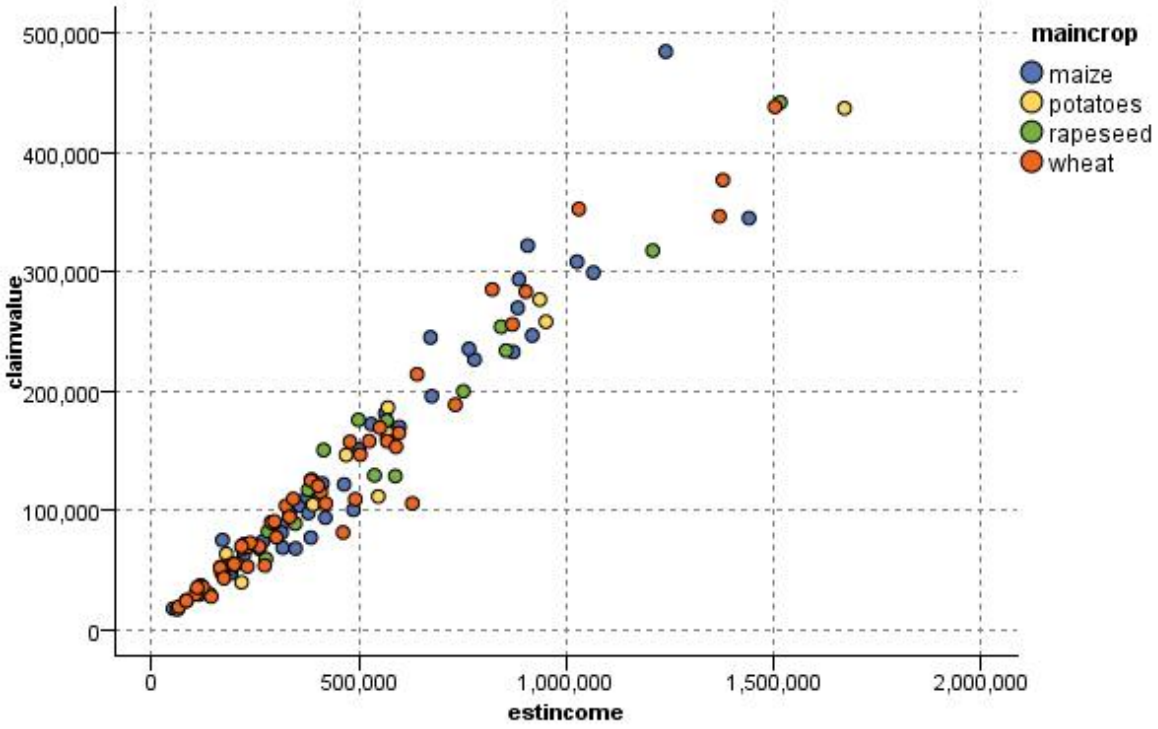


Рисунок 26. График взаимосвязи между оцененной прибылью и заявленной полезностью с категорией ведущей культуры в качестве наложения

Поскольку простые графики, графики нескольких зависимостей и графики оценок - это двумерные представления, выводимые по осям *Y* и *X*, при работе с ними можно легко определить регионы, пометить элементы или даже нарисовать полосы. Можно также сгенерировать узлы для данных, представленных этими регионами, полосами или элементами. Дополнительную информацию смотрите в разделе “Исследование графиков” на стр. 260.

Узел График зависимости от времени

Узлы График зависимости от времени позволяют просматривать один или несколько временных рядов, представленных на графике в зависимости от времени. Изображаемые на графике ряды должны содержать численные значения, и предполагается, что они определены в промежутке времени, в котором есть определенные периоды. Обычно до узла График зависимости от времени используется узел Интервалы времени, чтобы создать поле *TimeLabel*, по умолчанию используемое для отметки оси *x* на графиках. Дополнительную информацию смотрите в разделе Узел интервалов времени (объявлен устаревшим).

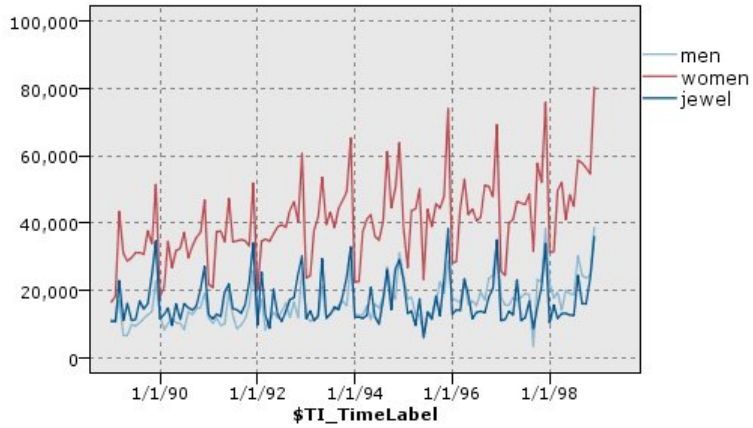


Рисунок 27. Вывод графика продаж мужской и женской одежды и украшений

Создание интервенций и событий

Из графика с зависимостью от времени можно создать поля Событие и Интервенция, сгенерировав в контекстном меню производный узел (флаговый или номинальный). Например, можно создать поле Событие на случай забастовки на железной дороге, значение которого будет True, если такое событие произошло, или False в ином случае. Для поля Интервенция (например, рост цен) можно использовать производный отсчет для обозначения даты увеличения, так что 0 соответствует старой цене, а 1 - новой. Дополнительную информацию смотрите в разделе “Узел извлечения” на стр. 144.

Вкладка графика временной зависимости

График. Предоставляет выбор способа нанесения на график данных временного ряда.

- **Выбранные ряды.** Наносит на график значения для выбранного временного ряда. Если вы выбрали эту опцию при построении графика доверительных интервалов, выключите переключатель **Нормализовать**.
- **Выбранные модели временных рядов**. Используемая в сочетании с моделью временного ряда, эта опция наносит на график все связанные поля (фактические и предсказанные значения, а также доверительные интервалы) для одного или нескольких выбранных временных рядов. Эта опция отключает некоторые другие опции в диалоговом окне. Это опция рекомендуется в случае нанесения на график доверительных интервалов.

Ряды. Выберите одно или несколько полей с данными временного ряда, которые вы хотите нанести на график. Данные должны быть числовыми.

Метка оси X. Выберите либо метку по умолчанию, либо одно поле для использование в качестве метки для оси *x* на графиках. При выборе опции По умолчанию система будет использовать поле TimeLabel, созданное из восходящего потока узла интервалов времени, или последовательные целочисленные значения, если узел интервалов времени отсутствует. Дополнительную информацию смотрите в разделе Узел интервалов времени (объявлен устаревшим).

Выводить ряды на отдельных панелях. Указывает, следует ли выводить каждый ряд на отдельной панели. Другой вариант - если не выбрать опцию панелей, все временные ряды будут нанесены на одну и ту же диаграмму, а сглаживатели будут недоступны. При построении графиков всех временных рядов на одной и той же диаграмме каждый ряд будет представлен различным цветом.

Нормализовать. Выберите эту опцию для масштабирования всех значений *Y* с переводом в диапазон 0–1 для вывода на диаграмме. Нормализация помогает изучить взаимосвязь между линиями, которые в противном случае были бы скрыты из-за различий в диапазонах значений для каждого ряда, и рекомендуется при

нанесении на одну и ту же диаграмму нескольких зависимостей или при сравнении графиков на примыкающих друг к другу панелях. (Если все значения данных попадают в один и тот же диапазон, нормализация не требуется.)

Вывод. Выберите один или несколько элементов для вывода на графике. Для выбора предлагаются линии, точки и сглаживатели (LOESS). Сглаживатели доступны, только если выбран вывод рядов на отдельных панелях. По умолчанию в качестве элемента выбирается линия. Перед вызовом узла диаграммы убедитесь, что выбран хотя бы один элемент графика; в противном случае система возвратит сообщение об ошибке, указывающее, что для нанесения на график ничего не выбрано.

Ограничить записи. Выберите эту опцию, если хотите ограничить число наносимых на график записей. Задайте в опции **Максимальное число записей к графику** число записей (читаемых с начала вашего файла данных), которые будут нанесены на график. По умолчанию в качестве этого числа задается 2000. Если вы хотите нанести на график последние *n* записей файла данных, вызовите перед этим узлом узел сортировки, чтобы упорядочить записи в возрастающем порядке по времени.

Вкладка Внешний вид графика временной зависимости

Перед созданием графиков можно задать опции внешнего вида.

Заголовок. Введите текст для использования в качестве заголовка диаграммы.

Подзаголовок. Введите текст для использования в качестве подзаголовка диаграммы.

Подпись. Введите текст для использования в качестве подписи диаграммы.

Метка X. Либо примите метку оси *x* (горизонтальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Метка Y. Либо примите метку оси *y* (вертикальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Показать сетку. Эта опция, включаемая по умолчанию, выводит позади графика сетку, упрощающую определение точек отсечения для регионов и полос. Линии сетки всегда выводятся белым цветом, если только не используется белый фон диаграммы; в этом случае они выводятся серым цветом.

Разметка . Только для графиков зависимостей от времени. Можно указать, выводить ли на графике значения времени по горизонтальной или по вертикальной оси.

Использование графика временной зависимости

После создания диаграммы графика временной зависимости появляются две опции для корректировки вывода этой диаграммы и генерирования узлов для дальнейшего анализа. Дополнительную информацию смотрите в разделе “Исследование графиков” на стр. 260.

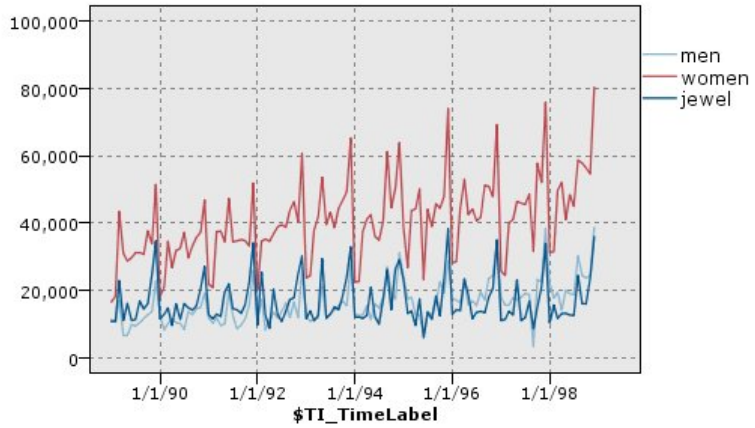


Рисунок 28. Вывод графика продаж мужской и женской одежды и украшений

Создав график временной зависимости, определив полосы и изучив результаты, вы можете, используя опции в меню Создать и контекстном меню, создать узлы выбора или извлечения. Дополнительную информацию смотрите в разделе “Генерирование узлов из диаграмм” на стр. 267.

Узел Распределение

Диаграмма распределения или таблица показывают появление в наборе данных символических (не числовых) значений, таких как тип закладной или пол. Типичное использование узла Распределение состоит в демонстрации дисбаланса данных, который может быть выправлен до создания модели с помощью узла Баланс. Узел Баланс можно сгенерировать автоматически при помощи меню Создать на диаграмме распределения или в окне таблицы.

Можно использовать также узел Graphboard (панель выбора диаграмм) для создания столбчатых диаграмм или диаграмм отсчетов. Однако в этом узле можно выбрать и другие опции. Дополнительную информацию смотрите в разделе “Доступные встроенные типы визуализации Панели выбора диаграмм” на стр. 190.

Примечание: Для вывода вхождений числовых значений необходимо использовать узел Гистограмма.

Вкладка График распределения

График. Выберите тип распределения. Выберите **Выбранные поля**, чтобы вывести распределение выбранного поля. Выберите **Все флаги (значения true)**, чтобы вывести распределение значений true для полей флагов в наборе данных.

Поле. Выберите номинальное или флаговое поле, для которого следует вывести распределение значений. В списке появятся только поля, которые не были определены явным образом как числовые.

Наложение. Выберите номинальное или флаговое поле, чтобы использовать его как наложение цвета, иллюстрирующего распределение значений в каждом значении заданного поля. Например, ответ на маркетинговую кампанию (*pep*) может использоваться как наложение для числа детей (*children*) в качестве иллюстрации реагирования, определяемого размером семьи. Дополнительную информацию смотрите в разделе “Эстетики, наложения, панели и анимация” на стр. 180.

Нормализовать по цвету. Выберите эту опцию для масштабирования столбиков так, чтобы в совокупности они занимали всю ширину диаграммы. Значения наложения приравниваются к соотношению для каждого столбика, что упрощает операции сравнения по категориям.

Сортировка. Выберите метод, используемый для вывода значений на диаграмме распределения. Выберите **По алфавиту**, чтобы использовать алфавитный порядок, или **По количеству**, чтобы вывести список значений по убыванию встречаемости.

Пропорциональная шкала. Выберите эту опцию для масштабирования значений так, чтобы значение с наибольшим количеством заполняло всю ширину графика. Все остальные столбики будут масштабироваться по этому значению. При отключении этой опции столбики будут масштабироваться в соответствии с общим количеством каждого значения.

Вкладка Внешний вид распределения

Перед созданием графиков можно задать опции внешнего вида.

Заголовок. Введите текст для использования в качестве заголовка диаграммы.

Подзаголовок. Введите текст для использования в качестве подзаголовка диаграммы.

Подпись. Введите текст для использования в качестве подписи диаграммы.

Метка X. Либо примите метку оси *x* (горизонтальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Метка Y. Либо примите метку оси *y* (вертикальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Показать сетку. Эта опция, включаемая по умолчанию, выводит позади графика сетку, упрощающую определение точек отсечения для регионов и полос. Линии сетки всегда выводятся белым цветом, если только не используется белый фон диаграммы; в этом случае они выводятся серым цветом.

Использование узла распределения

Узлы распределения используются для показа распределения символических значений в наборе данных. Часто они используются перед использованием узлов преобразования для исследования данных и исправления всевозможных дисбалансов. Например, если экземпляры респондентов без детей встречаются чаще, чем респондентов других типов, возможно, вы захотите сократить число этих экземпляров, чтобы можно было сгенерировать более полезные правила при дальнейших операциях исследования данных. Узел распределения поможет вам исследовать дисбалансы такого рода и решить, что с ними делать.

Узел распределения необычен тем, что генерирует и диаграмму, и таблицу для анализа данных.

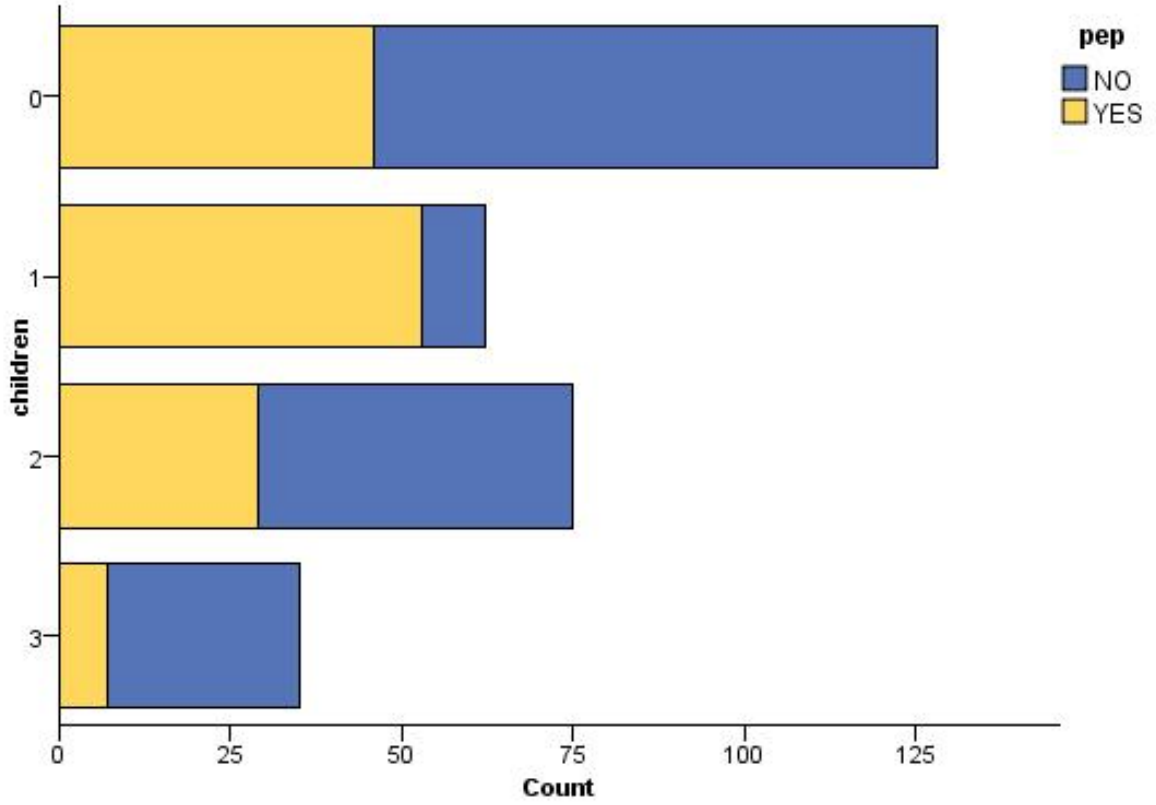


Рисунок 29. Диаграмма распределения, показывающая число людей с детьми или без детей, ответивших на маркетинговую компанию

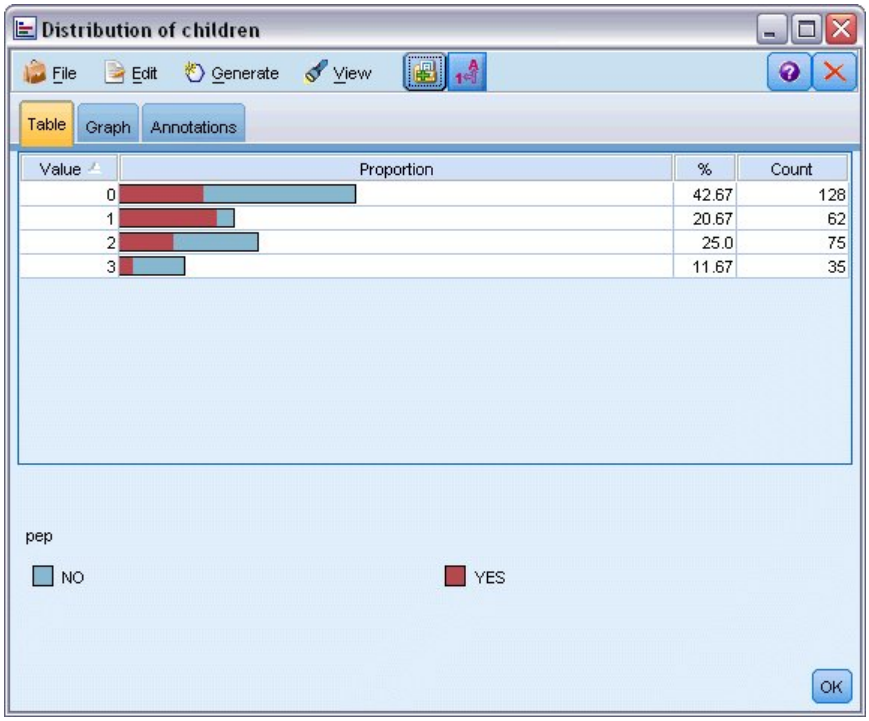


Рисунок 30. Таблица распределения, показывающая соотношение людей с детьми или без детей, ответивших на маркетинговую компанию

Создав диаграмму и таблицу распределения и изучив результаты, вы можете при помощи опций разных меню сгруппировать значения, скопировать их и сгенерировать ряд узлов для подготовки данных. Кроме того, информацию диаграммы и таблицы можно скопировать или экспортировать для использования в других прикладных программах, таких как MS Word или MS PowerPoint. Дополнительную информацию смотрите в разделе “Печать, сохранение, копирование и экспорт диаграмм” на стр. 283.

Чтобы выбрать и скопировать значения из таблицы распределения

1. Нажмите и удерживайте кнопку мыши, перемещая указатель по строкам, чтобы выбрать набор значений. При помощи меню Правка можно также **Выбрать все** значения.
2. В меню Правка выберите **Копировать таблицу** или **Копировать таблицу (включая имена полей)**.
3. Вставьте содержимое буфера обмена в нужную прикладную программу.

   Примечание: Столбики не копируются непосредственно. Вместо них копируются значения таблицы. Это подразумевает, что значения наложений в скопированной таблице выводиться не будут.

Чтобы сгруппировать значения из таблицы распределения

1. Выберите значения кнопкой мыши при нажатой клавише Ctrl.
2. В меню Правка выберите **Группа**.

Примечание: При формировании и расформировании групп значений диаграмма на вкладке Диаграмма автоматически перерисовывается с целью вывода изменений.

Кроме того, можно:

- Разгруппировать значения, выбрав имя группы в списке распределения и опцию **Снять группировку** в меню Правка.
- Отредактировать группы, выбрав имя группы в списке распределения и опцию **Изменить группу** в меню Правка. Эта опция открывает диалоговое окно, где можно перемещать значения в группу и из группы.

Опции меню Создать

С помощью опций меню Создать можно выбрать поднабор данных, получить поле флага, перегруппировать значения, переклассифицировать их или сбалансировать данные из диаграммы или таблицы. Эти операции генерируют узел подготовки данных и помещают его на холст потока. Для использования сгенерированного узла соедините его с существующим потоком. Дополнительную информацию смотрите в разделе “Генерирование узлов из диаграмм” на стр. 267.

Узел Гистограмма

Узлы Гистограмма показывают существующие значения для числовых полей. Они часто используются для исследования данных перед их обработкой и построением моделей. Аналогично узлу Распределение, узлы Гистограмма часто используются для устранения дисбаланса данных. Хотя для построения гистограмм можно использовать и узел Graphboard (панель выбора диаграмм), на этом узле есть больше опций для выбора. Дополнительную информацию смотрите в разделе “Доступные встроенные типы визуализации Панели выбора диаграмм” на стр. 190.

Замечание: Для вывода существующих значений символических полей необходимо использовать узел Распределение.

Вкладка График гистограммы

Поле. Выберите числовое поле, для которого следует вывести распределение значений. В списке будут указаны, которые не были определены явно как символические (категориальные).

Наложение. Выберите символическое поле для вывода категорий значений для заданного поля. При выборе поля наложения гистограмма преобразуется в накопительскую диаграмму с различным цветом, используемым для представления разных категорий поля наложения. Если используется узел гистограммы, предоставляется три типа наложений: цвет, панель и анимация. Дополнительную информацию смотрите в разделе “Эстетики, наложения, панели и анимация” на стр. 180.

Вкладка Опции гистограммы

Автоматический диапазон по X. Выберите эту опцию, чтобы использовать весь диапазон значений данных по этой оси. Выключите этот переключатель, чтобы использовать явный поднабор значений на основе заданных вами значений **Мин** и **Макс**. Введите значение или задайте его кнопками со стрелками. По умолчанию выбраны автоматические диапазоны для быстрого построения диаграмм.

Интервалы. Выберите либо **По числу**, либо **По ширине**.

- Выберите **По числу** для вывода фиксированного числа столбиков, ширина которых зависит от диапазона, а число интервалов задано. Укажите число интервалов для использования на диаграмме в опции **Число интервалов**. Для настройки этого числа используйте кнопки со стрелками.
- Выберите **По ширине**, чтобы создать диаграмму со столбиками фиксированной ширины. Число интервалов будет зависеть от заданной ширины и диапазона значений. Укажите ширину столбиков в опции **Ширина интервала**.

Нормализовать по цвету. Выберите эту опцию, чтобы скорректировать все столбики, задав для них один тот же вес, с выводом значений наложений в каждом столбике как процента от общего числа наблюдений.

Показать нормальную кривую. Выберите эту опцию для добавления на диаграмму нормальной кривой, показывающей среднее и дисперсию данных.

Разделить полосы для каждого цвета. Выберите эту опцию, для вывода на диаграмме значения оверлея в виде отдельной полосы.

Вкладка Внешний вид гистограммы

Перед созданием графиков можно задать опции внешнего вида.

Заголовок. Введите текст для использования в качестве заголовка диаграммы.

Подзаголовок. Введите текст для использования в качестве подзаголовка диаграммы.

Подпись. Введите текст для использования в качестве подписи диаграммы.

Метка X. Либо примите метку оси *x* (горизонтальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Метка Y. Либо примите метку оси *y* (вертикальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Показать сетку. Эта опция, включаемая по умолчанию, выводит позади графика сетку, упрощающую определение точек отсечения для регионов и полос. Линии сетки всегда выводятся белым цветом, если только не используется белый фон диаграммы; в этом случае они выводятся серым цветом.

Использование гистограмм

Гистограммы показывают распределение значений в числовом поле, диапазон значений которого представлен по оси *x*. Гистограммы работают подобно диаграммам собраний. Собрания показывают распределение значений для одного числового поля *относительно значений другого*, а не распределение значений для одного поля.

Создав диаграмму, можно проверить результаты и определить полосы для разбиения значений по оси *x* или определить регионы. Кроме того, можно пометить элементы на диаграмме. Дополнительную информацию смотрите в разделе “Исследование графиков” на стр. 260.

Опции в меню Создать позволяют создать узлы балансировки, выбора или извлечения при помощи данных на диаграмме, а именно в полосах, регионах или помеченных элементах. Диаграммы этого типа часто используются перед использованием узлов преобразования для исследования данных и исправления возможных дисбалансов посредством генерирования узла балансировки с диаграммы для использования в потоке. Можно также сгенерировать узел извлечения, чтобы добавить поле, показывающее, в какую полосу попадает каждая запись, или узел выбора, чтобы выбрать все записи в конкретном наборе или диапазоне значений. Такие операции помогают сосредоточиться на конкретном поднаборе данных для дальнейшего исследования. Дополнительную информацию смотрите в разделе “Генерирование узлов из диаграмм” на стр. 267.

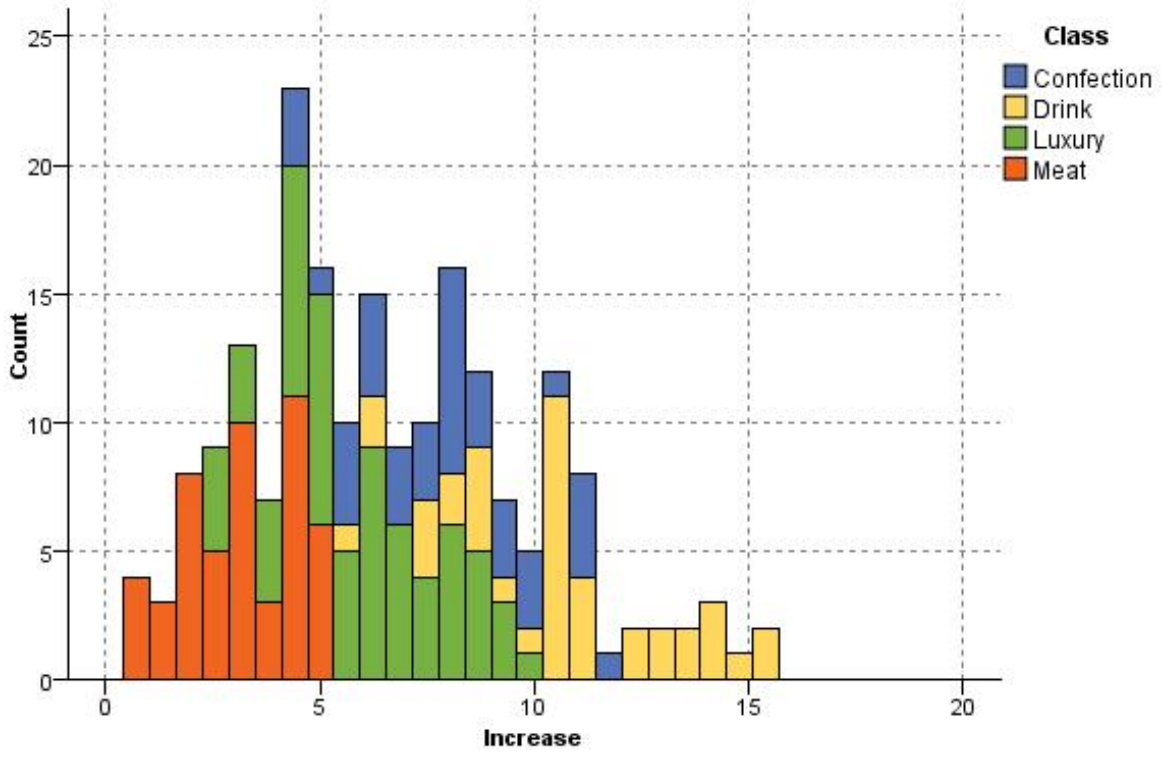


Рисунок 31. Гистограмма, показывающая распределение возросшего количества покупок по категориям из-за рекламной кампании

Узел Собрание

Собрания аналогичны гистограммам, однако собрания показывают распределение значений для одного числового поля относительно значений другого, а не распределение значений для одного поля. Собрание полезно для иллюстрации переменной или значений в поле, изменяющихся во времени. Используя 3D-представление, вы можете включить также символическую ось, показывающую распределения по категориям. Двумерные собрания показываются как составные столбчатые диаграммы с накоплением. Дополнительную информацию смотрите в разделе “Эстетики, наложения, панели и анимация” на стр. 180.

Вкладка График собрания

Сбор. Выберите поле, значения которого будут собраны и выведены по отношению к диапазону значений для поля, заданного в опции **Относительно**. В списке указываются только поля, которые не были определены как символические.

Относительно. Выберите поле, значения которого будут использоваться для вывода поля, заданного в опции **Сбор**.

По. Эта опция, включаемая при создании трехмерной диаграммы, позволяет выбрать номинальное или флаговое поле, используемое для вывода поля собрания по категориям.

Операция. Выберите, что будет представлять каждый столбик на диаграмме собрания. В состав опций входят **Сумма**, **Среднее**, **Максимум**, **Минимум** и **Среднеквадратичное отклонение**.

Наложение. Выберите символическое поле для вывода категорий значений для выбранного поля. Выбор поля наложения преобразует собрание и создает несколько столбиков различного цвета для каждой категории. У этого узла три типа наложений: цвет, панель и анимация. Дополнительную информацию смотрите в разделе “Эстетики, наложения, панели и анимация” на стр. 180.

Вкладка Опции собрания

Автоматический диапазон по X. Выберите эту опцию, чтобы использовать весь диапазон значений данных по этой оси. Выключите этот переключатель, чтобы использовать явный поднабор значений на основе заданных вами значений **Мин** и **Макс**. Введите значение или задайте его кнопками со стрелками. По умолчанию выбраны автоматические диапазоны для быстрого построения диаграмм.

Интервалы. Выберите либо **По числу**, либо **По ширине**.

- Выберите **По числу** для вывода фиксированного числа столбиков, ширина которых зависит от диапазона, а число интервалов задано. Укажите число интервалов для использования на диаграмме в опции **Число интервалов**. Для настройки этого числа используйте кнопки со стрелками.
- Выберите **По ширине**, чтобы создать диаграмму со столбиками фиксированной ширины. Число интервалов будет зависеть от заданной ширины и диапазона значений. Укажите ширину столбиков в опции **Ширина интервала**.

Вкладка Внешний вид собрания

Перед созданием графиков можно задать опции внешнего вида.

Заголовок. Введите текст для использования в качестве заголовка диаграммы.

Подзаголовок. Введите текст для использования в качестве подзаголовка диаграммы.

Подпись. Введите текст для использования в качестве подписи диаграммы.

Метка по факту. Либо примите метку, сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Метка собрания. Либо примите метку, сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Метка подписи. Либо примите метку, сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Показать сетку. Эта опция, включаемая по умолчанию, выводит позади графика сетку, упрощающую определение точек отсечения для регионов и полос. Линии сетки всегда выводятся белым цветом, если только не используется белый фон диаграммы; в этом случае они выводятся серым цветом.

Следующий пример показывает, где на трехмерном графике размещаются опции внешнего вида.

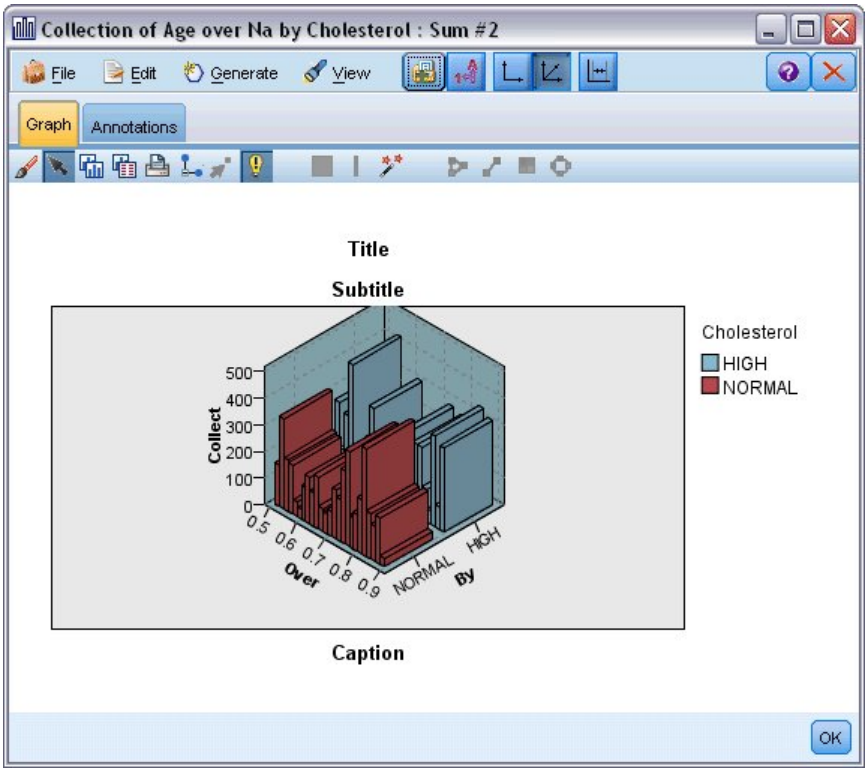


Рисунок 32. Размещение опций внешнего вида диаграмм на трехмерной диаграмме

Использование диаграммы собрания

Собрания показывают распределение значений для одного числового поля *относительно значений другого*, а не распределение значений для одного поля. Гистограммы работают подобно диаграммам собраний. Гистограммы показывают распределение значений в числовом поле, диапазон значений которого представлен по оси *x*.

Создав диаграмму, можно проверить результаты и определить полосы для разбиения значений по оси *x* или определить регионы. Кроме того, можно пометить элементы на диаграмме. Дополнительную информацию смотрите в разделе “Исследование графиков” на стр. 260.

Опции в меню Создать позволяют создать узлы балансировки, выбора или извлечения при помощи данных на диаграмме, а именно в полосах, регионах или помеченных элементах. Диаграммы этого типа часто используются перед использованием узлов преобразования для исследования данных и исправления возможных дисбалансов посредством генерирования узла балансировки с диаграммы для использования в потоке. Можно также сгенерировать узел извлечения, чтобы добавить поле, показывающее, в какую полосу попадает каждая запись, или узел выбора, чтобы выбрать все записи в конкретном наборе или диапазоне значений. Такие операции помогают сосредоточиться на конкретном поднаборе данных для дальнейшего исследования. Дополнительную информацию смотрите в разделе “Генерирование узлов из диаграмм” на стр. 267.

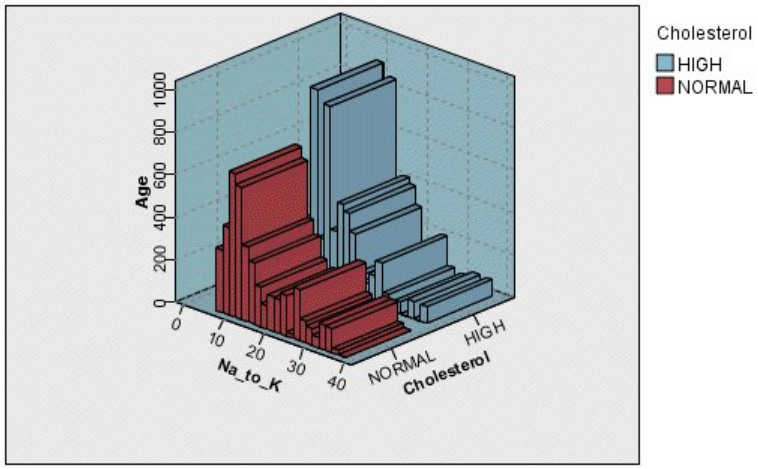


Рисунок 33. Трехмерная диаграмма собрания, показывающая суммарное Na/K по отношению к возрасту для высокого и нормального уровней холестерина

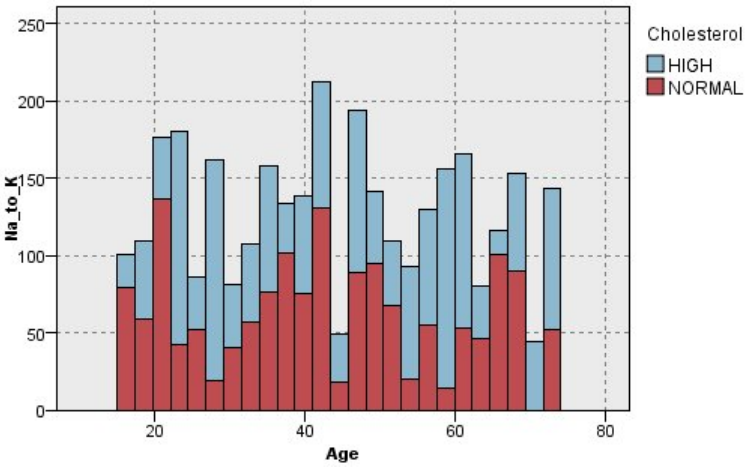


Рисунок 34. Диаграмма собрания, выводимая без оси z, но с выделенным цветом уровнем холестерина

Узел Web

Узлы Web показывают силу взаимосвязи между значениями двух или более символических полей. Такой график выводит соединения при помощи изменения типа линий для обозначения силы соединения. Узел Web можно использовать, например, для исследования взаимосвязи между покупкой различных товаров через Интернет и в традиционных точках продажи.

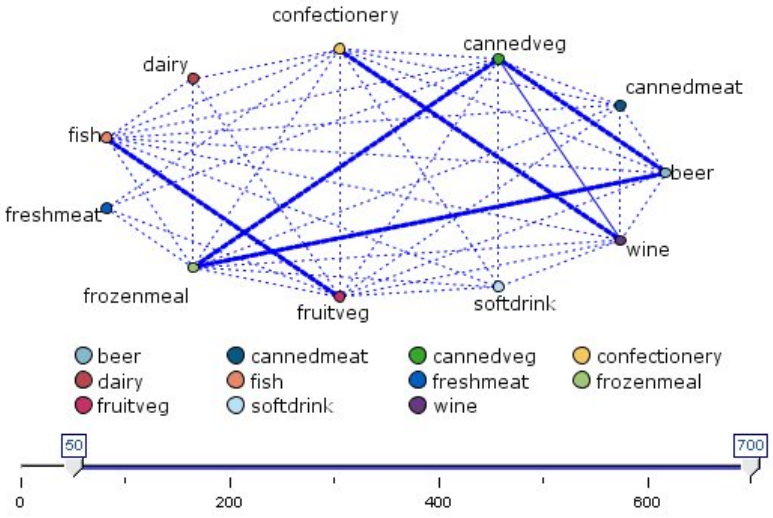


Рисунок 35. График Web, показывающий взаимосвязь между покупками бакалейных товаров

Узлы Directed Web

Узлы Directed Web аналогичны узлам Web, так как они показывают степень взаимосвязи между значениями символьных полей. Однако графики Directed Web показывают значения только одного или нескольких полей From с единственным полем To. Эти связи однонаправленные, то есть они определены только в одну сторону.

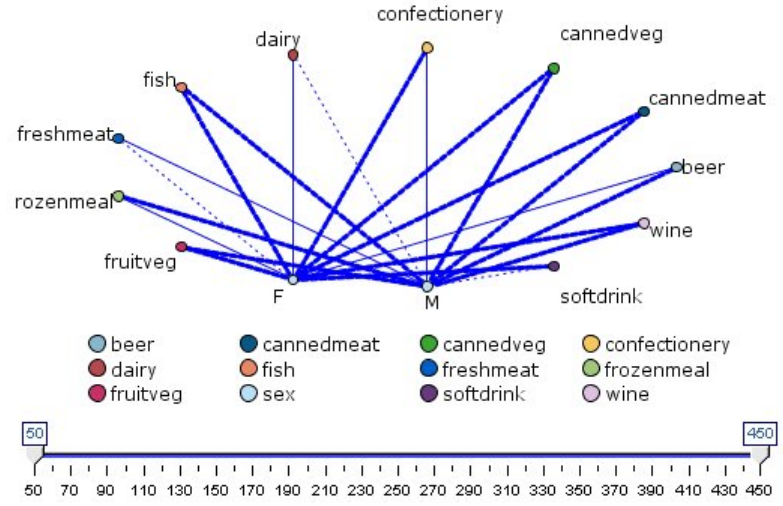


Рисунок 36. График Directed Web, показывающий взаимосвязь между покупкой бакалейных товаров и полом покупателя

Как и узлы Web, этот график показывает связи при помощи изменения типа линий для обозначения силы соединения. Узел Directed Web можно использовать, например, для изучения взаимосвязи между полом и склонностью к каким-либо покупкам.

Вкладка Сетевой граф

Сеть. Выберите для создания сетевого графа, иллюстрирующего силу взаимосвязей между всеми заданными полями.

Направленная сеть. Выберите для создания направленного сетевого графа, иллюстрирующего силу взаимосвязей между несколькими полями и значениями одного поля, например, гендерной и религиозной принадлежности. Когда выбрана эта опция, активируется вариант В поле и управляющий элемент Поля ниже переименовывается в Из полей для дополнительной ясности.

В поле (только направленный граф). Выбрать флаговое или номинальное поле, используемое для направленного графа. Перечисляются только поля, которые не были явно заданы как числовые.

Поля/Из полей. Выбрать поля для создания сетевого графа. Только поля, которые не были явно заданы как числовые. Используйте кнопку Средство выбора полей для выбора нескольких полей или полей по типу.

Примечание: Для направленного графа этот элемент управления используется, чтобы выбрать опцию Из полей.

Показывать только флаги true. Выбрать показ для флаговых полей только флагов true. Эта опция упрощает вывод графа и часто используется для данных, в которых появление положительных значений особенно важно.

Значения линий. Выберите тип порога из выпадающего списка.

- **Абсолютный** задает пороги на основе числа записей, в которых есть каждая пара значений.
- При опции **Общие проценты** показывается абсолютное число наблюдений, представленных связью, как долю всех появлений каждой пары значений, представленных в сетевом графе.
- **Процентные доля редкой пары поле/значение** и **Процентные доли частой пары поле/значение** обозначают, какие пары поле/значение использовать для оценки процентных долей. Например, предположим, что у 100 записей есть значение *drugY* для поля *Препарат* и только у десяти записей значение *НИЗКИЙ* для поля *BP*. Если у семи записей есть оба значения *drugY* и *НИЗКИЙ*, эта процентная доля равна 70% или 7%, в зависимости от того, на какое поле ссылаются, редкое (*BP*) или частое (*Препарат*).

Примечание: Для направленных сетевых графов третья и четвертая опции из указанных недоступны. Вместо этого вы можете выбрать **Процентная доля пар поле/значение типа "To"** и **Процентная доля пар поле/значение типа "From"**.

Сильные связи тяжелее. Выбирается по умолчанию, это стандартный способ показа связей между полями.

Слабые связи тяжелее. Выбрать для обращения смысла связей, показываемых жирными линиями. Эта опция часто используется для обнаружения мошенничества или исследования выбросов.

Вкладка Опции сетевых графов

Вкладка Опции для узлов Web содержит несколько дополнительных опций для настройки выходных графов.

Количество связей. Следующие опции используются для управления количеством связей, показываемых на выходном графе. Некоторые из этих опций, такие как **Слабые связи выше** и **Сильные связи выше**, доступны также в окне вывода графов. Для итогового графа можно использовать также ползунок, чтобы настроить количество выводимых связей.

- **Максимальное количество выводимых связей.** Задать число, обозначающее максимальное количество связей для показа на выходном графе. Для настройки этого значения используйте кнопки со стрелками.
- **Показать только связи выше.** Задать число, обозначающее минимальное значение, для которого будет показываться соединение в графе. Для настройки этого значения используйте кнопки со стрелками.
- **Показать все связи.** Задать для вывода всех связей независимо от минимальных и максимальных значений. Выбор этой опции может увеличить время обработки, если существует много полей.

Отклонять, если записей очень мало. Выбрать, чтобы игнорировать соединения, которые поддерживаются слишком малым числом записей. Задайте порог для этой опции, введя число в **Минимум записей на строку**.

Отклонять, если записей очень много. Выбрать, чтобы игнорировать сильно поддерживаемые соединения. Введите число в поле **Максимум записей на строку**.

Слабы связи ниже. Задать число, обозначающее порог для слабых (пунктирные линии) и обычных (простые линии) соединений. Все соединения со значением ниже рассматриваются как слабые.

Сильные связи выше. Задать число, обозначающее порог для сильных (жирные линии) и обычных (простые линии) соединений. Все соединения со значением выше рассматриваются как сильные.

Размер связи. Задать опции для управления размером связей:

- **Размер связи меняется непрерывно.** Выбрать для вывода диапазона размеров связей, отображающего различия в силе соединений на основе фактических значений данных.
- **Размер связи показывает сильные/нормальные/слабые категории.** Выбрать для вывода трех значений для силы соединения - сильного, нормального и слабого. Точки отсечения для этих категорий можно задать выше, а также на финальном графике.

Вывод в граф. Выбрать тип вывода в граф:

- **Круговой макет.** Выбрать для использования стандартного вывода в граф.
- **Сетевой макет.** Выбрать для использования алгоритма группировки совместно самых сильных связей. Одна предназначена для выделения сильных связей при помощи пространственного разделения и взвешенных линий.
- **Направленный макет.** Выбрать для создания направленного вывода в граф, который использует выбор **В поле** на вкладке График как фокус для направления.
- **Макет с сеткой.** Выбрать для создания вывода в граф, который раскладывается по равномерно распределенной сетке.

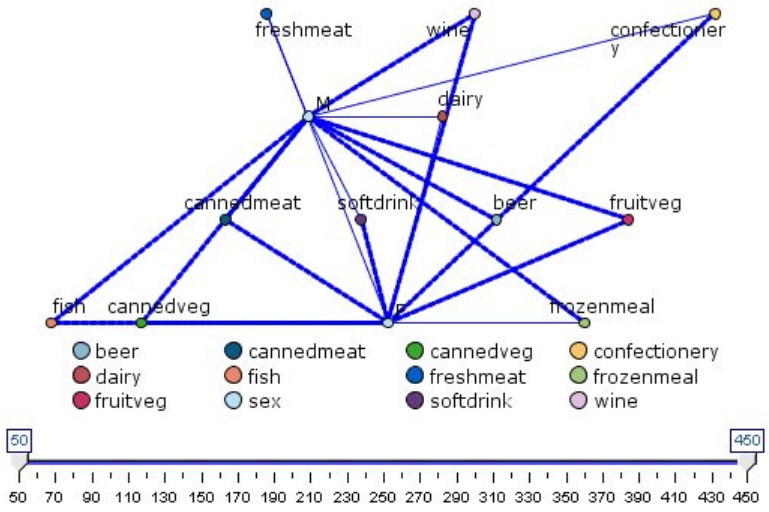


Рисунок 37. Граф, показывающий сильные соединения элементов frozenmeal и cannedveg с другими элементами бакалеи

Вкладка Внешний вид в Web

Перед созданием графиков можно задать опции внешнего вида.

Заголовок. Введите текст для использования в качестве заголовка диаграммы.

Подзаголовок. Введите текст для использования в качестве подзаголовка диаграммы.

Подпись. Введите текст для использования в качестве подписи диаграммы.

Показать пояснение. Можно указать, выводить ли легенду. Для графиков с большим числом полей легенда может улучшить внешний вид графика.

Использовать метки как узлы. Можно включить текст меток в каждый узел вместо того, чтобы выводить соседние метки. Для графиков с небольшим числом полей это может улучшить удобочитаемость диаграммы.

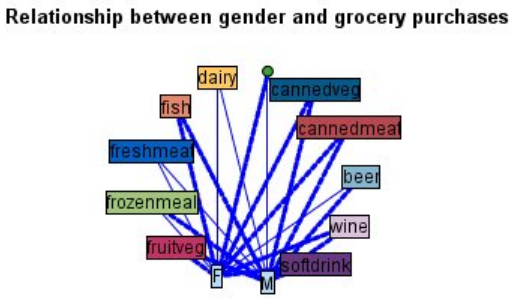


Рисунок 38. Диаграмма в Web, где показаны метки как узлы

Использование графа

Узлы Web показывают силу взаимосвязи между значениями двух или более символических полей. Соединения выводятся на графе различным типом линий, чтобы обозначить соединения с возрастающей силой. Узел Web можно использовать, например, для изучения взаимосвязи между уровнем холестерина, кровяным давлением и препаратом, оказавшимся эффективным при лечении пациента.

- Сильные соединения показаны жирными линиями. Это обозначает, что два значения сильно связаны и должны изучаться подробнее.
- Средние соединения показаны линиями нормальной толщины.
- Слабые соединения показаны пунктирной линией.
- Если между двумя значениями линии нет, это означает или тот факт, что данные два значения никогда не встречались в одной записи, или что эта комбинация принадлежит записям, количество которых ниже порога, заданного в диалоговом окне узла Web.

После создания узла Web есть несколько опций для настройки вывода графиков и генерирования узлов для дальнейшего анализа.

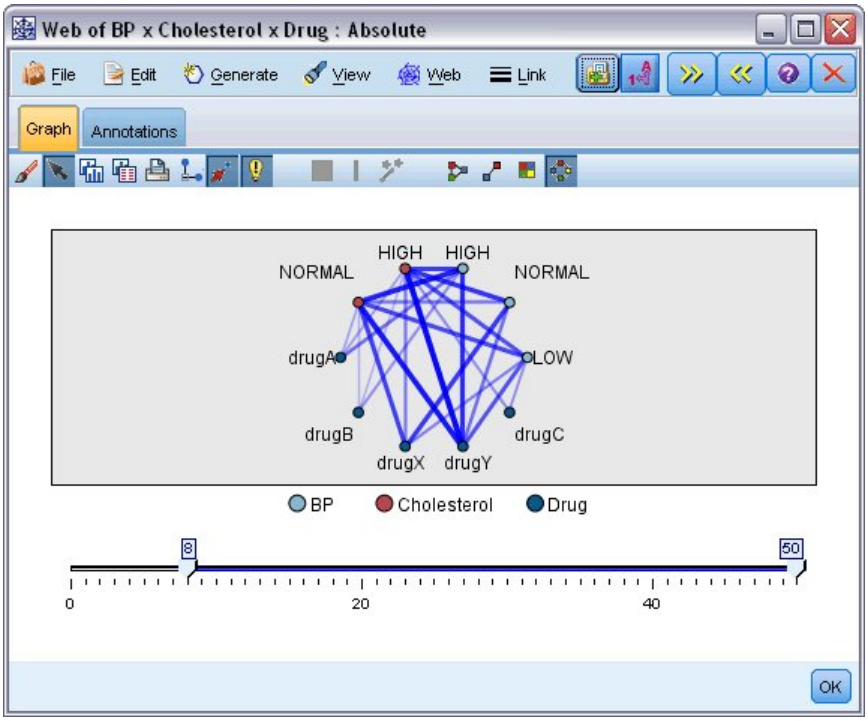


Рисунок 39. Граф, обозначающий количество сильных взаимосвязей, таких как кровяного давления с DrugX и уровня холестерина с DrugY

И для узлов Web, и для направленных узлов Web вы можете:

- Изменить макет вывода графа.
- Скрыть точки для упрощения вывода.
- Изменить пороги, управляющие стилем линий.
- Выделить линии между значениями для обозначения "выбранной" взаимосвязи.
- Сгенерировать узел Выбор для одной или нескольких "выбранных" записей или узел Флаг извлечения, связанной с одной или несколькими взаимосвязями графа.

Скорректировать точки

- **Переместить** точки, щелкнув кнопкой мыши по точке и перетащив ее в новое положение. Граф будет перерисован для отображения нового положения.
- **Скрыть** точки, щелкнув правой кнопкой мыши по точке графа и выбрав опцию **Скрыть** или **Скрыть и перепланировать** в контекстном меню. Опция **Скрыть** просто скрывает выбранную точку и все связанные с ней линии. При опции **Скрыть и перепланировать** граф перерисовывается с учетом всех внесенных вами изменений. Все перемещения вручную отменяются.
- **Показать** все скрытые точки, выбрав **Воспроизвести все** или **Воспроизвести все и перепланировать** в меню Сеть окна графика. При выборе опции **Воспроизвести все и перепланировать** граф перерисовывается, настраиваясь на включения всех ранее скрытых точек и их соединений.

Выбрать, то есть "выделить", линии

Выбранные линии выделяются красным цветом.

1. Для выбора одной линии щелкните по ней левой кнопкой.
2. Для выбора нескольких кнопок следуйте одному из следующих вариантов:

- Используя указатель мыши, нарисуйте круги вокруг точек, линии которых вы хотите выбрать.
- Удерживая клавишу Ctrl и щелкая левой кнопкой мыши по отдельным линиям, которые вы хотите выбрать.

Отменить выделение всех выбранных линий можно, щелкнув по фону графика или выбрав пункт **Очистить выделение** в меню Сеть в окне графа.

Просмотреть граф с использованием другого макета

Чтобы изменить макет графа, в меню Сеть выберите **Круговой макет**, **Сетевой макет**, **Направленный макет** или **Макет сетки**.

Чтобы включить или отключить ползунок связей

В меню Вид выберите **Ползунок связей**.

Выбрать или отметить флагом записи в одной взаимосвязи

1. Щелкните правой кнопкой мыши по линии, представляющей нужную взаимосвязь.
2. В контекстном меню выберите **Сгенерировать узел Выбор для связи** или **Сгенерировать узел извлечения для связи**.

Узел Выбор или узел Извлечение автоматически добавляется на холст потока с соответствующими заданными опциями и условиями:

- Узел Выбор выбирает все записи в данной взаимосвязи.
- Узел Извлечение генерирует флаг, обозначающий, удерживает ли выбранная взаимосвязь значение true для записей во всем наборе данных. Название поля флага строится соединением двух значений во взаимосвязи нижним подчеркиванием, например, *LOW_drugC* или *drugC_LOW*.

Выбрать или отметить флагом записи для группы взаимосвязей

1. Выберите линии выведенной структуры, представляющие нужные взаимосвязи.
2. В меню Генерировать в окне графика выберите **Выбрать узел ("And")**, **Выбрать узел ("Or")**, **Извлечь узел ("And")** или **Извлечь узел ("Or")**.

- Узлы "Or" дают дизъюнкцию условий. Это означает, что узел будет применен к записям, для которых присутствует любая из выбранных взаимосвязей.
- Узлы "And" дают конъюнкцию условий. Это означает, что узел будет применен к записям, для которых присутствуют все выбранные взаимосвязи. Если какие-то из выбранных взаимосвязей взаимно исключаются, возникнет ошибка.

После завершения вашего выбора узел Выбор или узел Извлечение автоматически добавляется на холст потока с соответствующими заданными опциями и условиями.

Корректировка порогов сетевых графов

После создания сетевого графа можно скорректировать пороги, управляющие стилями линий с помощью ползунка панели инструментов для изменения минимальной видимой линии. Вы можете также просмотреть дополнительные опции порогов, нажав желтую кнопку с двойной стрелкой на панели инструментов, чтобы раскрыть окно графа. Затем перейдите на вкладку **Элементы управления**, чтобы просмотреть дополнительные опции.

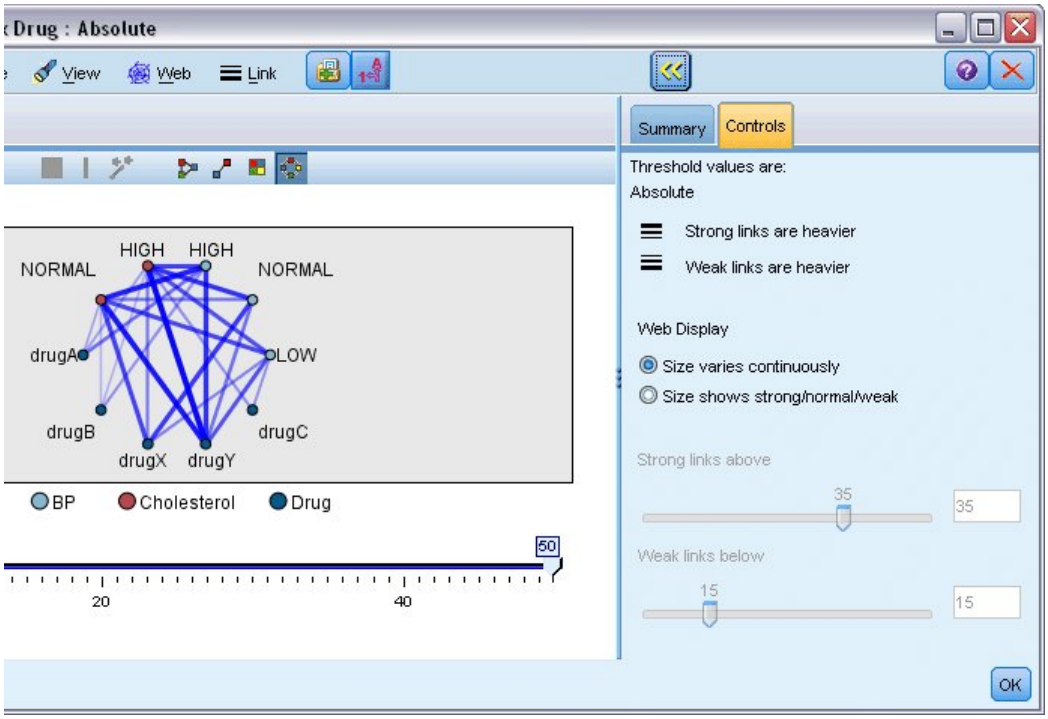


Рисунок 40. Вывод расширенного окна и опции порогов

Пороговые значения - . Показывает тип порога, выбранного при его создании в диалоговом окне узла Web.

Сильные связи тяжелее. Выбирается по умолчанию, это стандартный способ показа связей между полями.

Слабые связи тяжелее. Выбрать для обращения смысла связей, показываемых жирными линиями. Эта опция часто используется для обнаружения мошенничества или исследования выбросов.

Вывод сетевого графа. Задать опции для управления размером связей на выходном графе:

- **Размер меняется непрерывно.** Выбрать для вывода диапазона размеров связей, отображающего различия в силе соединений на основе фактических значений данных.
- **Размер отражает сильную/нормальную/слабую.** Выбрать для вывода трех значений для силы соединения - сильного, нормального и слабого. Точки отсечения для этих категорий можно задать выше, а также на финальном графе.

Сильные связи выше. Задать число, обозначающее порог для сильных (жирные линии) и обычных (простые линии) соединений. Все соединения со значением выше рассматриваются как сильные. Используйте ползунок, чтобы скорректировать значение или ввести число в поле.

Слабы связи ниже. Задать число, обозначающее порог для слабых (пунктирные линии) и обычных (простые линии) соединений. Все соединения со значением ниже рассматриваются как слабые. Используйте ползунок, чтобы скорректировать значение или ввести число в поле.

После корректировки порогов для сетевых графов можно перепланировать или перерисовать граф с новыми значениями порогов через меню графа, расположенное на панели инструментов сетевого графа. После того как найдены параметры, выявляющие наиболее значимые структуры, можно изменить исходные параметры на узле Web (называемом также родительским Web-узлом), выбрав в меню окна графа пункт **Изменить родительский узел**.

Создание сводки сетевого графа

Вы можете создать документ сводки сетевого графа, в котором перечисляются сильные, средние и слабые связи, нажав желтую кнопку с двойной стрелкой на панели инструментов, чтобы раскрыть окно сетевого графа. Затем перейдите на вкладку **Сводка** для просмотра таблиц для каждого типа связи. Таблицы можно раскрывать и сворачивать при помощи кнопок-переключателей для каждой из них.

Чтобы распечатать сводку, выберите следующий пункт в меню окна сетевого графа:

Файл > Печать сводки

Узел Оценка

Узел Оценка предоставляет простой способ оценки и сравнения моделей предсказания, чтобы выбрать лучшую модель для вашей прикладной программы. Диаграммы оценки показывают, как ведут себя модели при предсказании конкретных выходных данных. Они работают, сортируя записи на основе предсказанного значения и надежности предсказания, распределяя записи по группам равного размера **квантили**), а затем строя график значений бизнес-критерия для каждой квантили от максимального до минимального. Несколько моделей представляются разными линиями на графике.

Выходные данные обрабатываются с помощью определения конкретного значения или диапазона значений как **попаданий**. Обычно попадания обозначают успешность в каком-то смысле (например, в продажах) или важное событие (такое как конкретный медицинский диагноз). На вкладке Опции диалогового окна можно определить критерии попаданий, но можно использовать и критерии попаданий по умолчанию следующим образом:

- **Флаговые** поля вывода; попадания соответствуют значениям *true*.
- **Номинальные** поля вывода; попадание определяется первым значением в наборе.
- Для **непрерывных** полей вывода попадания - это значения больше средней точки диапазона поля.

Существует шесть типов диаграмм оценки, каждая из которых фокусируется на одном из различных критериев оценки.

Диаграммы выигрыша

Выигрыш определяется как доля числа попаданий в данной квантили. Выигрыш вычисляется в процентах: (количество попаданий в квантили / полное количество попаданий) × 100%.

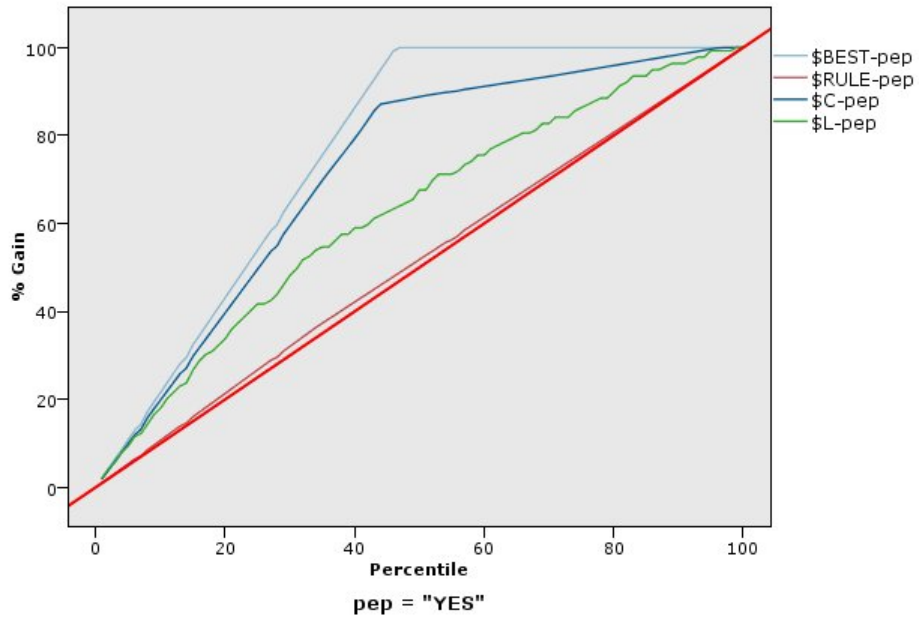


Рисунок 41. Диаграмма выигрыша (кумулятивная) с выводом базовой линии, наилучшей линии и бизнес-правила

Диаграммы роста

Рост сравнивает процентную долю записей в каждой квантили, представляющих собой попадания, с общей процентной долей попаданий в обучающих данных. Он вычисляется следующим образом: (число попаданий в квантили / число записей в квантили) / (полное число попаданий / полное число записей).

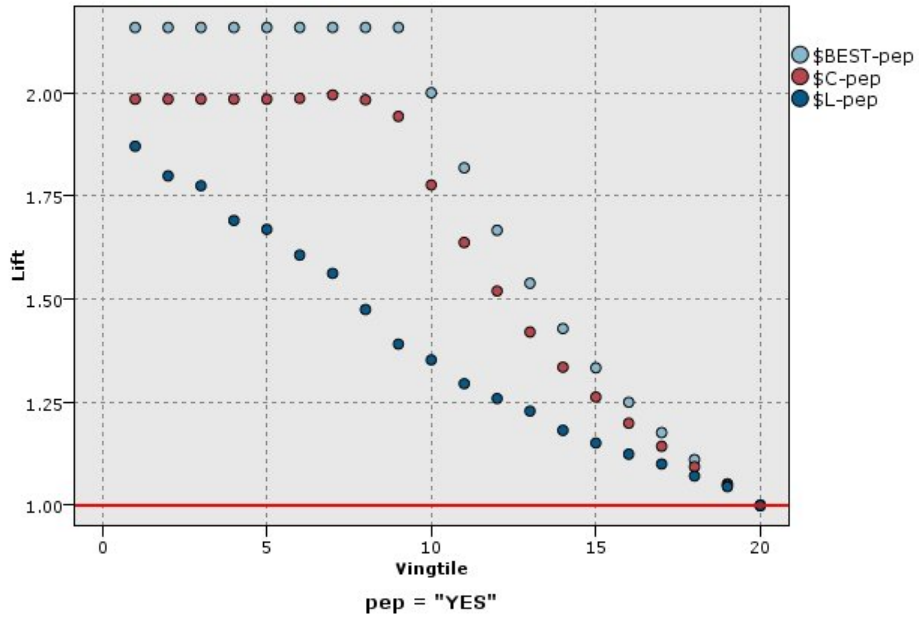


Рисунок 42. Диаграмма роста (кумулятивная), использующая точки и наилучшую линию

Диаграммы откликов

Отклик - это просто процентная доля попаданий по отношению ко всем записям квантили. Отклик вычисляется следующим образом: (число попаданий в квантили / число записей в квантили) × 100%.

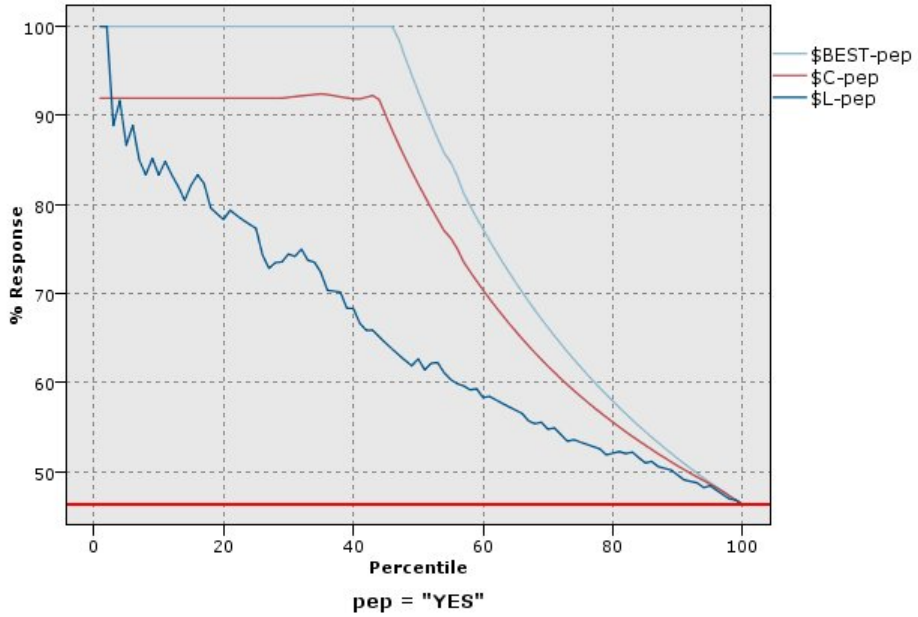


Рисунок 43. Диаграмма роста (кумулятивная) с наилучшей линией

Диаграммы прибыли

Прибыль равна **доходу** для каждой записи за вычетом **затрат**. Прибыли для квантили - просто сумма прибылей для всех записей в квантили. Предполагается, что доходы применимы только к записям, квалифицированным как попадания, а затраты - ко всем записям. Прибыли и затраты в данных можно фиксировать или определять значениями в полях. Прибыли вычисляются следующим образом: (сумма доходов для записей в квантили - сумма затрат для записей в квантили).

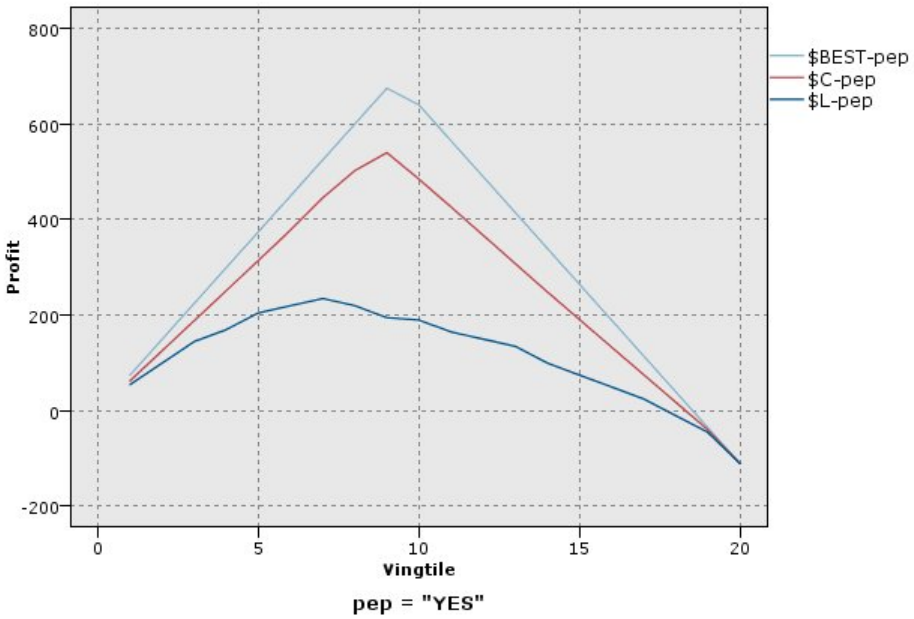


Рисунок 44. Диаграмма прибыли (кумулятивная) с наилучшей линией

Диаграммы ROI

Прибыль на инвестированный капитал (return on investment, ROI) аналогична прибыли, так как тоже учитывает объем прибыли и затрат. ROI сравнивает прибыли с затратами для квантили. ROI вычисляется следующим образом (прибыли для квантили / затраты для квантили) × 100%.

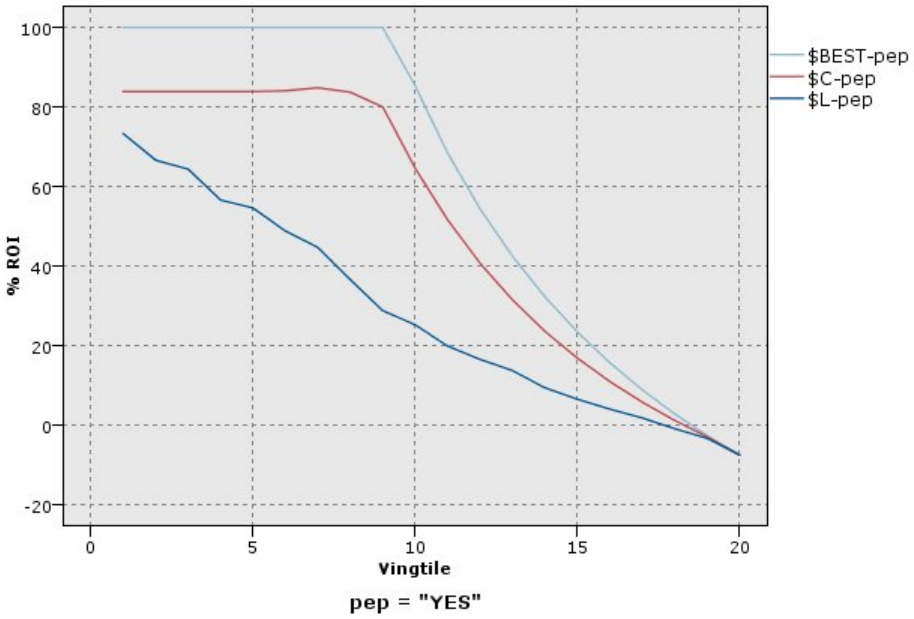


Рисунок 45. Диаграмма ROI (кумулятивная) с наилучшей линией

Диаграммы ROC

ROC (receiver operating characteristic, характеристика работы приемника) можно использовать только с бинарными классификаторами. ROC можно использовать для визуализации, организации и выбора классификаторов на основании их производительности. На диаграмме ROC показывается зависимость доли истинных положительных заключений от доли ложных положительных заключений классификатора. Диаграмма ROC изображает относительные соотношения между прибылями (истинные положительные заключения) и затратами (ложные положительные заключения). Истинное положительные заключение - это экземпляр попадания, классифицируемый как попадание. Поэтому доля истинных положительных заключений вычисляется как количество истинных положительных заключений, деленное на количество экземпляров, представляющих собой фактические попадания. Ложное положительное заключение - это экземпляр промаха, классифицируемый как попадание. Поэтому доля ложных положительных заключений вычисляется как количество ложных положительных заключений, деленное на количество экземпляров, представляющих собой фактические промахи.

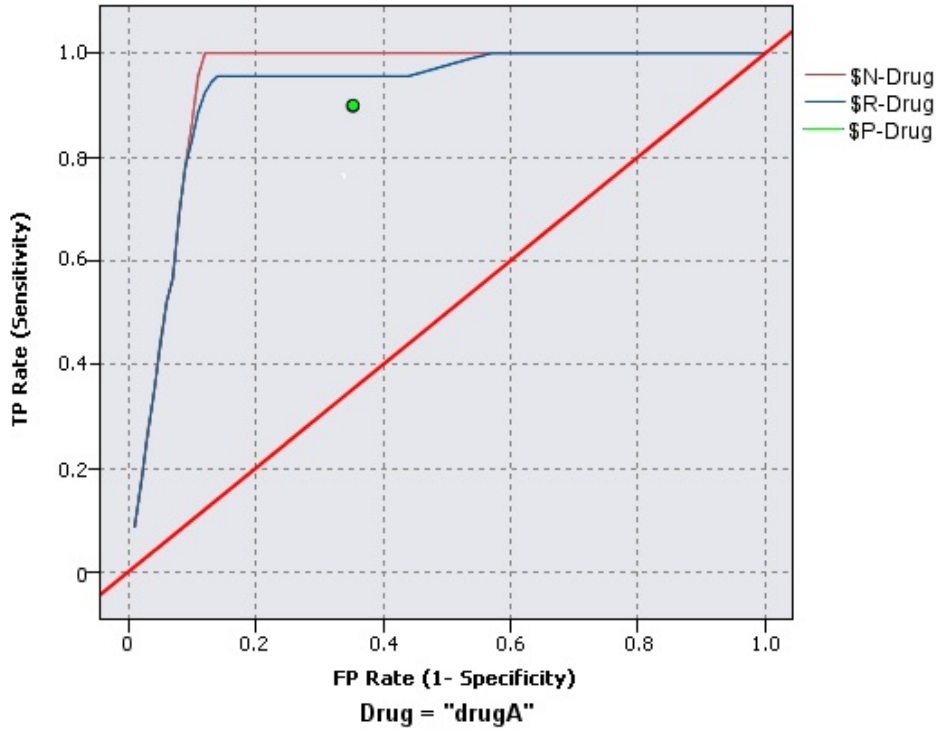


Рисунок 46. Диаграмма ROC с наилучшей линией

Диаграммы оценки могут также быть кумулятивными, так что значение в каждой точке равно значению для соответствующей квантили плюс для всех более высоких квантилей. Кумулятивные диаграммы обычно лучше представляют общую производительность моделей, в то время как некумулятивные диаграммы часто лучше обозначают конкретные проблемные области для моделей.

Вкладка График оценки

Тип диаграммы. Выберите один из следующих типов: **Выигрыш**, **Отклик**, **Рост**, **Доход**, **ROI** (return on investment - прибыль на инвестицию) или **ROC** (receiver operating characteristic, характеристика работы приемника).

Интегральный график. Выберите эту опцию, чтобы создать кумулятивную диаграмму. Значения на кумулятивной диаграмме выводятся для каждой квантили плюс для всех более высоких квантилей. (**Кумулятивный график** недоступен для диаграмм ROC).

Включить базовый уровень. Выберите эту опцию для включения в график базового уровня, указывающего совершенно случайное распределение совпадений, где доверительная вероятность становится иррелевантной. (Опция **Включить базовый уровень** для диаграмм Доход и ROI недоступна.)

Включить лучший уровень. Выберите эту опцию, чтобы включить в график лучшую линию, указывающую совершенную доверительную вероятность (где число совпадений = 100% наблюдений). (Опция **Включить базовый уровень** недоступна для диаграмм ROC).

Использовать критерии выигрыша для всех типов диаграмм. Выберите для использования критерия выигрыша (стоимость, прибыль и вес) при вычислении показателей оценки вместо обычного числа попаданий. Для моделей с определенными числовыми полями назначения, таких модель, которая предсказывает доход, полученный от заказчика, если он откликнется на предложение, значение поля назначения дает лучший показатель эффективности, чем количество попаданий. Выбор этой опции включает поля **Стоимости**, **Прибыль** и **Вес** для диаграмм Выигрыш, Отклик и Рост. Чтобы использовать критерий выигрыша для диаграмм этих трех типов, рекомендуется задать **Доход** как поле назначения, **Стоимость** равной 0,0 (то есть прибыль равна доходу) и пользовательское условие попадания "true" (то есть все записи рассматриваются как попадания). (Опция **Использовать критерии выигрыша для всех типов диаграмм** недоступна для диаграмм ROC).

Найти предсказанные поля / поля предикторов, применив. Выберите либо опцию **Метаданные выходных полей модели** для поиска предсказанных полей на диаграмме по их метаданным, либо опцию **Формат имени поля** для их поиска по имени.

Поля оценки диаграмм. Включите этот переключатель, чтобы включить средство выбора полей оценки. Затем выберите один или несколько полей оценки (полей диапазона или непрерывных полей), то есть, полей, которые не будут строго предсказательными моделями, но могут оказаться полезными для ранжирования записей в понятиях совпадения склонности. Узел оценки может сравнить любое поле или сочетание полей оценки с одной или несколькими предсказательными моделями. Типичный пример - сравнение нескольких полей RFM с лучшей предсказательной моделью.

Назначение. Выберите поле назначения при помощи средства выбора полей. Выберите инстанциированное флаговое или номинальное поле с несколькими значениями.

Примечание: Это поле назначения применимо только к полям оценки (предсказательные модели определяют свои собственные назначения) и игнорируется, если на вкладке Опции задан пользовательский критерий совпадения.

Разбить по разделам. Если для разбиения записей на обучающую, контрольную и проверочную выборки используется поле раздела, выберите эту опцию, чтобы для каждого раздела выводилась отдельная диаграмма оценки. Дополнительную информацию смотрите в разделе “Узел раздела” на стр. 167.

Примечание: При разбиении по разделам записи с пустыми значениями в поле раздела исключаются из оценки. Это никогда не становится проблемой, если используется узел Раздел, поскольку узлы Раздел не генерируют пустых значений.

График. Выберите размер квантилей для представления на диаграмме в выпадающем списке. В состав опций входят: **Квартили**, **Квинтили**, **Децили**, **Вингтили**, **Процентили** и **1000-тили**. (**График** недоступен для диаграмм ROC).

Стиль. Выберите **Линия** или **Точка**.

Для всех типов диаграмм, кроме диаграмм ROC, дополнительные управляющие элементы позволяют задавать стоимости, прибыль и веса.

- **Стоимости.** Задайте стоимость, связанную с каждой записью. Для стоимости можно выбрать опции **Фиксированная** или **Переменная**. Для фиксированных стоимостей задайте нужные значения. Для переменных стоимостей нажмите кнопку Выбор полей, чтобы выбрать поле стоимостей. (Для диаграмм ROC поле **Стоимости** недоступно.)
- **Прибыль.** Задайте прибыль, связанную с каждой записью, представляющей попадание. Для стоимости можно выбрать опции **Фиксированная** или **Переменная**. Для фиксированных прибылей задайте нужные значения. Для переменных прибылей нажмите кнопку Выбор полей, чтобы выбрать поле прибыли. (Для диаграмм ROC поле **Прибыль** недоступно.)
- **Вес.** Если записи в ваших данных представляют несколько блоков, можно использовать частотные веса, чтобы скорректировать результаты. Задайте вес, связанный с каждой записью, используя **Фиксированные** или **Переменные** веса. Для фиксированных весов задайте значение веса (количество блоков на запись). Для переменных весов нажмите кнопку Выбор полей, чтобы выбрать поле весов. (Для диаграмм ROC поле **Вес** недоступно.)

Вкладка Опции оценки

Вкладка Опции для диаграмм оценки обеспечивает гибкость определения совпадений, критериев скоринга и бизнес-правил, выводимых на диаграмме. Можно также задать опции для экспорта результатов оценки модели.

Пользовательское совпадение. Выберите эту опцию, чтобы задать пользовательское условие, используемое для указания совпадения. Эта опция полезна для определения результатов исследования вместо их определения по типу поля назначения и порядка значений.

- **Условие.** При выборе опции **Пользовательское совпадение** выше нужно задать выражение CLEM для условия совпадения. Например, `@TARGET = "YES"` - это допустимое условие, указывающее, что значение *Да* для поля назначения при оценке будет считаться совпадением. Заданное условие будет использоваться для всех полей назначения. Чтобы создать условие, введите его в поле или сгенерируйте выражение условия при помощи Построителя выражений. Если данные инстанциированы, можно будет вставить значения непосредственно из Построителя выражений.

Пользовательская оценка. Выберите эту опцию, чтобы задать условие, используемое для скоринга наблюдений перед назначением их для квантилей. Оценка по умолчанию вычисляется из предсказанного значения и доверительной вероятности. При помощи Построителя выражений создайте пользовательское выражение для скоринга.

- **Выражение.** Задайте выражение CLEM, используемое для скоринга. Например, если числовые результаты в диапазоне от 0 до 1 упорядочены так, что более низкие значения лучше более высоких, можно определить совпадение как `@TARGET < 0,5` и связанную с ним оценку как `1 - @PREDICTED`. Выражение оценки должно получать числовое значение. Чтобы создать условие, введите его в поле или сгенерируйте выражение условия при помощи Построителя выражений.

Включить бизнес-правило. Выберите эту опцию, чтобы задать условие правила, отражающее критерии исследования. Например, можно вывести на экран правило для всех наблюдений, где `закладная = "Y", а доход >= 33000`. Бизнес-правила отображаются на диаграмме и снабжаются в ключе меткой *Правило*. (Для диаграмм ROC **Включение бизнес-правила** не поддерживается.)

- **Условие.** Задайте выражение CLEM, используемое для определения бизнес-правила на диаграмме вывода. Просто введите его в поле или сгенерируйте выражение условия при помощи Построителя выражений. Если данные инстанциированы, можно будет вставить значения непосредственно из Построителя выражений.

Экспорт результатов в файл. Выберите эту опцию, чтобы экспортировать результаты оценки модели в текстовый файл с разделителями. Этот файл можно прочитать для выполнения специализированного анализа по вычисленным значениям. Задайте следующие опции для экспорта:

- **Имя файла.** Введите имя для файла вывода. Нажмите кнопку с многоточием **(...)** для просмотра нужной папки.

- **Ограничитель.** Введите символ, например знак запятой или пробела, для использования в качестве разделителя полей.

Включить имена полей. Выберите эту опцию, чтобы включать имена полей в качестве первой строки файла вывода.

Символ новой строки после каждой записи. Выберите эту опцию, чтобы начинать каждую запись в новой строке.

Вкладка Внешний вид оценки

Перед созданием графиков можно задать опции внешнего вида.

Заголовок. Введите текст для использования в качестве заголовка диаграммы.

Подзаголовок. Введите текст для использования в качестве подзаголовка диаграммы.

Текст. Либо примите текстовую метку, сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Метка X. Либо примите метку оси *x* (горизонтальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Метка Y. Либо примите метку оси *y* (вертикальной), сгенерированную автоматически, либо выберите **Пользовательская**, чтобы ее задать.

Показать сетку. Эта опция, включаемая по умолчанию, выводит позади графика сетку, упрощающую определение точек отсечения для регионов и полос. Линии сетки всегда выводятся белым цветом, если только не используется белый фон диаграммы; в этом случае они выводятся серым цветом.

Чтение результатов оценки модели

Интерпретация диаграммы оценки зависит от определенного экстента для типа диаграммы, но некоторые характеристики свойственны всем диаграммам оценки. Для кумулятивных диаграмм, чем выше линии, тем лучше модели, которые они представляют, особенно в левой части диаграммы. Во многих случаях при сравнении нескольких моделей линии будут пересекаться, так что одна модель будет выше в одной части диаграммы, а другая - выше в другой части диаграммы. В этом случае, принимая решение, какую модель выбрать, нужно посмотреть, какая часть выборки (определяющая точку на оси *x*) вам потребуется.

Большинство некумулятивных диаграмм будут очень похожи. Для хороших моделей кривые некумулятивных диаграмм должны быть высоки по отношению к левой стороне диаграммы и низки по отношению к правой стороны диаграммы. (Если на некумулятивной диаграмме выводится пилообразная кривая, ее можно сгладить, сократив число квантилей для построения графика и перепостроив диаграмму.) Провалы в левой части диаграммы или скачки в правой части могут указывать на области плохого предсказания модели. Горизонтальная линия через всю диаграмму указывает на модель, фактически не предоставляющую информацию.

Диаграммы выигрыша. Кумулятивные диаграммы выигрыша всегда начинаются с 0% и заканчиваются 100% в направлении слева направо. Для хороших моделей кривая диаграммы выигрыша будет круто подниматься до 100%, а затем выравниваться. Модели, не дающим информацию будут представлены кривой, тяготеющей к диагональной линии слева направо и снизу вверх (выводящейся на диаграмме, если выбрана опция **Включить базовый уровень**).

Диаграммы роста. Кривые кумулятивных диаграммам роста, как правило, начинаются выше 1,0 и постепенное понижаются, пока не достигнут 1,0, в направлении слева направо. Правый край диаграммы представляет весь набор данных, поэтому отношение числа совпадений в кумулятивных квантилях к числу совпадений в данных будет равно 1,0. Для хорошей модели кривая диаграммы роста должна начинаться

значительно выше 1,0 в левой части, оставаться высоко на горизонтальной полке по направлению вправо, а затем резко снижаться до 1,0 с правой стороны диаграммы. Для модели, не дающей информации, кривая будет колебаться около 1,0 по всей диаграмме. (Если выбрана опция **Включить базовый уровень**, на диаграмме для справки будет показана горизонтальная линия на уровне 1,0.)

Диаграммы откликов. Кумулятивные диаграммы откликов, как правило, очень похожи на диаграммы роста, за исключением масштабирования. Кривые диаграмм откликов обычно начинаются возле 100% и постепенно снижаются, пока не достигнут суммарного уровня отклика (итоговое число совпадений / общее число записей) на правом краю диаграммы. Для хорошей модели кривая отклика будет начинаться возле 100% на правом краю диаграммы, оставаться высоко на горизонтальной полке в направлении вправо, а затем резко спадать до суммарного уровня отклика на правом краю диаграммы. Для модели, не дающей информации, кривая будет колебаться около суммарного уровня отклика по всей диаграмме. (Если выбрана опция **Включить базовый уровень**, на диаграмме для справки будет показана горизонтальная линия на уровне суммарного отклика.)

Диаграммы прибылей. Кумулятивные диаграммы прибыли показывают сумму прибылей с ростом размера выбранной вами выборки в направлении слева направо. Кривые диаграмм прибылей обычно начинаются возле 0, монотонно растут в направлении слева направо, пока не достигнут пика или горизонтальной полки в середине диаграммы, а затем снижаются в сторону правого края диаграммы. Для хорошей модели кривая прибыли продемонстрирует четко выраженный пик где-нибудь в середине диаграммы. Для модели, не дающей информации, линия будет относительно прямой и может повышаться, понижаться или выравниваться в зависимости от применяемой структуры типа затраты/доход.

Диаграммы ROI. Кумулятивные диаграммы ROI (return on investment - прибыль на инвестицию), как правило, похожи на диаграммы откликов и диаграммы роста, за исключением масштабирования. Кривые диаграмм ROI обычно начинаются возле 0% и постепенно понижаются, пока не достигнут итогового значения ROI для всего набора данных (которое может быть отрицательным). Для хорошей модели кривая должна начинаться значительно выше 0%, оставаться высоко на горизонтальной полке в направлении вправо, а затем достаточно резко спадать до итогового значения ROI на правом краю диаграммы. Для модели, не дающей информации, кривая должна колебаться около итогового значения ROI.

Диаграммы ROC. Обычно у кривых ROC вид кумулятивной диаграммы выигрыша. При просмотре слева направо кривая начинается в точке с координатами (0,0) и заканчивается в точке (1,1). Диаграмма, которая круто растет в направлении точки (0,1), а затем выравнивается, обозначает хороший классификатор. Модель, случайным образом классифицирующая экземпляры как попадания или промахи, дает диаграмму, повышающуюся по диагонали от нижнего левого до верхнего правого угла (показывается на диаграмме, если выбрана опция **Включить базовый уровень**). Если у модели нет поля достоверности, она представлена одной точкой. Классификатор с оптимальным порогом классификации расположен ближе всего к точке (0,1), то есть к верхнему левому углу диаграммы. Это положение представляет большое число экземпляров, правильно классифицированных как попадание, и малое количество экземпляров, неправильно классифицированных как попадания. Точки над диагональю соответствуют хорошим результатам классификации. Точки под диагональю представляют плохие результаты классификации, которые хуже, чем случайная классификация экземпляров.

Использование диаграммы оценки

Изучение диаграммы оценки при помощи мыши аналогично использованию гистограммы или диаграммы собрания. Ось *x* представляет оценки модели по заданным квантилям, таким как вингтили или децили.

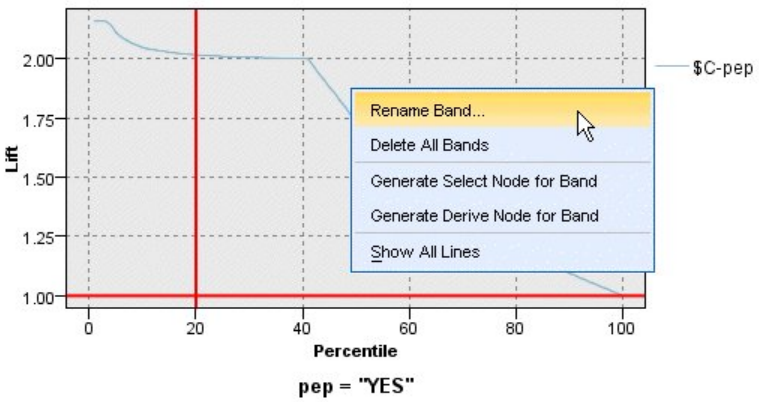


Рисунок 47. Работа с диаграммой оценки

Ось *x* можно разделить на полосы также, как и для гистограммы, при помощи значка разделителя, выводящего опции для автоматического разбиения оси на равные полосы. Дополнительную информацию смотрите в разделе “Исследование графиков” на стр. 260. Границы полос можно отредактировать вручную, выбрав в меню Правка опцию **Полосы диаграммы**.

Создав диаграмму оценки, определив полосы и изучив результаты, вы можете, используя опции в меню Создать и контекстном меню, автоматически создать узлы на основе вариантов, выбранных на диаграмме. Дополнительную информацию смотрите в разделе “Генерирование узлов из диаграмм” на стр. 267.

При генерировании узлов из диаграммы оценки будет предложено выбрать одну модель изо всех доступных моделей на диаграмме.

Выберите модель и нажмите кнопку **OK**, чтобы сгенерировать новый узел на холсте потока.

Узел визуализации карты

Узел Визуализация карты может принять несколько входных соединений и вывести геопространственные данные на карту как несколько слоев. Каждый слой - это одно геопространственное поле; например, базовым слоем может быть карта страны, выше может накладываться один слой дорог, один слой рек и один слой городов.

Хотя обычно большинство геопространственных наборов данных содержит одно геопространственное поле, при наличии нескольких геопространственных полей в одном вводе можно выбрать, какие поля показывать. Нельзя одновременно показывать два поля из одного входного соединения; однако можно скопировать в буфер и вставить данные входящего соединения и в одном поле выводить значения другого.

Вкладка График визуализации карты

Слои

В этой таблице представлена информация о входных данных для узла карт. Порядок слоев определяет порядок, в котором эти слои выводятся и на предварительном изображении карты, и для визуализации вывода при вызове узла. Верхняя строка таблицы - это ‘верхний’ слой, а нижняя строка - ‘базовый слой’; другими словами, каждый слой показывается на карте над слоем, непосредственно следующим после него в таблице.

Примечание: Когда слой в таблице содержит трехмерное геопространственное поле, выводятся только значения по координатам x и y. Координата z игнорируется.

Имя Имена создаются для каждого слоя автоматически и строятся по следующему формату:

тег[узел_источника:присоединенный_узел]. По умолчанию тег показывается как число, причем 1 представляет первый присоединенный источник ввода, 2 - второй и так далее. При необходимости нажмите кнопку **Изменить слой** в диалоговом окне Опции изменения слоев карты, чтобы изменить тег. Например, вы можете изменить тег на "дороги" или "города" для отображения содержания входных данных.

Тип Показывает значок типа измерения геопространственного поля, выбранного для слоя. Если входные данные содержат несколько полей с геопространственным типом измерений, их выбор по умолчанию использует следующий порядок сортировки:

1. Точки
2. Ломаная
3. Многоугольник
4. Несколько точек
5. Мультиломаная
6. Мультиполигон

Примечание: Если существует два поля с одинаковым типом измерений, по умолчанию выбирается первое поле в алфавитном порядке.

Идентификатор

Примечание: Этот столбец заполняется только для полей Точка и Несколько точек.
Показывает обозначение, используемое для полей Точка или Несколько точек. При необходимости нажмите кнопку **Изменить слой** в диалоговом окне Опции изменения слоев карты, чтобы изменить обозначение.

Цвет Показывает цвет, выбранный для представления слоя на карте. При необходимости нажмите кнопку **Изменить слой** в диалоговом окне Опции изменения слоев карты, чтобы изменить цвет. Цвет применяется к различным элементам в зависимости от типа измерения.

- Для полей Точка или Несколько точек цвет применяется к обозначению для слоя.
- Для ломаных и многоугольников цвет применяется ко всей форме. У многоугольников всегда есть черный контур; показанный в столбце цвет - это цвет, используемый для заполнения формы.

Просмотр

На этой панели показано предварительное изображение текущего выбора входных элементов в таблице **Слои**. На предварительном изображении учитывается порядок слоев, обозначения, цвет и все другие параметры вывода, связанные со слоями, и при возможности - изменяется вывод при изменении параметров. Если вы меняете некоторые параметры где-то еще в потоке, например, для использования геопространственных полей в качестве слоев или для уточнения некоторых подробностей, таких как связанные функции агрегации, для изменения предварительного изображения нужно нажать кнопку **Обновить данные**.

Используйте **Предварительный просмотр**, чтобы задать параметры вывода на экран, прежде чем запускать ваш поток. Чтобы не допустить больших задержек по времени, связанных с использованием большого набора данных, для предварительного просмотра используется выборка данных по каждому слою и построение изображения по первым 100 записям.

Изменение слоев карты

Диалоговое окно Опции изменения слоев карты можно использовать для внесения поправок в различные подробности о любом из слоев, показанных на вкладке **График** узла визуализации карт.

Подробности входных данных

Тег По умолчанию тег - это число; вы можете заменить это число на более значимый тег для простоты идентификации слоя на карте. Например, тег может быть названием входных данных, таким как "Города".

Поле слоя

Если в ваших входных данных несколько геопространственных полей, используйте эту опцию, чтобы выбрать поле, которое вы хотите вывести как слой на карте.

По умолчанию слои, из которых можно сделать выбор, располагаются в следующем порядке сортировки.

- Точки
- Ломаная
- Многоугольник
- Несколько точек
- Мультиломаная
- Мультиполигон

Параметры дисплея

Использовать группировку шестиугольниками

Примечание: Эта опция применима только к полям точек и нескольких точек.

Группировка шестиугольниками объединяет близкие точки (на основании их координат x и y) в одну точку для вывода на карте. Эта одна точка показывается как шестиугольник, но на самом деле обрабатывается, как многоугольник.

Так как шестиугольник обрабатывается как многоугольник, любые поля точек с включенной группировкой шестиугольниками рассматриваются как многоугольники. Это означает, что при выборе опции **Упорядочить по типу** в диалоговом окне узла карты все слои точек с примененной группировкой шестиугольниками будут обрабатываться над слоями многоугольников, но под слоями ломаных и мультиломаных.

Если вы используете группировку шестиугольниками для полей нескольких точек, такое поле сначала преобразуется в поле точек, для чего значения нескольких точек группируются для вычисления центральной точки. Центральные точки используются для вычисления группировок шестиугольниками.

Агрегирование

Примечание: Этот столбец доступен только при включении переключателя **Использовать группировку шестиугольниками** и при выборе опции **Наложение**.

Если поле **Наложения** выбрано для слоя точек, использующих группировку шестиугольниками, все значения в этом поле должны агрегироваться для всех точек в составе шестиугольника. Задайте функцию агрегирования для любых полей наложения, которые вы хотите применить к карте. Доступные функции агрегирования зависят от типа измерения.

- Функции агрегирования для количественного типа измерений с системой хранения Действительное или Целое число:
  - Сумма
  - Среднее значение
  - Минимум
  - Максимум
  - Медиана
  - Первая квартиль

- Третья квартиль

- Функции агрегирования для количественного типа измерений с системой хранения Врем, Дата или Отметка времени:
  - Среднее значение
  - Минимум
  - Максимум
- Функции агрегирования для номинального или категориального типа измерений:
  - Режим
  - Минимум
  - Максимум
- Функции агрегирования для типа измерений Флаг:
  - True, если любое - true
  - False, если любое - false

Цвет

Используйте эту опцию или для выбора стандартного цвета, применяемого ко всем возможностям геопространственного поля, или поля наложения, когда цвета применяются к возможностям на основании значений из другого поля в данных.

Если выбрать опцию **Стандарт**, можно выбрать цвет из палитры, показанной на панели **Порядок цветов категорий диаграммы** вкладки Вывод на экран в диалоговом окне Пользовательские опции.

Для опции **Наложение** можно выбрать любое поле из источника данных, содержащего геопространственное поле, выбранное как **Поле слоя**.

- Для номинальных и категориальных полей наложения цветовая палитра для выбора цветов та же, что показана для опций цвета **Стандарт**.
- Для количественных и порядковых полей наложения появится второй выпадающий список, из которого выбирается цвет. При выборе цвета наложение применяется изменением насыщения этого цвета в соответствии со значением в количественном или порядковым поле. Максимальное значение соответствует цвету, выбранному в выпадающем списке, а меньшие значения показаны соответственно меньшим насыщением.

Идентификатор

Примечание: Допустимо только для типов измерений Точки и Несколько точек.

Используйте эту опцию для выбора, применять ли **Стандартное** обозначение ко всем записям геопространственного поля, или обозначение **Наложения**, которые изменяют значок обозначения для точек на основании другого поля в данных.

При опции **Стандарт** для представления данных точек на карте можно выбрать одно из обозначений по умолчанию из выпадающего списка.

При опции **Наложение** можно выбрать любое номинальное, порядковое или категориальное поле из источника данных, содержащего геопространственное поле, выбранное как **Поле слоя**. Для каждого значения в поле наложения на карте выводится другое обозначение.

Например, ваши данные могут содержать поле точек, представляющее положения магазинов, а для наложения может использоваться поле типа магазинов. В этом примере все продовольственные магазины на карте могут обозначаться крестиком, а магазины электроники - квадратиком.

Размер

Примечание: Допустимо только для типов измерений Точка, Несколько точек, Ломаная и Мультиломаная.

Используйте эту опцию для выбора, применять ли **Стандартный** размер ко всем записям геопространственного поля, или размер **Наложения**, при котором изменяется размер значков обозначений или толщина линий на основании другого поля в данных.

При опции **Стандарт** можно выбрать значение ширины в пикселях. Доступные опции - это 1, 2, 3, 4, 5, 10, 20 или 30.

Для опции **Наложение** можно выбрать любое поле из источника данных, содержащего геопространственное поле, выбранное как **Поле слоя**. Толщина линии или точки меняется в зависимости от значения выбранного поля.

Прозрачность

Используйте эту опцию для выбора, применять ли **Стандартную** прозрачность ко всем записям геопространственного поля, или прозрачность **Наложения**, при которой изменяется прозрачность обозначения, линии или многоугольника на основании другого поля в данных.

При опции **Стандарт** можно выбрать один из уровней прозрачности от 0% (непроницаемый) до 100% (прозрачно) с шагом в 10%.

Для опции **Наложение** можно выбрать любое поле из источника данных, содержащего геопространственное поле, выбранное как **Поле слоя**. Для каждого значения в поле наложения на карте выводится свой уровень прозрачности. Прозрачность применима к цвету, выбранному из выпадающего списка для точки, линии или многоугольника.

Метка данных

Примечание: Эта опция недоступна, если включен переключатель **Использовать группировку шестиугольниками**.

Используйте эту опцию для выбора поля, которое будет использоваться как метки данных на карте. Например, при применении на слое многоугольников метка данных может быть полем названий, содержащим название для каждого многоугольника. Если выбрано поле названий, эти названия выводятся на карте.

Вкладка Вид визуализации карты

Перед созданием графиков можно задать опции внешнего вида.

Заголовок. Введите текст для использования в качестве заголовка диаграммы.

Подзаголовок. Введите текст для использования в качестве подзаголовка диаграммы.

Подпись. Введите текст для использования в качестве подписи диаграммы.

Исследование графиков

По сравнению с режимом редактирования, позволяющим изменять макет и внешний вид графиков, режим исследования предоставляет возможность аналитически исследовать данные и значения, представленные графиком. Главная цель исследования - это анализ данных и последующая идентификация значений при помощи полос, областей и меток для создания узлов Выбор, Получение и Баланс. Чтобы войти в этот режим, выберите в меню **Вид> Режим исследования** (или щелкните по значку на панели инструментов).

Некоторые графики могут использовать все инструменты исследования, для других же доступен только один такой инструмент. Режим исследования включает в себя:

- Определение и изменение полос, используемых для разделения значений по градуированной оси *x*. Дополнительную информацию смотрите в разделе “Использование полос” на стр. 261.
- Определение и изменение участков, используемых для идентификации группы значений в прямоугольных областях. Дополнительную информацию смотрите в разделе “Использование регионов” на стр. 264.

- Нанесение меток для элементов и снятие таких меток, чтобы вручную выбирать значения, которые можно будет использовать для создания узлов Выбор и Получение. Дополнительную информацию смотрите в разделе “Использование помеченных элементов” на стр. 266.
- Создание узлов при помощи идентифицированных вручную значений, областей, отмеченных элементов и Web-ссылок, которые будут использоваться в вашем потоке. Дополнительную информацию смотрите в разделе “Генерирование узлов из диаграмм” на стр. 267.

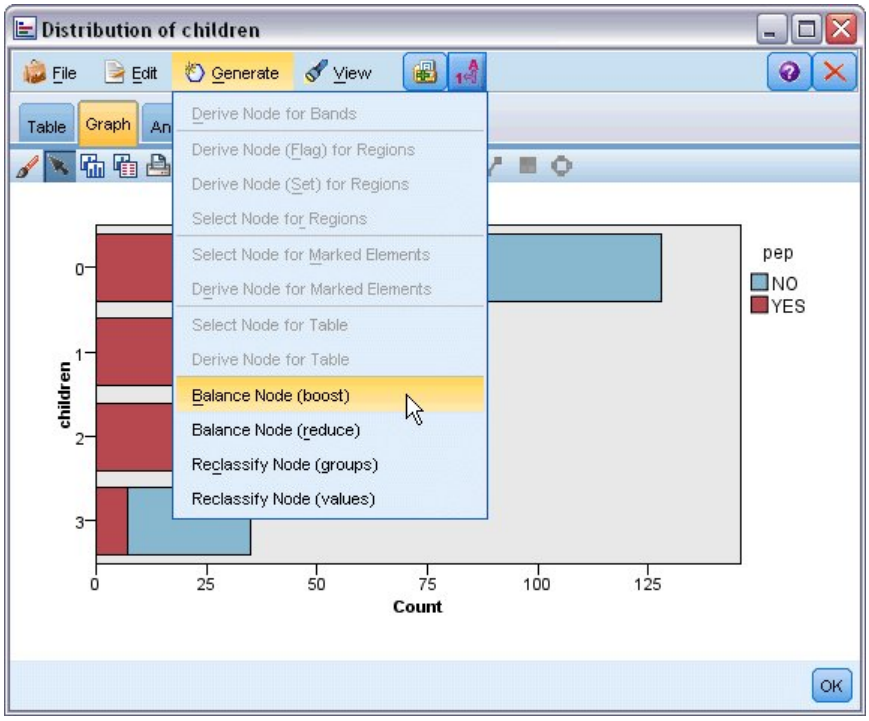


Рисунок 48. Диаграмма с выведенным меню генерирования

Использование полос

На диаграмме с полем масштабирования на оси *x* можно начертить вертикальные линии полос, чтобы разбить диапазон значений по оси *x*. Если у диаграммы несколько панелей, линия полосы, начерченная на одной панели, будет также представлена и на других панелях.

Не все диаграммы принимают полосы. Вот некоторые типы диаграмм, у которых могут быть полосы: гистограммы, столбчатые диаграммы и диаграммы распределений, графики (линейные, рассеяния, зависимостей от времени и так далее),диаграммы собраний и оценок. На диаграммах с несколькими панелями полосы появляются на всех панелях. А в некоторых случаях в матрице диаграммы рассеяния (SPLOM) вы увидите горизонтальную линию полосы, поскольку ось, по которой начерчена полоса поля/переменной, была перевернута.

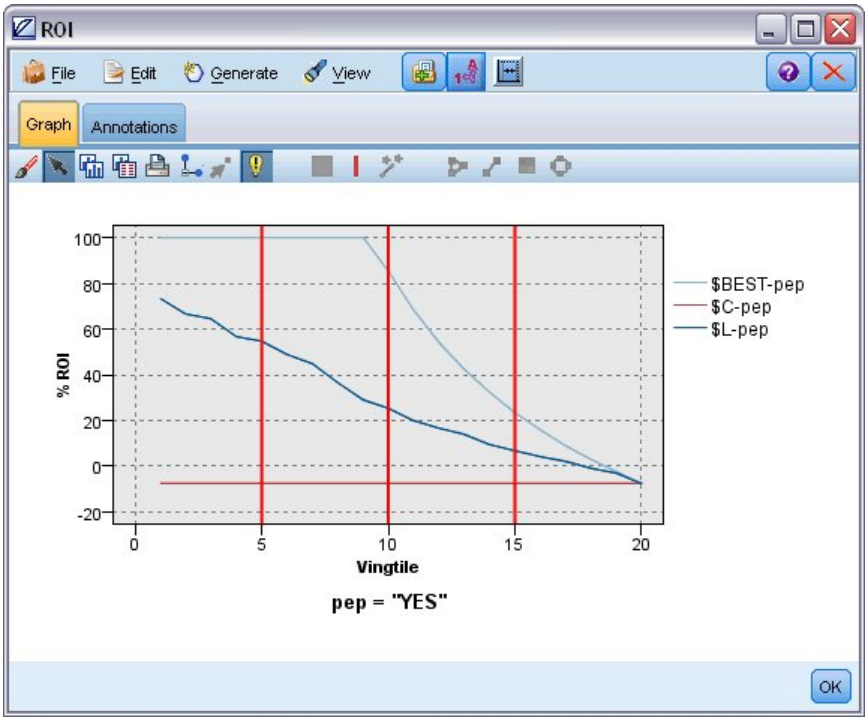


Рисунок 49. Диаграмма стремя полосами

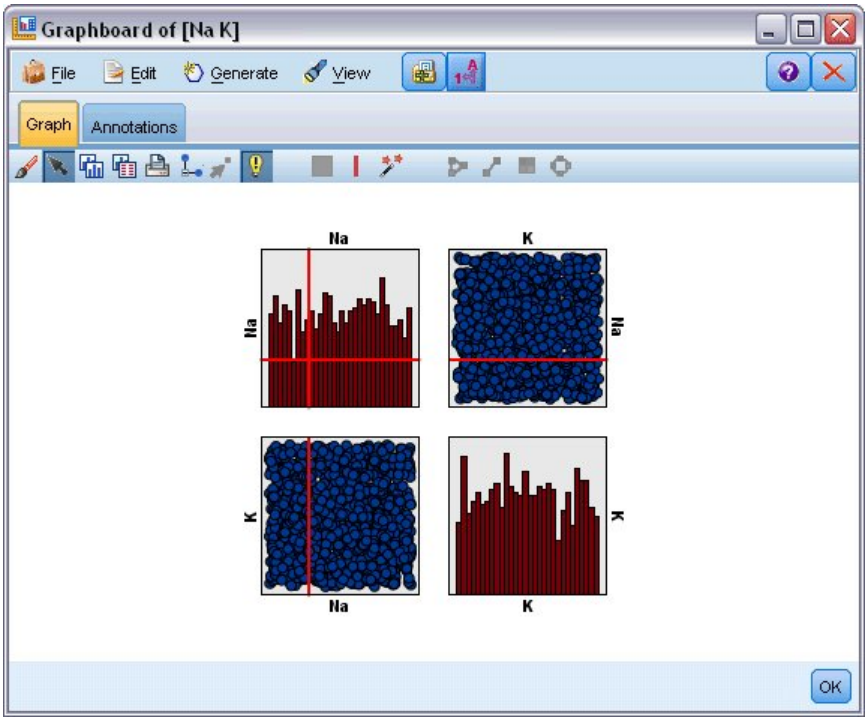


Рисунок 50. Матрица диаграммы рассеяния с полосами

Определение полос

На диаграмме без полос при добавлении линии полосы диаграмма разбивается на две полосы. Значение линии полосы при создании диаграммы в направлении слева направо называют также нижней границей

второй полосы. Подобным же образом на диаграмме с двумя полосами при добавлении линии полосы одна из двух полос разбивается на две, в результате чего получается три полосы. По умолчанию полосы именуются *bandN*, где *N* эквивалентно номерам полос по оси *x* слева направо.

После определения полосы можно изменить ее позицию на оси *x*. Щелкнув правой кнопкой мыши в полосе, можно вывести дополнительные ярлыки для задач, таких как переименование, удаление или генерирование узлов конкретно для этой полосы.

Чтобы определить полосы:

1. Убедитесь, что вы находитесь в режиме изучения. В меню выберите **Просмотр > Режим изучения**.
2. На панели инструментов режима изучения нажмите кнопку Начертить полосы.

Рисунок 51. Кнопка панели инструментов Начертить полосы

3. На диаграмме, принимающей полосы, щелкните по точке значений оси *x*, в которой вы хотите задать линию полосы.

Примечание: Другой вариант - щелкните по значку панели инструментов **Разбить диаграмму на полосы**, введите нужное вам число равных полос и нажмите кнопку **Разбить**.

Рисунок 52. Значок разделителя, используемый для расширения панели инструментов с добавлением опций для разбиения на полосы

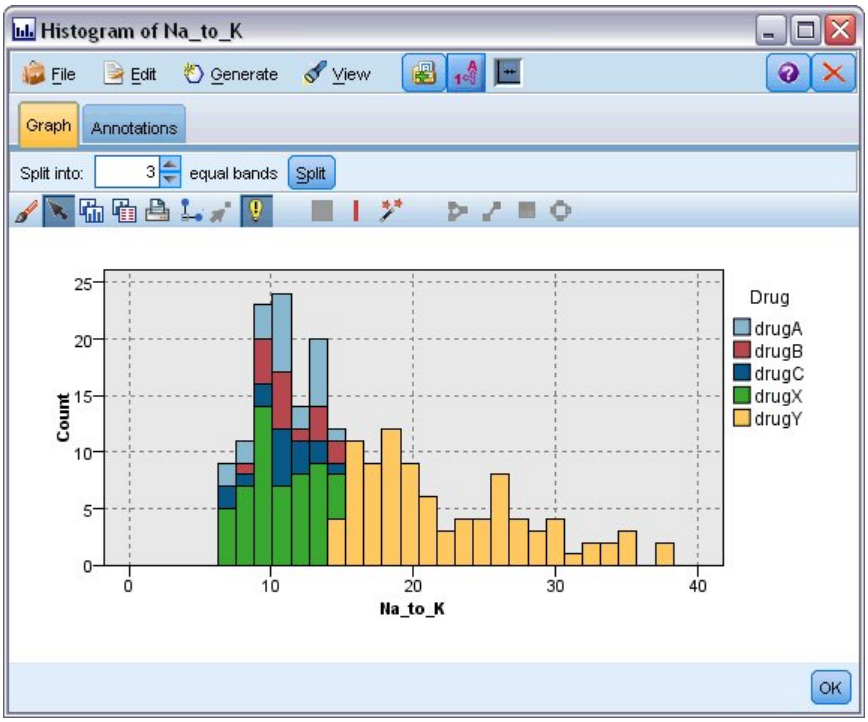


Рисунок 53. Создание панели инструментов равных полос с включенными полосами

Редактирование, переименование и удаление полос

Свойства существующих полос можно отредактировать в диалоговом окне Изменить полосы диаграмм или посредством контекстного меню на самой диаграмме.

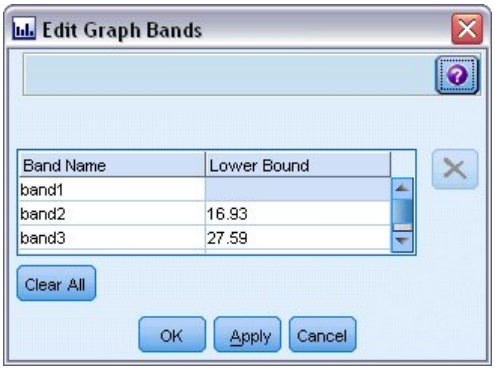


Рисунок 54. Диалоговое окно Изменить полосы диаграммы

Чтобы отредактировать полосы:

1. Убедитесь, что вы находитесь в режиме изучения. В меню выберите **Просмотр > Режим изучения**.
2. На панели инструментов режима изучения нажмите кнопку Начертить полосы.
3. В меню выберите **Изменить > Полосы диаграммы**. Откроется диалоговое окно Изменить полосы диаграммы.
4. Если на диаграмме несколько полей (например, на диаграммах SPLOM), нужное поле можно выбрать в выпадающем списке.
5. Добавьте новую полосу, введя имя и нижнюю границу. Чтобы начать новую строку, нажмите клавишу Enter.
6. Отредактируйте границу полосы, скорректировав значение **Нижняя граница**.
7. Переименуйте полосу, введя ее новое имя.
8. Удалите полосу, выбрав соответствующую строку в таблице и нажав кнопку удаления.
9. Нажмите кнопку **OK** , чтобы применить внесенные изменения и закрыть это диалоговое окно.

Примечание: Другой вариант - можно удалить и переименовать полосы непосредственно на диаграмме, щелкнув правой кнопкой мыши по линии полосы и выбрав нужную опцию в контекстных меню.

Использование регионов

На любой диаграмме с двумя количественными осями (или осями диапазонов) можно нарисовать регионы, чтобы сгруппировать значения в нарисованной вами прямоугольной области, которая называется регионом. **Регион** - это область диаграммы, описываемая максимальным и минимальным значениями на осях *X* и *Y*. Если у диаграммы несколько панелей, регион, нарисованный на одной панели, будет также представлен и на других панелях.

Не все диаграммы принимают регионы. Вот некоторые типы диаграмм, принимающих регионы: графики (линейные, рассеяния, пузырьковые, зависимостей от времени и так далее), диаграммы SPLOM и собраний. Эти регионы рисуются в пространстве X,Y и поэтому не могут быть определены в одномерных, трехмерных или анимированных графиках. На диаграммах с несколькими панелями регионы появляются на всех панелях. В случае матрицы диаграмм рассеяния (SPLOM) соответствующий регион появится на соответствующих верхних графиках, но не на диагональных графиках, поскольку на них выводится только одно поле масштаба.

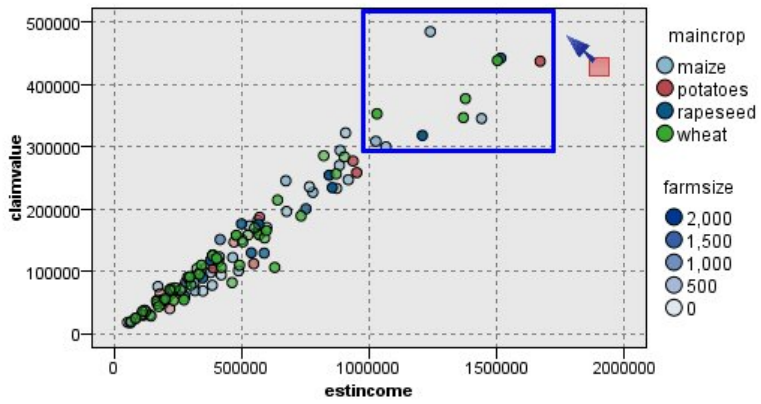


Рисунок 55. Определение региона высоких заявленных значений

Определение регионов

Всякий раз, когда вы определяете регион, вы создаете группировку значений. По умолчанию каждый новый регион называется*Region*<*N*>, где *N* соответствует числу уже созданных регионов.

После определения региона по нему можно щелкнуть правой кнопкой мыши, чтобы получить некоторые базовые ярлыки. Однако, щелкнув правой кнопкой мыши в самом регионе (а не по линии) можно вывести множество других ярлыков для задач, таких как переименование, удаление или генерирование узлов выбора и извлечения конкретно для этого региона.

Можно выбрать поднаборы записей по признаку их включения в конкретный регион или в один из нескольких регионов. Информацию о регионах можно также применить для записи, сгенерировав узел извлечения для присвоения флага записям на основе их включения в регион. Дополнительную информацию смотрите в разделе “Генерирование узлов из диаграмм” на стр. 267.

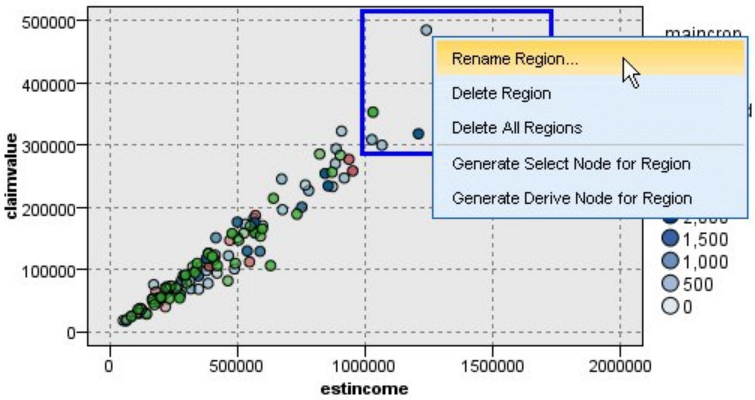


Рисунок 56. Изучение региона высоких заявленных значений

Чтобы определить регионы:

1. Убедитесь, что вы работаете в режиме изучения. В меню выберите **Просмотр > Режим изучения**.
2. На панели инструментов режима изучения нажмите кнопку Нарисовать регион.

Рисунок 57. Кнопка панели инструментов Нарисовать регион

3. На диаграмме, принимающей регионы, проведите указателем при нажатой кнопке мыши, чтобы очертить прямоугольный регион.

Редактирование, переименование и удаление регионов

Свойства существующих регионов можно отредактировать в диалоговом окне Изменить регионы диаграммы или посредством контекстного меню на самой диаграмме.

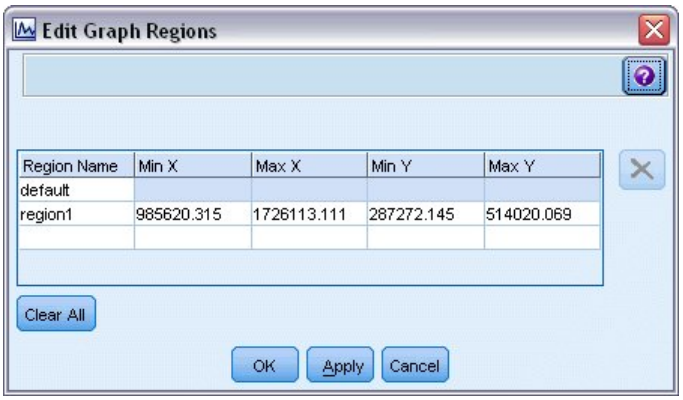


Рисунок 58. Задание свойств для определенных регионов

Чтобы отредактировать регионы:

1. Убедитесь, что вы работаете в режиме изучения. В меню выберите **Просмотр > Режим изучения**.
2. На панели инструментов режима изучения нажмите кнопку Нарисовать регион.
3. В меню выберите **Изменить > Регионы диаграммы**. Откроется диалоговое окно Изменить регионы диаграммы.
4. Если на диаграмме несколько полей (например, на диаграммах SPLOM), нужно будет определить поле для региона в столбцах *Поле A* и *Поле B*.
5. Добавьте новый регион в новую строку, введя имя, выбрав имена полей (если это требуется) и определив для каждого поля границы максимума и минимума. Чтобы начать новую строку, нажмите клавишу Enter.
6. Отредактируйте существующие границы региона, скорректировав значения **Мин** и **Макс** для *A* и *B*.
7. Переименуйте регион, изменив имя этого региона в таблице.
8. Удалите регион, выбрав соответствующую строку в таблице и нажав кнопку удаления.
9. Нажмите кнопку **OK** , чтобы применить внесенные изменения и закрыть это диалоговое окно.

Примечание: Другой вариант - можно удалить и переименовать регионы непосредственно на диаграмме, щелкнув правой кнопкой мыши по линии региона и выбрав нужную опцию в контекстных меню.

Использование помеченных элементов

Такие элементы, как столбики, сектора и точки, можно пометить на любой диаграмме. Линии, области и поверхности можно пометить только на диаграммах временных зависимостей, множественных зависимостей и диаграммах оценки, поскольку линии в этих случаях относятся к полям. Всякий раз, помечая элемент, вы фактически выделяете все данные, которые он представляет. На любой диаграмме, где одно наблюдение представлено в нескольких местах (например, на диаграмме SPLOM), нанесение меток синонимично рисованию кистью. Элементы можно помечать на диаграммах и даже в полосах в регионах.

Всякий раз, когда вы помечаете элемент, а затем возвращаетесь обратно в режим редактирования, нанесенные метки остаются видимы.

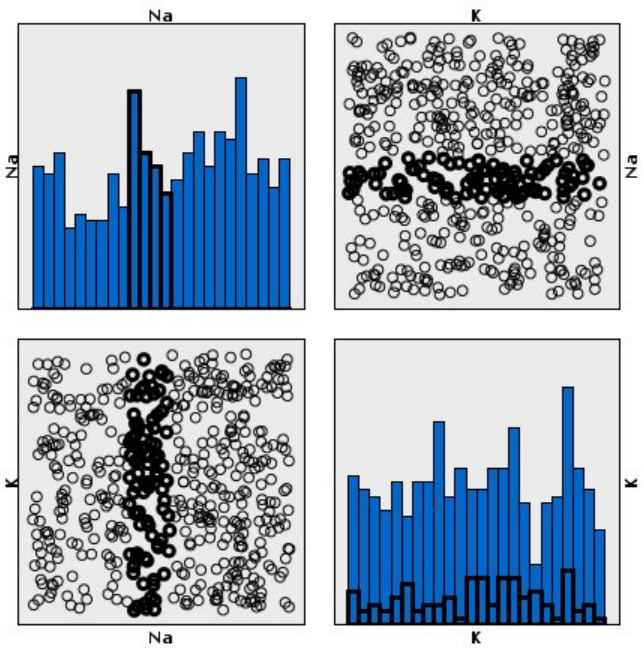


Рисунок 59. Нанесение меток на элементы в SPLOM

Наносить метки на элементы и снимать их с элементов можно, щелкая по элементам на диаграмме. При первом щелчке по элементу для нанесения метки он выделяется толстой цветной рамкой, указывающей, что он помечен. Если снова щелкнуть по элементу, цветная рамка исчезнет, и элемент снова станет непомеченным. Несколько элементов можно пометить, щелкнув по ним при нажатой клавише Ctrl или проведя по каждому из элементов, которые вы хотите пометить, указателем мыши при помощи "волшебной палочки". Учтите, что если щелкнуть кнопкой мыши по другой области или по элементу без нажатой клавиши Ctrl, все ранее помеченные элементы будут очищены.

Из помеченных на диаграмме элементов можно сгенерировать узлы выбора и извлечения. Дополнительную информацию смотрите в разделе “Генерирование узлов из диаграмм”.

Чтобы пометить элементы:

1. Убедитесь, что вы работаете в режиме изучения. В меню выберите **Просмотр > Режим изучения**.
2. На панели инструментов режима изучения нажмите кнопку Пометить элементы.
3. Щелкните по нужному элементу или проведите указателем при нажатой кнопке мыши, чтобы очертить линией регион, содержащий несколько элементов.

Генерирование узлов из диаграмм

Одна из самых мощных функциональных возможностей, предлагаемых диаграммами IBM SPSS Modeler - это возможность генерирования узлов из диаграмм или на основе вариантов выбора на диаграмме. Например, из диаграммы временного графика можно сгенерировать узлы извлечения и выбора на основе выбора или региона данных, создав тем самым эффективные подмножества данных. Например, эту мощную возможность можно использовать для обнаружения и исключения выбросов.

Всегда, когда можно начертить полосы, можно также сгенерировать и узел извлечения. На диаграммах с двумя количественными осями можно сгенерировать узлы извлечения или выбора из нарисованных на диаграмме регионов. На диаграмме с помеченными элементами можно сгенерировать узлы извлечения,

узлы выбора и (в некоторых случаях) узлы фильтра из этих элементов. Поддержка генерирования узлов балансировки включается для любой диаграммы, показывающей распределение количеств.

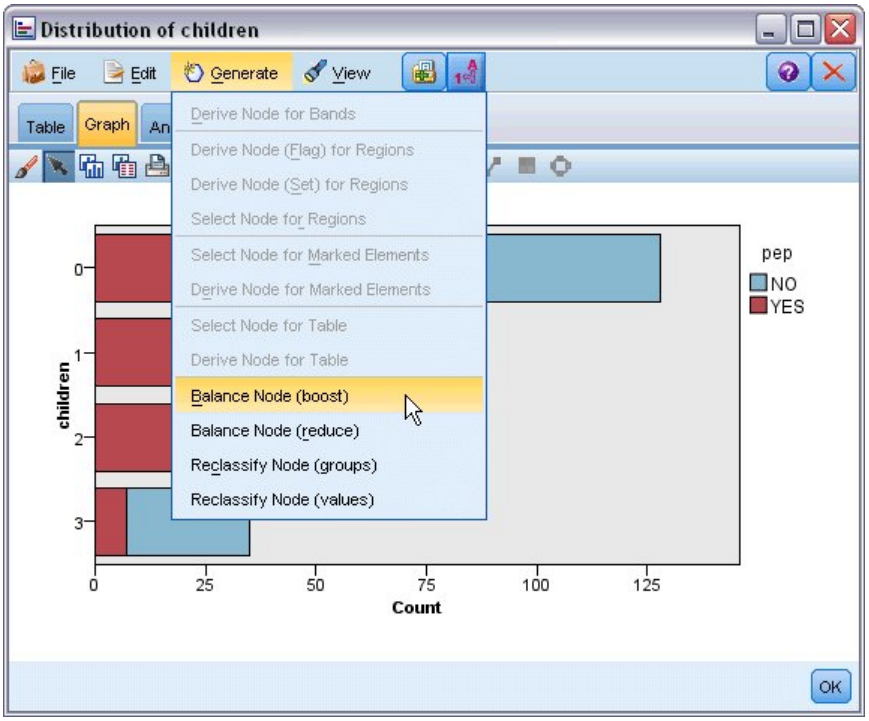


Рисунок 60. Диаграмма с выведенным меню генерирования

Всякий раз при генерировании узла он помещается непосредственно на холст потока, поэтому его можно соединить с существующим потоком. Из диаграмм можно сгенерировать узлы выбора, извлечения, балансировки, фильтра и переклассификации.

Узлы выбора

Узлы выбора могут быть сгенерированы для проверки включения записей в регион и исключения всех записей, попадающих за пределы региона, или обращения для обработки нисходящего потока.

- **Для полос**. Можно сгенерировать узел выбора, включающий записи в данную полосу или исключающий их из нее. Опция **Узел выбора только для полос** доступна только в контекстных меню, поскольку требуется выбор, какую полосу использовать на узле выбора.
- **Для регионов**. Можно сгенерировать узел выбора, включающий записи в регион или исключающий их из него.
- **Для помеченных элементов**. Можно сгенерировать узлы выбора для захвата записей, соответствующих помеченным элементам или ссылкам на диаграммы в Web.

Узлы извлечения

Узлы извлечения можно сгенерировать из регионов, полос и помеченных элементов. Узлы извлечения могут создаваться всеми диаграммами. В случае диаграмм оценки открывается диалоговое окно для выбора модели. В случае диаграмм в Web возможны опции **Узел извлечения (“And”)** и **Узел извлечения (“Or”)**.

- **Для полос**. Можно сгенерировать узел извлечения, создающий категории для каждого интервала, помеченного на оси, используя имена полос, перечисленных в диалоговом окне Изменить полосы, как имена категорий.
- **Для регионов**. Можно сгенерировать узел извлечения (**Извлечение в качестве флага**), создающий флаговое поле *in_region* с флагами, задаваемыми как *T* для записей в любом регионе и как *F* для записей за

пределами всех регионов. Можно также сгенерировать узел извлечения (**Извлечение в качестве набора**), создающий набор со значением для каждого региона с новым полем *region* для каждой записи, которое принимает в качестве своего значения имя региона, куда попадают записи. Записи, попадающие за пределы всех регионов, получают имя региона по умолчанию. Имена значений становятся именами регионов, перечисленных в диалоговом окне редактирования регионов.

- **Для помеченных элементов**. Можно сгенерировать узел извлечения, вычисляющий флаг, представляющий собой *True* для всех помеченных элементов и *False* для всех остальных записей.

Узлы балансировки

Узлы балансировки могут быть сгенерированы для исправления дисбалансов в данных, например, сокращения частоты встречаемости распространенных значений (при помощи опции меню **Узел балансировки (сокращение)**) или бустинга вхождений редко встречающихся значений (при помощи опции меню **Узел балансировки (повышение)**). Поддержка генерирования узлов балансировки включается для любой диаграммы, показывающей распределение количеств, например, для гистограммы, точечной диаграммы, диаграммы собрания, столбиковой диаграммы, круговой диаграммы частот и диаграммы с несколькими графиками.

Узлы фильтра

Узлы фильтра могут быть сгенерированы с целью переименования или фильтрации полей на основе линий или узлов, помеченных на диаграмме. В случае диаграмм оценки лучшая линия совпадений узел фильтра не генерирует.

Узлы переклассификации

Узлы переклассификации могут быть сгенерированы с целью перекодирования значений. Эта опция используется для диаграмм распределения. Сгенерировав узел переклассификации для **групп**, можно перекодировать конкретные значения выводимого поля в зависимости от их включения в группу (группы выбираются на вкладке **Таблицы** кнопкой мыши при нажатой клавише Ctrl). Кроме того, сгенерировав узел переклассификации для **значений**, можно перекодировать данные в существующий набор с многочисленными значениями, например, переклассифицировать данные в стандартный набор значений, чтобы объединить слиянием финансовые данные различных компаний для анализа.

Примечание: Если значения являются предопределенными, их можно передать в IBM SPSS Modeler как плоский файл и применить распределение для вывода всех значений. Затем для этого поля можно сгенерировать узел переклассификации (значений) непосредственно из диаграммы. Тогда все значения назначения будут помещены в столбец (выпадающий список) узла переклассификации *Новые значения*.

При задании опций для узла переклассификации таблица допускает явное отображение старого набора значений на новые, заданные вами:

- **Исходное значение.** Этот столбец содержит существующие значения для выдранных полей.
- **Новое значение.** Введите при помощи этого столбца новые значения категорий или выберите значение из выпадающего списка. При автоматическом генерировании узла переклассификации с помощью значений из диаграммы распределения эти значения включаются в выпадающий список. Это позволяет быстро отобразить существующие значения на известный набор значений. Например, медицинские организации иногда группируют диагнозы по разному на основе сети или локали. После их слияния или приобретения всем сторонам потребуется переклассифицировать новые или даже существующие данные сообразным способом. Вместо ввода вручную каждого значения назначения из очень длинного списка вы можете передать главный список значений в IBM SPSS Modeler, вызвать диаграмму распределения для поля *Диагноз* и сгенерировать для этого поля узел переклассификации (значений) непосредственно из диаграммы. Этот процесс сделает все значения назначения поля Диагноз доступными в выпадающем списке Новые значения.

Дополнительную информацию об узле переклассификации смотрите в разделе “Задание опций для узла переклассификации” на стр. 153.

Генерирование узлов из диаграмм

Сгенерировать узлы можно при помощи меню Создать в окне вывода диаграммы. Сгенерированный узел будет помещен на холст потока. Для использования этого узла соедините его с существующим потоком.

Чтобы сгенерировать узел из диаграммы:

1. Убедитесь, что вы работаете в режиме изучения. В меню выберите **Просмотр > Режим изучения**.
2. На панели инструментов режима изучения нажмите кнопку Регион.
3. Определите полосы, регионы или все помеченные элементы, необходимые для генерирования нужного узла.
4. В меню Создать выберите тип узла, который вы хотите создать. Будут доступны только те узлы, которые можно сгенерировать.

Примечание: Другой вариант - узлы можно также сгенерировать непосредственно из диаграммы, щелкнув правой кнопкой мыши и выбрав нужную опцию генерирования в контекстных меню.

Редактирование визуализаций

В то время как режим исследования позволяет аналитически исследовать данные и значения, представленные визуализацией, режим редактирования позволяет менять макет и внешний вид визуализации. Например, можно изменить шрифты и цвета, чтобы они соответствовали руководству по стилю оформления вашей организации. Чтобы войти в этот режим, выберите в меню **Вид> Режим редактирования** (или щелкните по значку на панели инструментов).

В режиме редактирования имеется несколько панелей инструментов, которые влияют на разные аспекты внешнего вида визуализации. Можно скрыть неиспользуемые панели, чтобы увеличить пространство диалогового окна, в котором показана диаграмма. Чтобы выбрать панель инструментов или отменить такой выбор, щелкните по имени соответствующей панели инструментов в меню Вид.

Примечание: Чтобы добавить дополнительные подробности визуализации, можно применить заголовок, сноску и метки осей. Дополнительную информацию смотрите в разделе “Добавление заголовков и сносок” на стр. 281.

Существует несколько параметров для редактирования визуализаций в **Режиме правки**. Вы можете:

- Редактировать и форматировать текст.
- Менять цвет заливки, прозрачность и штриховку рамок и графических элементов.
- Менять цвет и пунктир границ и линий.
- Производить вращение и менять форму и пропорции элементов точек.
- Менять размер графических элементов (таких, как столбцы и точки).
- Корректировать пространство вокруг элементов, используя внешние и внутренние поля.
- Указывать формат чисел.
- Менять настройки осей и шкал.
- Сортировать, исключать и сокращать категории по оси категорий.
- Устанавливать положение панелей.
- Преобразовывать системы координат.
- Менять статистики, типы графических элементов и модификаторы коллизий.
- Менять положение легенды.
- Применять таблицы стилей визуализации.

Далее описывается, как выполнять эти задачи. Также рекомендуется прочитать общие правила редактирования диаграмм.

Переключение в режим редактирования

Выберите в меню:

Вид > Режим редактирования

Общие правила редактирования визуализаций

Режим редактирования

Любое редактирование выполняется в режиме редактирования. Чтобы перейти в режим редактирования, выберите в меню:

Вид > Режим редактирования

Выделенный фрагмент

Доступные для редактирования параметры зависят от того, что выделено. В зависимости от того, что выделено, включаются разные параметры панели инструментов и свойства палитры. Только включенные параметры применимы к текущему выделению. Например, если выделена ось, в палитре свойств доступны вкладки Шкала, Основные деления и Вспомогательные деления.

Вот несколько советов по выделению элементов в визуализации:

- Щелкните по элементу мышью, чтобы выделить его.
- Выделите одним щелчком мыши графический элемент (например, точки на диаграмме рассеяния или столбцы на столбчатой диаграмме). После первоначального выделения щелкните повторно, чтобы сузить выделение до группы графических элементов или одного графического элемента.
- Нажмите клавишу Esc, чтобы отменить выделение.

Палитры инструментов

При выделении какого-либо элемента визуализации, палитры изменяются и автоматически приводятся в соответствие выделенному. Палитры содержат управляющие элементы для редактирования выделенного. Палитры могут являться панелями инструментов или панелями с несколькими управляющими элементами и вкладками. Палитры можно скрыть, поэтому стоит убедиться, что палитра показана. Проверьте в меню Вид, какие палитры сейчас показаны.

Можно изменить местоположение палитры, щелкнув и перетащив ее в пустое место на палитре панели инструментов или слева от других палитр. Визуальная обратная связь дает вам знать, где можно разместить палитру. Для палитр, не являющихся панелями инструментов, можно щелкнуть по кнопке Закрыть, чтобы скрыть палитру, или по кнопке Расстыковать, чтобы палитра была показана в новом окне. Нажмите кнопку Справка для вывода справки по определенной палитре.

Автоматические настройки

Некоторые настройки содержат параметр **-авто-**. Это означает, что применяются автоматические значения. То, какие автоматические настройки используются, зависит от конкретной визуализации и значений данных. Вы можете ввести значение, отменяющее автоматическую настройку. Если вы хотите восстановить автоматическую настройку, удалите текущее значение и нажмите клавишу Ввод. Эта настройка снова выводит **-auto-**.

Удаление/скрытие элементов

Вы можете удалять/скрывать различные элементы визуализаций. Например, можно скрыть легенду или метку оси. Чтобы удалить элемент, выделите его и нажмите клавишу Delete. Если удаление элемента недопустимо, ничего не произойдет. Если вы удалили элемент случайно, нажмите Ctrl+Z, чтобы отменить удаление.

Состояние

Некоторые панели инструментов отражают состояние текущего выделения, а другие - нет. Палитра свойств всегда отражает состояние. Если панель инструментов *не* отражает состояние, это упомянуто в теме, описывающей данную панель инструментов.

Редактирование и форматирование текста

Вы можете редактировать имеющийся текст и менять форматирование всего текстового блока. Обратите внимание, что редактировать текст, который напрямую связан со значениями данных, невозможно. Например, невозможно редактировать метки делений, поскольку содержимое меток извлекается из исходных данных. Можно однако форматировать любой текст в визуализации.

Как отредактировать имеющийся текст

1. Щелкните дважды по текстовому блоку. Таким образом будет выделен весь текст. В это время все панели инструментов отключатся, так как изменение какой-либо другой части визуализации во время редактирования текста - невозможно.
2. Введите текст для замены существующего. Также можно повторно щелкнуть по тексту, чтобы в нем появился курсор. Расположите указатель там, где вы хотите ввести дополнительный текст.

Как отформатировать текст

1. Выделите рамку, содержащую текст. Не нужно дважды щелкать по тексту.
2. Отформатируйте текст, используя панель шрифтов. Если эта панель инструментов не включена, убедитесь, что выделена только *рамка* , содержащая текст. Если выделен сам текст, эта панель инструментов будет отключена.

Вы можете менять шрифт:

- Цвет
- Семейство (например, Arial или Verdana)
- Размер (единицей размера является точка (pt), если вы не укажете другую единицу, например, пику (pc))
- Толщина
- Выравнивание по отношению к текстовой рамке

Форматирование применяется ко всему тексту, находящемуся в рамке. Невозможно изменить форматирование отдельных букв или слов в любом отдельном блоке текста.

Изменение цветов, штриховки, пунктира и прозрачности

Многие элементы визуализации имеют заливку и границу. Наиболее очевидным примером является столбец столбчатой диаграммы. Цвет столбцов является цветом заливки. Также столбцы могут иметь сплошную черную границу.

Есть и другие, менее бросающиеся в глаза элементы, имеющие цвета заливки. Если цвет заливки прозрачен, вы можете и не догадываться о наличии заливки. Например, обратите внимание на текст метки оси. Кажется, что он является "парящим" текстом, но на самом деле он содержится в рамке, имеющей прозрачный цвет заливки. Вы увидите эту рамку, выделив метку оси.

Любая рамка визуализации может иметь заливку и стиль границы, включая рамку вокруг всей визуализации. Кроме того, любая заливка имеет связанный с ней уровень непрозрачности/прозрачности, который может регулироваться.

Как изменять цвета, штриховку, пунктир и прозрачность

1. Выберите элемент для форматирования. Например, выберите столбцы в столбчатой диаграмме или рамку, содержащую текст. Если визуализация разбита категориальной переменной или полем, вы также можете выбрать группу, соответствующую отдельной категории. Это позволяет менять стиль по умолчанию, назначенный данной группе. Например, можно менять цвет одной из составных групп на составной столбчатой диаграмме.
2. Чтобы изменить цвет заливки, цвет границы или штриховку заливки, используйте панель цветов.

   Примечание: Эта панель инструментов не отображает состояния текущего выбора.

   Для изменения цвета или заливки можно щелкнуть по кнопке для выбора показанного параметра или щелкнуть по раскрывающейся стрелке для выбора другого параметра. Что касается цветов, обратите внимание на цвет, который выглядит как белый с красным, с диагональной полосой. Это прозрачный цвет. Вы можете использовать его, например, для скрытия границ столбцов гистограммы.

   - Первая кнопка управляет цветом заливки. Если цвет связан с количественным или порядковым полем, эта кнопка изменяет цвет заливки для цвета, связанного с самым высоким значением в данных. Также можно использовать таблицу Цвет на палитре свойств для изменения цвета, связанного с самым низким значением и отсутствующими данными. Цвет элементов изменится инкрементально с низкого до высокого, поскольку значения базовых данных вырастут.
   - Вторая кнопка управляет цветом границы.
   - Третья кнопка управляет штриховкой заливки. Штриховка заливки использует цвет границы. Поэтому она видна только в том случае, если виден цвет границы.
   - Четвертым управляющим элементом является ползунок и текстовое поле, которые управляют прозрачностью цвета заливки и штриховки. Меньший процент означает меньшую непрозрачность и большую прозрачность. 100% - полностью непрозрачный.
3. Чтобы изменить пунктир границы или линии, используйте панель линий.

   Примечание: Эта панель инструментов не отображает состояния текущего выбора.

   Как и на других панелях инструментов, вы можете щелкнуть кнопку для выбора показанного параметра или щелкнуть раскрывающуюся стрелку, чтобы выбрать другой параметр.

Вращение и изменение формы и пропорций элементов точек

Вы можете вращать элементы точек, назначать им другую предустановленную форму или менять пропорции (отношение ширины к высоте).

Как изменять элементы точек

1. Выделите элементы точек. Производить вращение и менять форму и пропорции отдельных элементов точек невозможно.
2. Чтобы изменять точки, используйте панель символов.
   - Первая кнопка позволяет менять форму точек. Щелкните раскрывающуюся стрелку и выберите предустановленную форму.
   - Вторая кнопка позволяет вращать точки, устанавливая для них определенное положение по окружности. Щелкните по раскрывающейся стрелке и перетащите образец в нужное положение.
   - Третья кнопка позволяет менять пропорции. Щелкните раскрывающуюся стрелку, затем щелкните и тащите появившийся прямоугольник. Форма прямоугольника представляет пропорции.

Изменение размера графических элементов

Размер графических элементов визуализации можно изменять. Графическими элементами, помимо прочих, являются столбцы, линии и точки. Если размер графического элемента определяется переменной или полем, указанный размер является *минимальным* .

Как изменить размер графических элементов

1. Выделите графический элемент, размер которого хотите изменить.

2. Используйте ползунок или введите определенный размер для параметра, доступного на панели символов. Единицами размера являются пиксели, если вы не укажете другие единицы (см. ниже полный список сокращений единиц). Также можно указать процент (например, 30%), и графический элемент будет использовать указанный процент доступного пространства. Доступное пространство зависит от типа графического элемента и определенной визуализации.

Таблица 35. Сокращения допустимых единиц

Сокращение	**Единица**
см	сантиметр
дюймов	дюйм
mm	миллиметр
pc	пика
pt	пункт
px	пиксель

Задание внешних и внутренних полей

Если пространстве вокруг рамки или внутри нее слишком мало или слишком велико, настройки внешних и внутренних полей рамки можно изменить. **Внешним полем** является пространство между рамкой и окружающими ее элементами. **Внутренним полем** является пространство между границей рамки и *содержимым* рамки.

Как задать внешние и внутренние поля

1. Выделите рамку, для которой вы хотите задать внешние и внутренние поля. Это может быть текстовая рамка, рамка вокруг легенды или даже рамка с данными, показывающая графические элементы (такие, как столбцы и точки).
2. Используйте вкладку Поля на палитре свойств, чтобы задать параметры. Все размеры указываются в пикселях, если не выбраны другие единицы (например, сантиметры или дюймы).

Форматирование чисел

Вы можете задать формат для чисел в метках делений на количественной оси или для меток значений данных, содержащих числа. Например, можно задать, чтобы числа, показанные на метках делений, показывались бы в тысячах.

Как задать форматы чисел

1. Выделите метки делений непрерывной оси или метки значений данных, если они содержат числа.
2. Перейдите на вкладку **Формат** на палитре свойств.
3. Выберите нужные вам опции форматирования чисел:

   Префикс . Символ, который будет показан перед числами. Например, введите знак доллара ($), если числа - это зарплаты в долларах США.

   Суффикс . Символ, который будет показан после чисел. Например, введите знак процента (%), если числа представляют собой проценты.

   Мин. количество цифр. Минимальное количество разрядов для вывода на экран в целочисленной части десятичного представления. Если действительное значение не содержит минимального количества разрядов, целочисленная часть значения будет дополнена нулями.

   Макс. количество цифр. Максимальное количество разрядов для вывода на экран в целочисленной части десятичного представления. Если действительное значение превышает максимальное количество разрядов, целочисленная часть значения будет заменена звездочками.

Мин. количество цифр. Минимальное количество разрядов для вывода на экран в десятичной части десятичного или экспоненциального представления. Если действительное значение не содержит минимального количества разрядов, десятичная часть значения будет дополнена нулями.

Макс. количество цифр. Максимальное количество разрядов для вывода на экран в десятичной части десятичного или экспоненциального представления. Если действительное значение превышает максимальное количество разрядов, десятичная часть округляется до соответствующего количества разрядов.

Экспоненциальное представление . Показывать ли числа в экспоненциальном представлении. Экспоненциальное представление полезно для очень больших и очень малых чисел. **-auto-** позволяет прикладной программе самой определить, когда нужно использовать экспоненциальное представление.

Масштабирование . Коэффициент масштабирования - это число, на которое делится исходное значение. Используется, когда числа большие, а метки делений не должны быть слишком длинными. Если вы меняете формат чисел меток делений, не забудьте отредактировать заголовок оси, чтобы объяснить, как следует интерпретировать числа. Например, предположим, что на количественной оси отображаются зарплаты с метками 30 000, 50 000, 70 000. Можно ввести коэффициент масштабирования `1000` , чтобы метки превратились в 30, 50 и 70, соответственно. Затем следует отредактировать заголовок количественной оси, добавив в него текст "в тысячах".

Скобки для отрицательных значений . Следует ли показывать скобки вокруг отрицательных значений.

Группирование . Показывать ли символ между группами разрядов. Символ, который будет использоваться для разделения групп разрядов зависит от региональных параметров вашего компьютера.

Изменение настроек осей и шкал

Можно использовать несколько параметров для изменения осей и шкал.

Как изменить оси и параметры шкалы

1. Выделите любую часть оси (например, метку оси или метки делений).
2. Используйте вкладки Шкала, Основные деления и Вспомогательные деления в палитре свойств для изменения настроек осей и шкал.

Вкладка Шкала

Примечание: Вкладка Шкала не появляется для графиков, когда данные предварительно объединены (например, для гистограмм).

Тип. Здесь указывается линейная шкала или преобразованная. Преобразования шкалы помогают лучше понять данные, а также делать предположения, необходимые для статистических выводов. В диаграммах рассеяния преобразованную шкалу можно использовать, если связь между зависимой переменной или полем и независимыми переменными или полями нелинейная. Преобразования также можно применять, чтобы сделать несимметричную гистограмму более симметричной и более похожей на нормальное распределение. Обратите внимание, что изменятся только шкала, на которой показаны данные; сами данные не изменяются.

- **линейная.** Задает линейную шкалу без преобразований.
- **логарифмическая.** Задает логарифмические шкалы с десятичным логарифмом. Чтобы вмещать нулевые и отрицательные значения, это преобразование использует модифицированную версию логарифмической функции. Функция "безопасного логарифмического преобразования" задается формулой `sign(x) * log(1 + abs(x))`. Таким образом, значение `safeLog(-99)` равно:

  ```
  sign(-99) * log(1 + abs(-99)) = -1 * log(1 + 99) = -1 * 2 = -2
  ```

- **степенная.** Задает шкалу со степенным преобразованием, с использованием показателя степени 0,5. Для обработки отрицательных значений, это преобразование использует модифицированную версию степенной функции. Функция "безопасного степенного преобразования" задается формулой `sign(x) * pow(abs(x), 0.5)`. Таким образом, значение `safePower(-100)` равно:

  ```
  sign(-100) * pow(abs(-100), 0.5) = -1* pow(100, 0.5) = -1 * 10 = -10
  ```

Мин/Макс/Удобное наименьшее/Удобное наибольшее. Задает диапазон шкалы. Выбор **Удобного наименьшего** и **Удобного наибольшего** позволяет приложению выбрать соответствующую шкалу на основании данных. Минимум и максимум называются "удобными", потому что они обычно являются целыми значениями, большими или меньшими максимального и минимального значений данных. Например, если данные составляют диапазон от 4 до 92, удобные наименьшее и наибольшее для шкалы могут составлять 0 и 100, а не минимум и максимум действительных данных. Будьте внимательны, чтобы не установить диапазон, который окажется слишком узким и скроет важные значения. Также учтите, что установить определенный минимум и максимум невозможно, если выбран параметр **Включить ноль**.

Нижнее поле/Верхнее поле. Создает поля на нижнем и верхнем концах оси. Поля отсчитываются перпендикулярно выделенной оси. Единицей является пиксель, если не выбраны другие единицы (например, сантиметры или дюймы). Например, если установить **Верхнее поле** равным 5 для вертикальной оси, вдоль рамки данных появится горизонтальное поле шириной 5 пикселей.

Перевернутая. Задает, является ли шкала перевернутой.

Включить ноль. Указывает, что шкала должна содержать ноль. Этот параметр обычно используется для столбчатых диаграмм, чтобы обеспечить начало столбцов от 0, а не от значения, близкого к высоте наименьшего столбца. Если выбран этот параметр, **Мин** и **Макс** отключены, так как установка пользовательского минимума и максимума для диапазона шкалы в этом случае невозможна.

Вкладка Основные деления/Вспомогательные деления

Деления или **засечки** представляют собой линии, показанные на оси. Они указывают значения в определенных интервалах или категориях. **Основными делениями** являются засечки с метками. Они длиннее других засечек. **Вспомогательными делениями** являются засечки, показанные между основными делениями. Некоторые параметры являются специфическими для типа делений, но большинство параметров доступны как для основных, так и для вспомогательных делений.

Показать деления. Задает, будут ли показаны на диаграмме основные или вспомогательные деления.

Показывать линии сетки. Задает, будут ли показаны линии сетки на основных или вспомогательных делениях. **Линии сетки** представляют собой линии, пересекающие всю диаграмму от оси до оси.

Положение. Указывает положение делений относительно оси.

Длина. Задает длину делений. Единицей является пиксель, если не выбраны другие единицы (например, сантиметры или дюймы).

База. *Применяется только к основным рискам.* Задает значение, на котором появляется первое основное деление.

Дельта. *Применяется только к основным рискам.* Указывает разность между основными делениями. То есть, основные деления появляются на каждом значении, кратном *n* , где *n* представляет собой значение дельты.

Деления. *Применяется только к вспомогательным рискам.* Задает количество делений вспомогательных делений между основными делениями. Количество вспомогательных делений равно количеству делений минус один. Например, допустим, что основные деления расположены на 0 и 100. Если для количества вспомогательных делений ввести 2, будет использоваться *одно* вспомогательное деление на 50, делящее диапазон 0–100 и создающее *два* раздела.

Редактирование категорий

Редактировать категории на категориальной оси можно несколькими способами:

- Путем изменения сортировки порядка вывода на экран категорий;
- Путем исключения исключать определенных категорий;

- Путем добавления категорий, которых нет в наборах данных.
- Сворачивая/объединяя небольшие категории в одну категорию.

Как изменить порядок сортировки категорий

1. Выделите категориальную ось. Палитра Категории показывает категории на оси.

   Примечание: Если палитра невидима, убедитесь, что она включена. В меню Вид IBM SPSS Modeler выберите **Категории**.
2. В палитре Категории в раскрывающемся списке выберите вариант сортировки:

   Пользовательские. Сортирует категории на основании того, в каком порядке они показаны в палитре. Используйте клавиши со стрелками для перемещения категорий в начало списка, вверх, вниз и в конец списка.

   Данные. Сортирует категории на основании того, в каком порядке они появляются в наборе данных.

   Имя. Сортирует категории по алфавиту, используя названия, показанные в палитре. Это может быть значение или метка, в зависимости от того, что выбрано на панели инструментов: вывод на экран значений или вывод на экран меток.

   Значение. Сортирует категории по лежащему в основе значению данных, используя значения, показанные в палитре в скобках. Только источники данных с метаданными (например, файлы данных IBM SPSS Statistics) поддерживают этот параметр.

   Статистики Сортирует категории на основании статистик, рассчитанных для каждой категории. Примеры статистик включают количества, проценты и средние. Этот параметр доступен только в том случае, если в диаграмме используются статистики.

Как добавить категорию

По умолчанию доступны только те категории, показанные в наборе данных. При необходимости в визуализацию можно добавлять категории.

1. Выделите категориальную ось. Палитра Категории показывает категории на оси.

   Примечание: Если палитра невидима, убедитесь, что она включена. В меню Вид IBM SPSS Modeler выберите **Категории**.
2. В палитре Категории нажмите кнопку добавления категорий:

Рисунок 61. Кнопка Добавить категорию

3. В диалоговом окне Добавить новую категорию введите имя категории.
4. Щелкните по **OK**.

Как исключить определенные категории

1. Выделите категориальную ось. Палитра Категории показывает категории на оси.

   Примечание: Если палитра невидима, убедитесь, что она включена. В меню Вид IBM SPSS Modeler выберите **Категории**.
2. В палитре Категории выберите название категории в списке Включить и нажмите кнопку X. Чтобы переместить категорию обратно, выделите ее в списке Исключены, а затем щелкните по стрелке справа от списка.

Как свернуть/объединить небольшие категории

Категории, количество наблюдений в которых настолько мало, что такие категории не стоит показывать по отдельности, можно объединять. Например, если в круговой диаграмме много категорий, стоит подумать о том, чтобы свернуть категории, процент наблюдений в которых менее десяти. Свернуть категории можно

только для аддитивных статистик. Например, невозможно складывать средние значения, поскольку они не являются аддитивными. Следовательно, если в качестве статистики используется среднее значение, объединение/сворачивание категорий, невозможно.

1. Выделите категориальную ось. Палитра Категории показывает категории на оси.

   Примечание: Если палитра невидима, убедитесь, что она включена. В меню Вид IBM SPSS Modeler выберите **Категории**.

2. В палитре Категории выберите **Свернуть** и задайте процент. Все категории, для которых процент от общего количества меньше указанного числа, будут объединены в одну категорию. Процент основываться на статистиках, показанных на диаграмме. Сворачивание доступно только для статистик, основанных на количестве и сложении (суммировании).

Изменение ориентации панелей

Если вы используете панели для визуализации, вы можете изменить их ориентацию.

Как изменить ориентацию панелей

1. Выделите любую часть визуализации.
2. Перейдите на вкладку **Панели** на палитре свойств.
3. Выберите параметр из списка **Компоновка** :

Таблица. Компонует панели в виде таблицы, в которой каждому отдельному значению назначены строка или столбец.

Транспонированные Компонует панели в виде таблицы, но также меняет исходные строки и столбцы. Этот вариант не является эквивалентом транспонирования самой диаграммы. Обратите внимание, что ось *x* и ось *y* не меняются при выборе этого варианта.

Список. Компонует панели в виде списка, в котором каждая ячейка представляет сочетание значений. Столбцам и строкам больше не назначаются отдельные значения. Этот вариант позволяет при необходимости разрывать панели.

Преобразование систем координат

Многие визуализации показаны в плоской прямоугольной системе координат. Систему координат при необходимости можно преобразовать. Например, можно применить к системе координат полярное преобразование, добавить эффекты косоугольного отбрасывания теней или транспонировать оси. Также можно отменить любые из этих преобразований, если они уже применены к текущей визуализации. Например, круговая диаграмма начерчена в полярной системе координат. Вы можете отменить преобразование к полярным координатам и вывести круговую диаграмму как как одну составную столбчатую диаграмму в системе декартовых координат.

Как преобразовать систему координат

1. Выделите систему координат, которую нужно преобразовать. Выделение системы координат производится путем выделения рамки вокруг диаграммы.
2. Перейдите на вкладку **Координаты** на палитре свойств.
3. Выберите преобразования, которые хотите применить к системе координат. Также можно отменить выбор преобразования, чтобы отменить его.

   Транспонирование. Изменение ориентации осей называется **транспонированием** . Оно похоже на перемене мест вертикальной и горизонтальной осей в двухмерной визуализации.

   Полярное . Полярное преобразование размещает графические элементы под определенным углом и на определенном расстоянии от центра диаграммы. Круговая диаграмма - это одномерная визуализация с преобразованием к полярным координатам, при которой отдельные столбцы рисуются под конкретным углом. Радарная диаграмма - это двумерная визуализация с преобразованием к полярным координатам,

при которой графические элементы изображаются под конкретными углами и на определенном расстоянии от центра изображения. Трехмерная визуализация будет также содержать дополнительное измерение глубины.

Косоугольное . Косоугольное преобразование добавляет трехмерный эффект к графическим элементам. Это преобразование добавляет глубину графическим элементам, но эта глубина является исключительно декоративной. На него не влияют конкретные значения данных.

Сохранение пропорций . Применение преобразования сохранения пропорций означает, что одно и то же расстояние по каждой шкале представляет одно и то же различие значений данных. Например, 2 см на обеих шкалах представляют разницу 1000.

Врезка % до преобразования . Если после преобразования оси оказываются обрезанными, может оказаться полезным перед применением преобразования добавить в диаграмму врезки. Врезки уменьшают размерности на определенный процент до того, как к системе координат будут применены какие-либо преобразования. Вы можете управлять наименьшим *x*, наибольшим *x*, наименьшим *y* и наибольшим *y* измерениями (именно в таком порядке).

Врезка % после преобразования . Если нужно изменить пропорции диаграммы, после применения преобразования можно добавить врезки в диаграмму. Врезки уменьшают измерения на определенный процент после того, как к системе координат применяются какие-либо преобразования. Эти врезки можно применить даже в том случае, если никакие преобразования к диаграмме не применяются. Вы можете управлять наименьшим *x*, наибольшим *x*, наименьшим *y* и наибольшим *y* измерениями (именно в таком порядке).

Изменение статистик и графических элементов

Графический элемент можно преобразовать к другому типу, изменить статистику, используемую для изображения графического элемента, или задать модификатор коллизий, определяющий, что происходит при перекрытии графических элементов.

Как преобразовать графический элемент

1. Выделите графический элемент, который хотите преобразовать.
2. Перейдите на вкладку **Элемент** в палитре свойств.
3. Выберите новый тип графического элемента в списке Тип.

Таблица 36. Тип графических элементов

Тип графических элементов	Описание
Точки	Маркер, идентифицирующий конкретную точку данных. Элемент точка используется в диаграммах рассеяния и других аналогичных визуализациях.
Интервал	Прямоугольная фигура, которая прорисовывается для конкретного значения данных и заполняет пространство между началом координат и другим значением данных. Элемент столбец используется в столбчатых диаграммах и в гистограммах.
Линейчатую	Линия, которая соединяет значения данных.
Путь	Линия, которая соединяет значения данных в порядке их следования в наборе данных.
С областями	Пространство между линией, соединяющая элементы данных, и началом координат.
Многоугольник	Многогранная фигура, в которой заключена область данных. Элемент многоугольник можно использовать в диаграмме рассеяния с группировкой или в карте.
Схема	Элемент, состоящий из ящика с усами и маркеров, которые обозначают выбросами. Элемент схема используется для ящичных диаграмм с усами.

Как изменить статистику

1. Выберите графический элемент, статистику которого хотите изменить.
2. Перейдите на вкладку **Элемент** в палитре свойств.

Как задать модификатор коллизий

Модификатор коллизий определяет, что происходит при наложении графических элементов.

1. Выделите графический элемент, для которого вы хотите задать модификатор коллизий.
2. Перейдите на вкладку **Элемент** в палитре свойств.
3. В раскрывающемся списке Модификатор выберите модификатор коллизий. **-авто-** предоставляет приложению определить, какой модификатор коллизий будет уместен для типа графического элемента и статистики.

   Наложение . Графические элементы прорисовываются друг над другом, если они имеют одинаковое значение.

   Стыкование. Стыкует графические элементы, которые накладывались бы друг на друга, при наличии одинаковых значений данных.

   Отклонение. Графические элементы размещаются рядом с другими графическими элементами, показанными на том же значении, а не накладываются друг на друга. Графические элементы располагаются симметрично. Это значит, что графические элементы перемещаются на противоположные стороны от центрального положения. Отклонение очень похоже на кластеризацию.

   Стопка . Графические элементы размещаются рядом с другими графическими элементами, показанными на том же значении, а не накладываются друг на друга. Графические элементы располагаются асимметрично. Это значит, что графические элементы размещаются один над другим; при этом нижний графический элемент находится на определенном значении шкалы.

   Дрожание (нормальное) . Графические элементы размещаются на одном значении данных случайным образом с помощью нормального распределения.

   Дрожание (равномерное) . Графические элементы размещаются на одном значении данных случайным образом с помощью равномерного распределения.

Изменение положения легенды

Если диаграмма содержит легенду, она обычно показана справа от диаграммы. При необходимости можно изменить ее положение.

Как изменить положение легенды

1. Выделите легенду.
2. Перейдите на вкладку **Легенда** на палитре свойств.
3. Выберите положение.

Копирование визуализации и данных визуализации

Палитра Общие содержит кнопки для копирования визуализаций и их данных.

Рисунок 62. Кнопка копирования визуализации

Копирование визуализации. Это действие копирует визуализацию в буфер обмена как изображение. Доступно несколько форматов изображений. При вставке изображения в другое приложение, можно выбрать вариант Специальная вставка, а затем выбрать для вставки один из доступных форматов изображений.

Рисунок 63. Кнопка копирования данных визуализации

Копирование данных визуализации. Это действие копирует исходные данные, используемые для создания визуализации. Данные копируются в буфер обмена как обычный текст или текст в формате HTML. При вставке данных в другое приложение можно выбрать вариант Специальная вставка, а затем выбрать для вставки один из этих форматов.

Сочетания клавиш в редакторе диаграмм

Таблица 37. Сочетания клавиш

Клавиши	Функция
Ctrl+Пробел	Переключиться между режимами изучения и редактирования
Удалить	Удалить элемент визуализации
Ctrl+Z	Отменить
Ctrl+Y	Повторить
F2	Показать структуру для выбора элементов диаграммы

Добавление заголовков и сносок

Для всех типов диаграмм можно добавлять уникальные заголовки, сноски и метки осей, что помогает идентифицировать содержимое диаграмм.

Добавление заголовков диаграмм

1. Выберите в меню пункт **Изменить > Добавить заголовок диаграммы**. Над диаграммой появится текстовое окно, содержащее поле **<ЗАГОЛОВОК>**.
2. Проверьте, включен ли режим Изменение. Выберите пункт меню **Просмотр > Режим изменения**.
3. Щелкните два раза по тексту **<TITLE>**.
4. Введите требуемый заголовок и нажмите клавишу Enter.

Добавление сносок к диаграммам

1. Выберите в меню пункт **Изменить > Добавить сноску к диаграмме**. Над диаграммой появится текстовое окно, содержащее поле **<СНОСКА>**.
2. Проверьте, включен ли режим Изменение. Выберите пункт меню **Просмотр > Режим изменения**.
3. Щелкните два раза по тексту **<FOOTNOTE>**.
4. Введите требуемый заголовок и нажмите клавишу Enter.

Использование таблиц стилей диаграмм

Основной информацией о выводе диаграммы, такой как цвета, шрифты, обозначения и толщина линий, управляет таблица стилей. С IBM SPSS Modeler поставляется таблица стилей по умолчанию; однако вы можете изменить ее по необходимости. Например, может существовать корпоративная цветовая схема для презентаций, которую вы захотите использовать использовать в диаграммах. Дополнительную информацию смотрите в разделе “Редактирование визуализаций” на стр. 270.

В узлах диаграммы можно использовать режим редактирования, чтобы произвести стилевые изменения во внешнем виде диаграмм. Затем можно использовать меню **Изменить > Стили** и сохранить изменения как

таблицу стилей для применения ко всем диаграммам, которые будут последовательно генерироваться на текущем узле диаграмм, или как новую таблицу стилей по умолчанию для всех создаваемых в IBM SPSS Modeler диаграмм.

Есть пять опций таблиц стилей, доступных в подпункте **Стили** меню Изменить:

- **Сменить таблицу стилей.** Здесь выводится список различных хранимых таблиц стилей, которые можно выбрать для изменения вида ваших диаграмм. Дополнительную информацию смотрите в разделе “Применение таблиц стилей”.
- **Сохранить стили в узле.** Здесь хранятся изменения к выбранным стилям диаграмм, чтобы использовать их для любой из следующих диаграмм, создаваемых в одном узле диаграмм текущего потока.
- **Сохранить как стили по умолчанию.** Здесь хранятся изменения к выбранным стилям диаграмм, чтобы использовать их для любой из следующих диаграмм, создаваемых в любом узле диаграмм любого потока. После выбора этой опции можно использовать возможность **Применить стили по умолчанию**, чтобы изменить все другие существующие диаграммы и использовать одинаковые стили.
- **Применить стили по умолчанию.** Здесь выбранные стили диаграмм изменяются на те стили, которые в данный момент сохранены как стили по умолчанию.
- **Применить исходные стили.** Здесь стили диаграмм изменяются обратно к первоначальным, поставляемым как исходные стили по умолчанию.

Применение таблиц стилей

Можно применить таблицу стилей визуализации, в которой описаны стилистические свойства визуализации. Например, в таблице стилей могут быть описаны шрифты, типы заливки, цвета и другие свойства. Таблицы стилей обеспечивают ускоренный доступ к правкам, которые иначе выполнялись бы вручную. Однако имейте в виду, что таблица стилей позволяет изменять только *стили* . Другие изменения, такие как положение легенды или диапазон шкалы не сохраняются в таблицах стилей.

Применение таблицы стилей

1. Выберите в меню:

   Правка > Стили > Переключить таблицу стилей
2. В диалоговом окне Смена таблицы стилей выберите таблицу стилей.
3. Нажмите **Применить** , чтобы применить таблицу стилей к визуализации, не закрывая диалоговое окно. Нажмите **OK** , чтобы применить таблицу стилей и закрыть диалоговое окно.

Диалоговое окно Сменить/выбрать таблицу стилей.

В таблице в верхней части диалогового окна перечислены все доступные в данный момент таблицы стилей визуализаций. Некоторые таблицы стилей установлены предварительно, а другие, возможно, были созданы в IBM SPSS Visualization Designer (отдельная программа) .

Внизу диалогового окна показываются примеры визуализаций с гипотетическими данными. Выберите одну из таблиц стилей, чтобы применить ее стили к примерам визуализаций. С помощью этих примеров можно понять, как таблица стилей повлияет на визуализацию.

В диалоговом окне также доступны следующие параметры.

Существующие стили . По умолчанию таблица стилей может перезаписывать все стили в визуализации. Это можно изменить.

- **Перезаписать все стили** . При применении таблицы стилей происходит перезапись всех стилей в визуализации, включая стили, измененные в визуализации в течение текущего сеанса редактирования.
- **Сохранить измененные стили** . При применении таблицы стилей происходит перезапись только тех стилей, которые *не* были изменены в визуализации в течение текущего сеанса редактирования. Стили, которые были изменены в течение текущего сеанса редактирования, сохраняются.

Управление . Управление шаблонами визуализаций, таблицами стилей и картами на компьютере. На вашем локальном компьютере можно импортировать, экспортировать, переименовывать и удалять шаблоны визуализации , таблицы стилей и отображения. Дополнительную информацию смотрите в разделе “Управление шаблонами, таблицами стилей и файлами карт” на стр. 209.

Местоположение. Изменение папки, в которой хранятся шаблоны визуализаций, таблицы стилей и карты Текущее местоположение показывается справа от кнопки. Дополнительную информацию смотрите в разделе “Указание местоположения для хранения шаблонов, таблиц стилей и карт” на стр. 208.

Печать, сохранение, копирование и экспорт диаграмм

У каждой диаграммы есть несколько опций, позволяющих сохранить или напечатать диаграмму, а также экспортировать ее в другой формат. Большинство этих опций доступны в меню Файл. Кроме этого, в меню Правка можно выбрать копирование диаграммы или ее данных для использования в другой прикладной программе.

Печать

Для печати диаграммы используйте пункт меню **Печать** или соответствующую кнопку. Перед печатью можно использовать пункты меню **Конфигурирование страницы** и **Предварительный просмотр** для задания опций печати и предварительного просмотра вывода.

Сохранение диаграмм

Чтобы сохранить диаграмму в файле вывода IBM SPSS Modeler (*.*cou*), выберите пункт меню **Файл > Сохранить** или **Файл > Сохранить как**.

или

Чтобы сохранить диаграмму в репозитории, выберите пункт меню **Файл > Сохранить данные вывода**.

Копирование диаграмм

Чтобы скопировать диаграмму для использования в другой прикладной программе, например, в MS Word или MS PowerPoint, выберите пункт меню **Изменить > Копировать диаграмму**.

Копирование данных

Чтобы скопировать данные для использования в другой прикладной программе, например, в MS Excel или MS Word, выберите пункт меню **Изменить > Копировать данные**. По умолчанию данные будут сформатированы как HTML. В другой прикладной программе используйте возможность **Специальная вставка**, чтобы просмотреть другие опции форматирования при вставке.

Экспорт диаграмм

Опция **Экспорт диаграмм** позволяет экспортировать диаграмму в одном из следующих форматов: Bitmap (*.bmp*), JPEG (*.jpg*), PNG (*.png*), HTML (*.html*) или документ ViZml (*.xml*) для использования в других прикладных программах IBM SPSS Statistics.

Чтобы экспортировать диаграммы, выберите пункт меню **Файл > Экспортировать диаграмму**, а затем выберите формат.

Экспорт таблиц

Опция **Экспортировать таблицу** позволяет экспортировать таблицу в одном из следующих форматов: с разделением символами табуляции (*.tab*), с разделением запятыми (*.csv*) или HTML (*.html*).

Чтобы экспортировать таблицы, выберите пункт меню **Файл > Экспортировать таблицу**, а затем выберите формат.

Глава 6. Узлы вывода

Обзор узлов вывода

Узлы вывода предоставляют средние значения для получения информации об используемых данных и моделях. Эти узлы предоставляют также механизм для экспорта данных в различных форматах для взаимодействия с другими используемыми программными инструментами.

Доступны следующие узлы вывода:

Узел Таблица выводит данные в табличном формате, которые можно также записать в файл. Это полезно всякий раз, когда вам нужно проверить значения своих данных или экспортировать их в просто читаемую форму.

Узел Матрица создает таблицу, показывающую взаимосвязи между полями. Чаще всего он используется для показа взаимосвязи между двумя символическими полями, но он же может показывать взаимосвязи между флаговыми или числовыми полями.

Узел Анализ оценивает способность прогнозирующих моделей генерировать точные предсказания. Узлы анализа выполняют различные сравнения между предсказанными и фактическими значениями для одного или нескольких слепков моделей. Они могут сравнивать также прогнозирующие модели друг с другом.

Узел Аудит данных предоставляет всесторонний первый взгляд на данные, в том числе сводную статистику, гистограммы и распределение для каждого поля, а также информацию о выбросах, значениях отсутствия и экстремумах. Результаты выводятся в виде простой для чтения матрицы, которую можно отсортировать и использовать для генерирования узлов полноразмерных графиков и подготовки данных.

Узел Преобразование позволяет выбрать и предварительно просмотреть результаты преобразований, прежде чем применить их к выбранным полям.

Узел Статистика предоставляет базовую сводную информацию о числовых полях. Здесь вычисляется сводная статистика для индивидуальных полей и корреляции между полями.

Узел средних значений сравнивает независимые группы или пары связанных полей для проверки, существует ли между ними существенное различие. Например, можно сравнить средние прибыли до и после рекламной кампании, или сравнить прибыли от клиентов, не получавших рекламы, и клиентов, участвовавших в программе продвижения товара.

Узел отчетов создает форматированные отчеты, содержащие фиксированный текст, а также данные и другие выражения, полученные из данных. Вы задаете формат отчета, используя текстовые шаблоны, чтобы определить конструкции фиксированного текста и вывода данных. Вы можете предоставить пользовательское форматирование текста с помощью тегов HTML в шаблоне и задав опции на вкладке Вывод. Значения данных и другой условный вывод можно включить в отчет с использованием выражений CLEM в шаблоне.

Узел Задать глобальные значения просматривает данные и вычисляет сводные значения, которые можно использовать в выражениях CLEM. Например, можно использовать этот узел для вычисления статистических показателей для поля с именем *age*, а затем использовать общее среднее *age* в выражениях CLEM, вставив функцию `@GLOBAL_MEAN(age)`.

Узел подгонки имитации исследует статистическое распределение данных в каждом поле и генерирует (или обновляет) узел генерирования имитации, используя для каждого поля оптимально подогнанное распределение. Затем узел генерирования имитации можно использовать для генерирования данных имитации.

Узел оценки имитации оценивает заданное предсказанное поле назначения и представляет информацию о распределении и корреляции этого поля назначения.

Управление выводом

Менеджер вывода показывает диаграммы, графики и таблицы, сгенерированные за сеанс IBM SPSS Modeler. Вывод всегда можно открыть повторно, щелкнув по нему дважды кнопкой мыши в менеджере; перезапускать соответствующий поток или узел не требуется.

Чтобы просмотреть Менеджер вывода

Откройте меню Вид и выберите **Менеджеры**. Щелкните по вкладке **Поля вывода**.

В менеджере вывода можно:

- Вывести существующие объекты вывода, такие как гистограммы, диаграммы оценки и таблицы.
- Переименовать объекты вывода.
- Сохранить объекты вывода на диске или в IBM SPSS Collaboration and Deployment Services Repository (если он доступен).
- Добавить файлы вывода в существующий проект.
- Удалить несохранные объекты вывода из текущего сеанса.
- Открыть сохраненные объекты вывода или получить их из IBM SPSS Collaboration and Deployment Services Repository (если он доступен).

Для обращения к этим опциям щелкните правой кнопкой мыши в любом месте вкладки Поля вывода.

Просмотр вывода

На экране вывод отображается в окне браузера вывода. Окно браузера вывода содержит свой собственный набор меню, позволяющих вывести вывод на печать, сохранять его или экспортировать его в другой формат. Имейте в виду, что конкретный состав опций определяется типом вывода.

Печать, сохранение и экспорт вывода. Доступна следующая дополнительная информация:

- Для печати вывода опцию или кнопку меню **Печать**. Перед печатью можно использовать пункты меню **Конфигурирование страницы** и **Предварительный просмотр** для задания опций печати и предварительного просмотра вывода.
- Чтобы сохранить вывод в файле вывода IBM SPSS Modeler (*.cou*), в меню файл выберите **Сохранить** или **Сохранить как**.
- Чтобы сохранить вывод в другом формате, например, в тестовом или HTML, в меню Файл выберите **Экспорт**. Дополнительную информацию смотрите в разделе “Экспорт вывода” на стр. 288.

  Обратите внимание на то, что эти форматы можно выбрать только при обеспечении осмысленного экспорта данных, содержащихся в выводе. Например, содержимое дерева решений можно экспортировать как текст, но содержимое модели K-средних в виде текста будет бессмысленным.
- Для сохранения вывода в совместно используемом репозитории, чтобы другие пользователи могли его просмотреть при помощи IBM SPSS Collaboration and Deployment Services Deployment Portal, выберите в меню Файл опцию **Опубликовать в Web**. Имейте в виду, что, что для этой опции требуется отдельная лицензия к IBM SPSS Collaboration and Deployment Services.

Выбор ячеек и столбцов. Меню Правка содержит различные опции для выбора, отмены выбора и копирования ячеек и столбцов образом, соответствующим текущему типу вывода. Дополнительную информацию смотрите в разделе “Выбор ячеек и столбцов” на стр. 289.

Генерирование новых узлов. Меню Создать позволяет сгенерировать новые узлы на основе содержимого браузера вывода. Состав этих опций непостоянен в зависимости от типа вывода и элементов, выбранных в выводе в текущий момент. Подробности об опциях генерирования узлов для конкретного типа вывода смотрите в документации к этому выводу.

Опубликовать в Web

Возможность Опубликовать в Web позволяет публиковать вывод потока определенных типов в центральном совместно используемом IBM SPSS Collaboration and Deployment Services Repository, образующем основу IBM SPSS Collaboration and Deployment Services. Если вы примените эту опцию, другие пользователи, которым нужно просмотреть данный вывод, смогут это сделать, используя доступ в Интернет и учетную запись IBM SPSS Collaboration and Deployment Services; установка IBM SPSS Modeler им не потребуется.

В приведенной ниже таблице перечислены узлы IBM SPSS Modeler, поддерживающие возможность Опубликовать в Web. Вывод с этих узлов хранится в IBM SPSS Collaboration and Deployment Services Repository в формате объектов вывода (*.cou*), и его можно просмотреть непосредственно в IBM SPSS Collaboration and Deployment Services Deployment Portal.

Другие типы вывода можно просмотреть, только если на компьютере будет установлена нужная прикладная программа (например IBM SPSS Modeler для объектов потока).

Таблица 38. Узлы, поддерживающие публикацию в Web.

Тип узла	**Узел**
Диаграммы	все
Вывод	Таблица
	Матрица
	Аудит данных
	Преобразование
	Средние
	Анализ
	Статистика
	Отчет (HTML)
IBM SPSS Statistics	Вывод статистики

Публикация вывода в Web

Чтобы опубликовать вывод в Web:

1. В потоке IBM SPSS Modeler вызовите один из узлов, перечисленных в таблице. При этом в будет создан объект вывода (например, объект таблицы, матрицы или отчета) в новом окне.
2. В этом окне объекта вывода выберите:

   Файл > Опубликовать в Web

   Примечание: Если вы хотите всего лишь экспортировать простые файлы HTML для использования в стандартном Web-браузере, в меню Файл выберите **Экспорт** и выберите опцию **HTML**.
3. Соединитесь с IBM SPSS Collaboration and Deployment Services Repository.

   После успешного соединения откроется диалоговое окно Репозиторий: сохранить, где предлагается ряд опций хранения.
4. Выбрав нужные вам опции хранения, нажмите кнопку **Сохранить**.

Просмотр опубликованного вывода в Web

Для использования этой возможности у вас должна быть сконфигурирована учетная запись IBM SPSS Collaboration and Deployment Services. Если у вас установлена нужная прикладная программа для типа объектов, который вы хотите просмотреть, (например, IBM SPSS Modeler или IBM SPSS Statistics), вывод появится не в браузере, а в самой этой программе.

Чтобы просмотреть опубликованный вывод в Web:

1. Укажите в своем браузере *http://<хост_репозитория>:<порт_репозитория>/peb*

   где *хост_репозитория* и *порт_репозитория* - это имя хоста и номер порта для хоста IBM SPSS Collaboration and Deployment Services.
2. Введите сведения для входа в систему IBM SPSS Collaboration and Deployment Services.
3. Нажмите кнопку **Репозиторий содержимого**.
4. Перейдите к объекту, который вы хотите просмотреть, или найдите его.
5. Щелкните по имени объекта. Для объектов некоторых типов, таких как диаграммы, возможна задержка при обработке вывода объекта в браузере.

Просмотр вывода в браузере HTML

С вкладки Дополнительно слепков линейных моделей, логистических и моделей типа PCA/фактор выводимую информацию можно просмотреть в отдельном браузере, таком как Internet Explorer. Информация выводится в виде HTML, что позволяет сохранить ее в другом месте, например, во внутрикорпоративной сети или на сайте в Интернете.

Чтобы вывести информацию в браузере, нажмите кнопку запуска, расположенную под значком модели в верхней левой части вкладки Дополнительно слепка модели.

Экспорт вывода

В окне браузера вывода можно выбрать экспорт вывода в другой формат, например, текстовый или HTML. Состав форматов вывода зависит от типа вывода, но в основном подобен опциям типов файлов, доступным при выборе опции **Сохранить в файле** на узле, используемом для генерирования вывода.

Примечание: Эти форматы можно выбрать только при обеспечении осмысленного экспорта данных, содержащихся в выводе. Например, содержимое дерева решений можно экспортировать как текст, но содержимое модели K-средних в виде текста будет бессмысленным.

Как экспортировать вывод

1. В браузере вывода откройте меню Файл и выберите **Экспорт**. Затем выберите тип файла, который вы хотите создать:

- **Разделитель - табуляция (*.tab).** Эта опция генерирует файл с форматированным текстом, содержащий значения данных. Этот стиль часто бывает полезен для генерирования простого текстового представления информации, которую можно импортировать в другие прикладные программы. Эта опция доступна для узлов таблицы, матрицы и средних.
- **Разделитель - запятая (*.dat).** Эта опция генерирует текстовый файл с разделителями-запятыми, содержащий значения данных. Этот стиль часто бывает полезен как быстрый способ сгенерировать файл данных, который можно импортировать в электронные таблицы или или другие прикладные программы анализа данных. Эта опция доступна для узлов таблицы, матрицы и средних.
- **Транспонированный с разделителем - табуляцией (*.tab).** Эта опция идентична опции Разделитель - табуляция, но данные транспонируются таким образом, что строки представляют поля, а столбцы представляют записи.
- **Транспонированный с разделителем - запятой (*.dat).** Эта опция идентична опции Разделитель - запятая, но данные транспонируются таким образом, что строки представляют поля, а столбцы представляют записи.
- **HTML (*.html).** Эта опция записывает вывод в формате HTML в один или несколько файлов.

Выбор ячеек и столбцов

Ряд узлов, включая узел таблицы, узел матрицы и узел средних, генерируют табличный вывод. Эти выходные таблицы можно просмотреть и управлять ими похожими способами, выполняя такие операции, как выбор ячеек, копирование в буфер обмена всей или части таблицы, генерирование новых узлов на основе выбранных в текущим момент вариантов, а также сохранение таблицы и вывод ее на печать.

Выбор ячеек. Чтобы выбрать ячейку, щелкните по ней. Чтобы выбрать диапазон ячеек, щелкните в одном углу нужного диапазона и перетащите указатель мыши в другой угол диапазона, после чего отпустите кнопку мыши. Чтобы выбрать весь столбец, щелкните по его заголовку. Чтобы выбрать несколько столбцов, щелкните по их заголовкам при нажатой клавише Shift или Ctrl.

При новом выборе старый выбор стирается. Если при выборе удерживать нажатой клавишу Ctrl, новый выбор будет добавлен к уже существующему выбору, который при этом не будет стерт. Этим способом можно выбрать несколько несмежных областей таблицы. Мен Правка содержит также опции **Выбрать все** и **Очистить выбор**.

Переупорядочение столбцов. Браузеры вывода узла таблицы и узла средних позволяют перемещать столбцы в таблице, щелкнув по заголовку столбца и перетащив его в нужное положение. Одновременно можно переместить только один столбец.

Узел таблицы

Узел таблицы создает таблицу, возвращающую список значений исследуемых данных. В нее включаются все поля и все значения в потоке, предоставляя удобный способ проверки значений данных или их экспорта в удобочитаемой форме. Дополнительно можно выделить записи, соответствующие определенному условию.

Примечание: Если вы работаете не с маленькими наборами данных, рекомендуется выбрать подмножество данных для передачи на узел таблицы. Узел таблицы не может выполнять правильный вывод, когда число записей превосходит размер, допустимый для содержания в структуре вывода (например, 100 миллионов строк).

Вкладка Параметры узла таблицы

Выделять записи, где. Записи в таблице можно выделить, введя выражение CLEM, оцениваемое как true для выделяемых записей. Эта опция доступна, только если выбрана опция **Вывод на экран**.

Вкладка Формат узла таблицы

Вкладка Формат содержит опции. используемые для задания форматирования на уровне отдельных полей. Эта вкладка совместно используется с узлом типа. Дополнительную информацию смотрите в разделе “Вкладка Параметры форматов полей” на стр. 140.

Вкладка Вывод узлов вывода

Для узлов, генерирующих вывод в табличном стиле, вкладка Вывод позволяет задать формат и положение этих результатов.

Имя вывода. Задает имя вывода, генерируемого при вызове узла. Опция **Авто** выбирает имя на основе узла, генерирующего вывод. Дополнительно можно выбрать **Пользовательское**, чтобы задать другое имя.

Вывод на экран (по умолчанию)**.** Создает объект вывода для просмотра в оперативном режиме. Этот объект вывода появляется на вкладке Выходные поля окна менеджера при вызове узла вывода.

Вывод в файл. Сохраняет вывод в файле при вызове узла. При выборе этой опции введите имя файла (или перейдите в каталог и задайте имя файла при помощи кнопки средства выбора файлов) и выберите тип файла. Имейте в виду, что некоторые типы файлов могут быть недоступны для определенных типов вывода.

Данные выводятся в формате кодирования системы по умолчанию, который задается на панели управления Windows или (при работе в распределенном режиме) на компьютере сервера.

- **Данные (с разделителем-табулятором) (*.tab).** Эта опция генерирует файл с форматированным текстом, содержащий значения данных. Указанный стиль часто бывает полезен для генерирования простого текстового представления информации, которую можно импортировать в другие прикладные программы. Эта опция доступна для узлов таблицы, матрицы и средних.
- **Данные (с разделителем-запятой) (*.dat).** Эта опция генерирует текстовый файл с разделителями-запятыми, содержащий значения данных. Указанный стиль часто бывает полезен как быстрый способ генерирования файла данных, который можно импортировать в электронные таблицы или или другие прикладные программы анализа данных. Эта опция доступна для узлов таблицы, матрицы и средних.
- **HTML (*.html).** Эта опция записывает вывод в формате HTML в один или несколько файлов. Для табличного вывода (с узлов таблицы, матрицы или средних) набор файлов HTML содержит панель содержания со писком имен полей и данные в таблице HTML. Эта таблица может быть разбита на несколько файлов HTML, если число строк в ней превышает спецификацию **Строк на страницу**. В этом случае панель содержимого содержит ссылки на страницы и предоставляет средства навигации по таблице. Для нетабличного вывода создается один файл HTML, содержащий результаты узла.

  Примечание: Если вывод HTML содержит форматирование только для первой страницы, выберите опцию **Разбить вывод на страницы** и скорректируйте спецификацию **Строк на страницу** так, чтобы включить весь вывод в одну страницу. Либо, если шаблон вывода для узлов, таких как узел отчета, содержит пользовательскую страницу HTML, убедитесь, задан **Пользовательский** тип формата.
- **Текстовый файл (*.txt).** Эта опция генерирует текстовый файл, содержащий значения данных. Указанный стиль часто бывает полезен для генерирования вывода, который можно импортировать в другие прикладные программы, такие как текстовые редакторы или программы презентаций. Для некоторых узлов эта опция недоступна.
- **Объект вывода (*.cou).** Объекты вывода, сохраненные в этом формате, можно открыть и просмотреть в IBM SPSS Modeler, добавить в проекты и опубликовать, а также отслеживать при помощи IBM SPSS Collaboration and Deployment Services Repository.

Представление вывода. Для узла средних можно указать, отображать ли по умолчанию на экране простой или расширенный вывод. Имейте в виду, что между этими двумя представлениями можно также переключаться при просмотре сгенерированного вывода. Дополнительную информацию смотрите в разделе “Браузер вывода узла средних” на стр. 311.

Формат. Для узла отчета можно выбрать, форматировать ли вывод автоматически или при помощи HTML, включенного в шаблон. Выберите **Пользовательский**, чтобы разрешить форматирование HTML в шаблоне.

Заголовок. Для узла отчета можно задать необязательный текст заголовка, который появится в верхней части вывода отчета.

Выделите вставленный текст. Для узла отчета выберите эту опцию для выделения текста, генерируемого выражениями CLEM в шаблоне отчета. Дополнительную информацию смотрите в разделе “Вкладка Шаблон узла отчета” на стр. 313. Эта опция не рекомендуется, если используется **Пользовательский** формат.

Строк на страницу. Для узла отчета задайте число строк для включения в каждую страницу в режиме форматирования **Авто**.

Транспонировать данные. Эта опция транспонирует данные перед экспортом таким образом, чтобы строки представляли поля, а столбцы - записи.

Примечание: Для больших опции выше могут оказаться недостаточно эффективными, особенно при работе с удаленным сервером. В таких случаях использование узла вывода файлов обеспечит более высокую производительность. Дополнительную информацию смотрите в разделе “Узел экспорта плоских файлов” на стр. 342.

Браузер таблиц

Браузер таблиц выводит табличные данные и позволяет выполнять стандартные операции, в том числе выбор и копирование ячеек, переупорядочение столбцов, сохранение и распечатку таблицы. Дополнительную информацию смотрите в разделе “Выбор ячеек и столбцов” на стр. 289. Это те же операции, которые можно выполнить при предварительном просмотре данных в узле.

Экспорт табличных данных. Экспортировать данные из браузера таблиц можно при следующем выборе:

Файл > Экспорт

Дополнительную информацию смотрите в разделе “Экспорт вывода” на стр. 288.

Данные экспортируются в системном формате кодировки по умолчанию, который задан на панели управления Windows, или, при запуске в распределенном режиме, на компьютере сервера.

Поиск в таблице. Кнопка поиска на главной панели инструмента (со значком бинокля) активирует панель инструментов поиска, позволяя проводить поиск конкретных значений в таблице. В таблице поиск можно проводить в прямом или обратном направлении, задавать учет или неучет регистра (кнопка **Aa**), а также прерывать поиск при иго выполнении, нажав кнопку прерывания поиска.

Генерирование новых узлов. Меню Создать содержит операции генерирования узлов.

- **Узел выбора ("Записи").** Генерирует узел Выбор, где выбираются записи, для которых отбираются ячейки в таблице.
- **Выбрать ("And").** Генерирует узел Выбор, где выбираются записи, содержащие *все* значения, отобранные в таблице.
- **Выбрать ("Or").** Генерирует узел Выбор, где выбираются записи, содержащие *некоторые* значения, отобранные в таблице.
- **Извлечь ("Записи").** Генерирует узел извлечения для создания нового флагового поля. Флаговое поле содержит *T* для записей, соответствующих любой выбранной ячейке в матрице, и *F* для остальных записей.
- **Извлечь ("And").** Генерирует узел извлечения для создания нового флагового поля. Флаговое поле содержит *T* для записей, содержащих *все* значения, выбранные в таблице, и *F* для остальных записей.

- **Извлечь ("Or").** Генерирует узел извлечения для создания нового флагового поля. Флаговое поле содержит *T* для записей, содержащих *некоторые* значения, выбранные в таблице, и *F* для остальных записей.

Узел матрицы

Узел матрицы позволяет создать таблицу, показывающую взаимосвязи между полями. Чаще всего она используется для показа взаимосвязи между двумя категориальными полями (флаговыми, номинальными или порядковыми), но ее можно также использовать для показа взаимосвязей между непрерывными полями (числового диапазона).

Вкладка Параметры узла матрицы

На вкладке Параметры можно задать опции для структуры матрицы.

Поля. Выберите тип выбора полей при помощи следующих опций:

- **Выбранные.** Эта опция позволяет выбрать категориальное поле для строк и поле для столбцов в матрице. Строки и столбцы матрицы определяются списком значений для выбранного категориального поля. Ячейки этой матрицы содержат итожащие статистики, выбранные ниже.
- **Все флаги (значения true).** Эта опция запрашивает в требовании матрицу с одной строкой и одним столбцом для каждого флагового поля в данных. Ячейки этой матрицы содержат количества, выраженные положительными числами двойной точности, для каждого сочетания флагов. Другими словами, для каждой строки, соответствующей *купленному хлебу*, и столбцу, соответствующему *купленному сыру*, ячейка в пересечении этой строки и столбца будет содержать число записей, для которых и *купленный хлеб*, и *купленный сыр* являются значениями true.
- **Все числовые** Эта опция запрашивает в требовании матрицу с одной строкой и одним столбцом для каждого числового поля. Ячейки этой матрицы представляют сумму перекрестных произведений для соответствующей пары полей. Другими словами, для каждой ячейки в матрице значения для поля строки и поля столбца будут перемножаться для каждой записи, а затем суммироваться по записям.

Включить пропущенные значения. Включает в вывод строк и столбцов пользовательские пропущенные значения (пробельные символы) и системные пропущенные значения ($null$). Например, если значение *Нет* было определено для выбранного столбца как пользовательское пропущенное значение, в таблицу будет включен отдельный столбец, помеченный меткой *Нет*, точно также, как и любая другая категория (при допущении, что это значение фактически присутствует в данных). Если эта опция не включена, столбец *Нет* исключается, независимо от того, как часто он встречается.

Примечание: Опция для включения пропущенных значений применяется, только если выбранные поля являются полями комбинационной таблицы. Пробельные значения отображаются на пустые ($null$) и исключаются из агрегации для поля функции, если используется режим **Выбранные** и содержимое задано как **Функция**; и для всех числовых полей это выполняется, если используется режим **Все числовые**.

Содержимое ячеек. Если вы выбрали поля **Выбранные** выше, можно задать статистику для использования в ячейках матрицы. Выберите статистику, основанную на количествах, или выберите поле наложения для суммирования значений числового поля на основе значений полей строки и столбца.

- **Комбинационная таблица.** Значения ячеек представляют собой количества и/или процентные доли записей, у которых есть соответствующее сочетание значений. Нужные вам сводки комбинационной таблицы можно задать при помощи опций на вкладке Внешний вид. Выводится также глобальное значение хи-квадрат наряду с со значимостью. Дополнительную информацию смотрите в разделе “Браузер вывода узла матрицы” на стр. 293.
- **Функция.** Если выбрать функцию сводки, значения ячеек будут функцией от выбранных значений поля наложения для наблюдений, обладающих подходящими значениями строки и столбца. Например, в случае поля строки *Регион*, поля столбца *Продукт* и поля наложения *Доход* ячейка в строке *Северо-восток* и столбце *Виджет* будет содержать суммарный (средней, минимальный или максимальный) доход для виджетов в северо-восточном регионе. Функция сводки по умолчанию - **Среднее**. Для суммирования в

поле функции можно выбрать другую функцию. Доступны варианты: **Среднее**, **Сумма**, **SDev** (среднеквадратичное отклонение), **Макс** (максимум) и **Мин** (минимум).

Вкладка Внешний вид узла матрицы

В кладка Внешний вид позволяет управлять опциями сортировки и выделения для матриц комбинационных таблиц.

Строки и столбцы. Управляет сортировкой заголовков строк и столбцов в матрице. Значение по умолчанию - **Без сортировки**. Выберите **По возрастанию** или **По убыванию** для сортировки заголовков строк и столбцов в указанном направлении.

Наложение. Позволяет выделить в матрице значения экстремумов. Значения выделяются основе количеств в ячейках (для матриц комбинационных таблиц) или вычисленных значений (для функциональных матриц).

- **Выделить верхние.** Можно затребовать выделение в матрице самых высоких значений (красным). Задайте число выделяемых значений.
- **Выделить нижние.** Можно также затребовать выделение в матрице самых низких значений (зеленым). Задайте число выделяемых значений.

Примечание: Для этих двух опций выделения совпадающие значения могут привести к выделению большего числа значений, чем затребовано. Например, при наличии матрицы с шестью нулями в ячейках, если вы затребуете **Выделить 5 нижних**, будут выделены все шесть нулей.

Содержимое ячеек комбинационной таблицы. Для комбинационных таблиц можно задать итожащие статистики, содержащиеся в матрице, для матриц комбинационных таблиц. Если вкладке Параметры выбрана опция **Все числовые** или **Функция**, эти опции недоступны.

- **Количества.** В ячейки включается число записей со значением строки, у которых есть соответствующее значение столбца. Это - единственное содержимое ячейки по умолчанию.
- **Ожидаемые значения.** Ожидаемое значение для числа записей в ячейке при допущении отсутствия взаимосвязи между строками и столбцами. Ожидаемые значения основаны на следующей формуле:

  p(*значение строки*) * p(*значение столбца*) * *общее число записей*
- **Остатки.** Разность между наблюденными и ожидаемыми значениями.
- **Процент строк.** Процент всех записей со значением строки, у которых есть соответствующее значение столбца. Проценты суммируются до 100 в строках.
- **Процент столбцов.** Процент всех записей со значением столбца, у которых есть соответствующее значение строки. Проценты суммируются до 100 в столбцах.
- **Общий процент.** Процент всех записей, у которых есть соответствующее сочетание значения столбца и значения строки. Проценты суммируются до 100 по всей матрице.
- **Включить итоги строк и столбцов.** Добавляет в матрицу строку и столбец для итогов по столбцам и строкам).
- **Применить параметры.** (Только для браузера вывода) Позволяет внести изменения во внешний вид вывода узла матрицы без закрытия и повторного открытия браузера вывода. Внесите на этой вкладке браузера вывода изменения, нажмите эту кнопку, а затем выберите вкладку Матрица, чтобы посмотреть результаты внесения изменений.

Браузер вывода узла матрицы

Браузер матрицы выводит данные комбинационной таблицы и позволяет выполнять операции с матрицей, включая выбор ячеек, копирование всей или части матрицы в буфер обмена, генерирование на основе выбора матрицы новых узлов, сохранение матрицы и вывод ее на печать. Браузер матрицы можно также использовать для отображения на экране вывода определенных моделей, таких как модели наивного критерия Байеса в Oracle.

В меню Файл и Правка предоставляются обычные опции для печали, сохранения и экспорта вывода, а также для выбора и копирования данных. Дополнительную информацию смотрите в разделе “Просмотр вывода” на стр. 286.

Хи-квадрат. Для комбинационной таблицы двух категориальных полей ниже в таблице выводится также глобальный хи-квадрат Пирсона. Этот критерий указывает вероятность того, что два поля не связаны, на основе разницы между наблюденными количествами и количествами, ожидаемыми вами, если взаимосвязь отсутствует. Например, при отсутствии взаимосвязи между удовлетворенностью покупателя и положением магазина можно ожидать похожие показатели удовлетворенности для всех магазинов. Но если покупатели в определенных магазинах постоянно сообщают о более высоких показателях, чем в других, можно подозревать, что это не случайное совпадение. Чем больше разница, тем ниже вероятность того, что это результат случайной ошибки отдельной выборки.

- Критерий хи-квадрат показывает вероятность того, что два поля не связаны; в этом случае все различия между наблюденными и ожидаемыми частотами будут результатом отдельной случайности. Если эта вероятность очень мала (обычно меньше 5%), то говорят, что взаимосвязь между двумя полями значительная.
- При наличии всего одной строки или столбца (односторонний критерий хи-квадрат) число степеней свободы равно числу ячеек минус единица. Для двустороннего критерия хи-квадрат число степеней свободы равно числу строк минус единица умножить на число столбцов минус единица.
- Будьте осторожны, интерпретируя статистику хи-квадрат, если какие-либо из ожидаемых частот в ячейках будут меньше пяти.
- Критерий хи-квадрат доступен только для комбинационной таблицы двух полей. (Если на вкладке Параметры выбрана опция **Все флаговые** или **Все числовые**, этот критерий не выводится.)

Меню Создать. Меню Создать содержит операции генерирования узлов. Эти операции доступны только для матриц комбинационных таблиц, причем в матрице нужно выбрать хотя бы одну ячейку.

- **Выбрать узел.** Генерирует узел выбора, который выбирает записи, соответствующие любой выбранной ячейке в матрице.
- **Узел извлечения (флаг).** Генерирует узел извлечения для создания нового флагового поля. Флаговое поле содержит *T* для записей, соответствующих любой выбранной ячейке в матрице, и *F* для остальных записей.
- **Узел извлечения (набор).** Генерирует узел извлечения для создания нового номинального поля. Номинальное поле содержит по одной категории для каждого непрерывного набора выбранных ячеек в матрице.

Узел Анализ

Узел Анализ позволяет оценить возможность модели генерировать точные предсказания. Узлы Анализ выполняют различные операции сравнения между предсказанными значениями и фактическими значениями (вашим полем назначения) для одного или нескольких слепков моделей. Узлы Анализ могут также использоваться для выполнения сравнения одних предсказательных моделей с другими.

При вызове узла анализа сводка результатов анализа автоматически добавляется в раздел Анализ на вкладке Сводка для каждого слепка модели в выполненном потоке. Подробные результаты анализа появляются на вкладке Поля вывода окна менеджера или могут быть записаны непосредственно в файл.

Примечание: Поскольку узлы Анализ выполняют сравнение предсказанных значений с фактическими значениями, их полезно использовать только с предсказательными моделями (теми, которым требуется поле назначения). Для неконтролируемых моделей, таких как алгоритмы кластеризации, фактические результаты, доступные для использования в качестве основы сравнения, отсутствуют.

Вкладка Анализ узла Анализ

На вкладке Анализ можно задать подробности анализа.

Матрицы совпадений (для символических или категориальных назначений). Шаблон совпадений между каждым сгенерированным (предсказанным) полем и его полем назначения для категориальных назначений (флаговых, номинальных или порядковых). Выводится таблица со строками, определенными по фактическим значениям, столбцами, определенными по предсказанным значениям, и числом записей, каждая ячейка которой содержит этот шаблон. Это полезно для выявления в предсказании систематических ошибок. Если с выходным полем связано несколько сгенерированных полей, но сгенерированных по разным моделям, наблюдения, где эти поля согласованы и несогласованы, подсчитываются, и выводятся их итоговые количества. Для наблюдений, где они согласованы, выводится другой набор правильных/неправильных статистик.

Оценка производительности. Показывает статистику оценки производительности для моделей с категориальными выходными полями. Эта статистика (представляемая для каждой категории выходных полей) является мерой среднего объема информации (в битах) из модели для предсказания числа записей, принадлежащих к данной категории. Она учитывает трудность проблемы классификации, поэтому точные предсказания для редко встречающихся категорий будут получать более высокий индекс оценки производительности, чем точные предсказания для распространенных категорий. Если модель дает результаты не лучше, чем приблизительная оценка для категории, индекс оценки производительности для этой категории будет равен 0.

Показатели оценки (AUC и Gini; только для двоичных классификаторов). Для двоичных классификаторов эта опция возвращает показатели оценки AUC (area under curve - площадь под кривой) и коэффициент Джини. Оба эти показателя вычисляются вместе для каждой двоичной модели. Значения этих показателей выводятся в таблице в браузере вывода анализа.

Показатель оценки AUC вычисляется как площадь под кривой ROC (receiver-operator characteristic curve - график зависимости чувствительности от частоты ложно положительных заключений) и является скалярным представлением ожидаемого выполнения классификатора. Значение AUC всегда находится между 0 и 1, и чем оно больше, тем лучше классификатор. Диагональная кривая ROC между координатами (0,0) и (1,1) представляет случайный классификатор со значением AUC 0,5. Поэтому у реального классификатора не будет значения AUC меньше 0,5.

Показатель оценки коэффициент Джини иногда используется как показатель оценки, альтернативный AUC; эти две меры тесно связаны. Коэффициент Джини вычисляется как двойная площадь между кривой ROC и диагональю (или как Gini = 2AUC - 1). Коэффициент Джини всегда находится между 0 и 1, и чем он больше, тем лучше классификатор. При маловероятном условии, что кривая ROC находится ниже диагонали, коэффициент Джини будет отрицательным.

Показателе совпадений (если доступны). Для моделей, генерирующих поле доверительной вероятности, эта опция сообщает статистику по значениям доверительной вероятности и их взаимосвязи с предсказаниями. У этой опции есть два параметра:

- **Порог для.** Сообщает уровень доверительной вероятности, выше которого точность будет равной заданному проценту.
- **Повысить точность.** Сообщает уровень, выше которого точность будет повышена в соответствии с заданным показателем. Например, при общей точности 90% и заданном для этой опции значении 2,0 сообщаемое значение будет соответствовать доверительной вероятности, требуемой для точности 95%.

Найти предсказанные поля / поля предикторов, применив. Определяет способ сопоставления предсказанных полей с исходным полем назначения.

- **Метаданные выходных полей модели.** Сопоставляет предсказанные поля с полем назначения на основе информации о полях модели, допуская сопоставление, даже если предсказанное поле было переименовано. К информации о полях модели для любого предсказанного поля можно также обратиться из диалогового окна Значения при помощи узла Тип. Дополнительную информацию смотрите в разделе “Использование диалогового окна Значения” на стр. 134.

- **Формат имени поля.** Сопоставляет поля на основе соглашения об именовании. Например, предсказанные значения, сгенерированные по слепку модели C5.0 для назначения с именем *отклик*, должны находиться в поле с именем *$C-отклик*.

Разделить по разделам. Если для разбиения записей на обучающую, контрольную и проверочную выборки используется поле раздела, выберите эту опцию, чтобы результаты выводились для каждого раздела по отдельности. Дополнительную информацию смотрите в разделе “Узел раздела” на стр. 167.

Примечание: При разделении по разделам записи с пустыми значениями в поле раздела исключаются из анализа. Это никогда не становится проблемой, если используется узел Раздел, поскольку узлы Раздел не генерируют пустых значений.

Пользовательский анализ. Вы можете задать свое собственное аналитическое вычисление, используемое для оценки моделей. При помощи выражений CLEM задайте, что следует вычислить для каждой записи и как объединить оценки на уровне записей в общую оценку. При помощи функций **@TARGET** и **@PREDICTED** создайте ссылку на значение назначения (фактическое выходное значение) и предсказанное значение соответственно.

- **If.** Если потребуется использование других вычислений, зависящих от некоторого условия, задайте условное выражение.
- **Then.** Задайте это вычисление, если условие If = true.
- **Else.** Задайте это вычисление, если условие If = false.
- **Использовать.** Выберите статистику для вычисления общей оценки по отдельным оценкам.

Анализ порогов по полям. Показывает категориальные поля, доступные для разложения анализа. В дополнение к общему анализу будет представлен отдельный анализ для каждой категории каждого поля разложения.

Браузер вывода анализа

В браузере вывода анализа выводятся результаты выполнения узла Анализ. Обычные опции сохранения, экспорта и печати доступны в меню Файл. Дополнительную информацию смотрите в разделе “Просмотр вывода” на стр. 286.

При первом просмотре вывода анализа результаты развернуты. Чтобы скрыть результаты после их просмотра, сверните отдельные результаты, которые вы хотите скрыть, при помощи находящегося слева от них элемента управления расширением или сверните все результаты, нажав кнопку **Свернуть все**. Чтобы снова увидеть результаты после их сворачивания, выведите нужные результаты при помощи находящегося слева от них элемента управления расширением или выведите все результаты, нажав кнопку **Развернуть все**.

Результаты для выходного поля. Вывод анализа содержит раздел для каждого выходного поля, для которого существует соответствующее поле предсказания, созданное обобщенной моделью.

Сравнение. В разделе выходного поля есть подраздел для каждого поля предсказания, связанного с этим выходным полем. Для категориальных выходных полей верхний уровень этого раздела содержит таблицу, показывающую число и процент правильных и неправильных предсказаний и общее число записей в потоке. Для числовых выходных полей в этом разделе выводится следующая информация:

- **Минимальная ошибка.** Показывает минимальную ошибку (разность между наблюденными и предсказанными значениями).
- **Максимальная ошибка.** Показывает максимальную ошибку.
- **Средняя ошибка.** Показывает среднеарифметическое (среднее) для ошибок по всем записям. Оно указывает, существует ли в модели систематическое **смещение** (более сильная тенденция к переоценке, чем к недооценке, и наоборот).
- **Средняя абсолютная ошибка.** Показывает среднее для абсолютных значений ошибок по всем записям. Указывает среднюю величину ошибки, независимо от направления.

- **Среднеквадратичное отклонение.** Показывает среднеквадратичное отклонение ошибок.
- **Линейная корреляция.** Показывает линейную корреляцию между предсказанными и фактическими значениями. Этот статистический показатель изменяется между –1,0 и 1,0. Значения, близкие к +1,0, указывают на сильную положительную связь, поэтому большие предсказанные значения связаны с малыми фактическими значениями. Значения, близкие к -1,0, указывают на сильную отрицательную связь, поэтому большие предсказанные значения связаны с малыми фактическими значениями и наоборот. Значения, близкие к 0,0, указывают на слабую связь, поэтому предсказанные значения почти независимы от фактических значений. *Примечание*: Пробельная запись здесь указывает, что вычисление линейной корреляции в данном случае невозможно, поскольку либо фактические, либо предсказанные значения - константы.
- **Вхождения.** Показывает число записей, использованных при анализе.

Матрица совпадений. Для категориальных входных полей, если в опциях анализа затребована матрица совпадений, здесь появится подраздел, содержащий эту матрицу. Строки представляют фактические наблюденные значения, а столбцы - предсказанные значения. В ячейке таблицы указывается число записей для каждого сочетания предсказанного и фактического значений.

Оценка производительности. Для категориальных входных полей, если в опциях анализа затребована статистика оценки производительности, здесь появятся результаты оценки производительности. Каждая выходная категория указывается с соответствующей ей статистикой оценки производительности.

Отчет о значениях доверительной вероятности. Для категориальных входных полей, если в опциях анализа затребованы значения доверительной вероятности, здесь появятся эти значения. Для значений доверительной вероятности модели сообщаются следующие статистики:

- **Диапазон.** Показывает диапазон значений доверительной вероятности (наименьшее и наибольшее значения) для записей в данных потока.
- **Среднее для правильных.** Показывает среднюю доверительную вероятность для записей, которые классифицируются правильно.
- **Среднее для неправильных.** Показывает среднюю доверительную вероятность для записей, которые классифицируются неправильно.
- **Всегда верно выше.** Показывает порог доверительной вероятности, выше которого предсказания всегда верны, и процент наблюдений, соответствующих этому критерию.
- **Всегда неверно ниже.** Показывает порог доверительной вероятности, ниже которого предсказания всегда неверны, и процент наблюдений, соответствующих этому критерию.
- **X% точности выше.** Показывает уровень доверительной вероятности, при котором точность равна *X%*. *X* приблизительно равно значению, заданному в опциях анализа для опции **Порог для**. Для некоторых моделей невозможно выбрать значение доверительной вероятности, соответствующее точному порогу, заданному в опциях (обычно из-за кластеров схожих наблюдений с одинаковым значением доверительной вероятности, близким к порогу). Сообщаемый порог - это ближайшее значение к заданному критерию точности, который может быть получен с одним порогом значений доверительной вероятности.
- **X-кратное повышение точности.** Показывает значение доверительной вероятности, при котором точность в *X* раз выше, чем для всего набора данных. *X* равно значению, заданному в опциях анализа для опции **Повысить точность**.

Согласуемость между. Если в поток включены несколько сгенерированных моделей, предсказывающих одно и то же выходное поле, вы увидите также статистику по **согласуемости** между предсказаниями, сгенерированными моделью. В нее включается число и процент записей, для которых предсказания согласуются (для категориальных выходных полей), или статистику сводки ошибок (для непрерывных выходных полей). Для категориальных полей в нее включается анализ предсказаний, сравненных с фактическими значениями для поднабора записей, по которым модели согласуются (генерируют одно и то же предсказанное значение).

Показатели оценки. Для двоичных классификаторов; если в опциях анализа затребованы показатели оценки, в этом разделе выводятся значения показателей оценки AUC и коэффициента Джини. В таблице содержится по одной строке для каждой модели двоичного классификатора. Таблица показателей оценки выводится не для каждой модели, а для каждого выходного поля.

Узел Аудит данных

Узел Аудит данных обеспечивает первое всестороннее представление данных, перемещаемых в IBM SPSS Modeler, в виде удобной для чтения матрицы, данные которой можно отсортировать и использовать для генерирования полноразмерных диаграмм и ряда узлов для подготовки данных.

- На вкладке Аудит выводится отчет, содержащий итожащие статистики и диаграммы распределения, которые могут помочь получить предварительное представление о данных. С этим отчетом выводится также значок хранения перед именем поля.
- На вкладке Качество в отчете аудита выводится информация о выбросах, экстремумах и пропущенных значениях и предлагаются инструменты для обработки этих значений.

Использование узла Аудит данных

Узел Аудит данных можно присоединить непосредственно к узлу источника или нисходящему потоку с инстанциированного узла типа. Можно также сгенерировать ряд узлов подготовки данных на основе результатов. Например, можно сгенерировать узел Фильтр, исключающий поля со слишком большим числом пропущенных значений, что полезно при моделировании, и сгенерировать надузел, импутирующий пропущенные значения для каких-либо или всех остающихся полей. Именно здесь реализуется полезная мощность аудита, позволяющая не только оценить состояние данных, но и выполнить на основе этой оценки соответствующее действие.

Скрининг или выборка данных. Поскольку начальный аудит особенно эффективен при работе с “большими данными”, узел выборки можно использовать для сокращения времени обработки при предварительном исследовании посредством выбора только поднабора данных. Узел Аудит данных можно также использовать в сочетании с другими узлами, такими как Выбор возможностей и Выявление аномалий, на изыскательских этапах анализа.

Вкладка Параметры узла Аудит данных

На кладке Параметры можно задать параметры для аудита.

По умолчанию. Можно просто присоединить этот узел к потоку и нажать кнопку **Выполнить**, чтобы сгенерировать отчет аудита для всех полей на основе значений параметров по умолчанию следующим образом:

- Если параметры узла Тип отсутствуют, в отчет включаются все поля.
- При наличии параметров узла Тип (независимо от того, инстанциированы ли они) в вывод включаются все *входные* поля, поля *назначения* и поля *двойного назначения*. При наличии одного поля *назначения* используйте его в качестве поля наложения. Если задано несколько полей *назначения*, наложение по умолчанию не задается.

Использовать пользовательские поля. Выберите эту опцию, чтобы выбрать поля вручную. При помощи кнопки Средство выбора полей, находящейся справа, выберите поля по отдельности либо по типам.

Поле наложения. Поле наложения используется для построения диаграмм миниизображений, выводимых в отчете аудита. В случае непрерывного поля (числового диапазона) вычисляется также двумерные статистики (ковариационная и корреляционная). Если представлено одно поле *назначения* на основе параметров узла Тип, оно используется в качестве поля наложения по умолчанию, как описано выше. Альтернативно можно выбрать опцию **Использовать пользовательские поля**, чтобы задать наложение.

Вывод. Позволяет указать, будут ли доступны диаграммы в выводе, и выбрать статистику, выводимую по умолчанию.

- **Диаграммы.** Выводит диаграмму для каждого выбранного поля: диаграмму распределения (столбчатую), гистограмму или диаграмму рассеяния, сообразно используемым данным. Диаграммы выводятся в начальном отчете как миниизображения, но можно также сгенерировать полноразмерные диаграммы и узлы диаграмм. Дополнительную информацию смотрите в разделе “Браузер вывода аудита данных” на стр. 300.
- **Базовая/расширенная статистика** Задает уровень статистики, включаемой в вывод по умолчанию. Этот параметр определяет начальный вывод на экран, но в выводе доступна вся статистика, независимо от этого параметра. Дополнительную информацию смотрите в разделе “Вывести статистики” на стр. 300.

Медиана и мода. Вычисляет медиану и моду для всех полей в отчете. Имейте в виду, что при больших наборах данных эти статистики могут увеличить время обработки, поскольку их вычисление занимает больше времени по сравнению с другими. Если вычисляется только медиана, записанное значение в некоторых случаях может быть основано на выборке 2000 записей (а не на всем наборе данных). Эта выборка выполняется на уровне отдельных полей в тех случаях, где в противном случае были бы превышены пределы памяти. Когда действует выборка, результаты помечаются в этом качестве в выводе (*Выборочная медиана*, а не просто *Медиана*). Все остальные статистики всегда вычисляются по всему набору данных.

Пустые поля или поля без типа. Если поля без типа используются с инстанциированными данными, в отчет аудита они не включаются. Чтобы включить в отчет поля без типа (в том числе пустые поля), выберите **Очистить все значения** на любых узлах Тип восходящего потока. Эта опция гарантирует, что данные не будут инстанциированы, что приведет к включению в отчет всех полей. Например, это может быть полезно, если вы хотите получить полный список всех полей или сгенерировать узел фильтра, который исключит пустые поля. Дополнительную информацию смотрите в разделе “Фильтрация полей при исследовании данных” на стр. 303.

Вкладка Качество аудита данных

Вкладка Качество на узле Аудит данных содержит опции для обработки пропущенных значений, выбросов и значений экстремумов.

Пропущенные значения

- **Число записей с допустимыми значениями.** Выберите эту опцию, чтобы вывести число записей с допустимыми значениями для каждого оцененного поля. Имейте в виду, что пустые (не определенные) значения, пробельные значения, пробелы и пустые строки всегда обрабатываются как недопустимые значения.
- **Пороговое число записей с недопустимыми значениями.** Выберите эту опцию, чтобы вывести число записей с каждым типом допустимых значений для каждого поля.

Выбросы и значения экстремумов

Метод определения выбросов и значений экстремумов. Поддерживаются два метода:

Среднеквадратичное отклонение от среднего значения. Определяет выбросы и экстремумы на основе числа среднеквадратичных отклонений от среднего значения. Например, если у вас есть поле со средним значением 100 и среднеквадратичным отклонением 10, вы должны задать 3,0 чтобы любое значение меньше 70 или больше 100 обрабатывалось как выброс.

Межквартильная зона. Определяет выбросы и экстремумы на основе межквартильной зоны, представляющей собой диапазон, в который попадают две центральных квартили (между 25-й и 75-й процентилями). Например, на основе значения параметра по умолчанию 1,5 нижний порог для выбросов будет равен Q1 – 1,5 * IQR, а верхний порог будет равен Q3 + 1,5 * IQR. Имейте в виду, что использование этой опции может понизить производительность для больших наборов данных.

Браузер вывода аудита данных

Браузер аудита данных - это мощный инструмент для получения обзора данных. На вкладке Аудит выводятся диаграммы миниизображений, значки хранения и статистика для всех полей, а на вкладке Качество выводится информация о выбросах, экстремумах и пропущенных значениях. На основе исходных диаграмм и итожащих статистик можно решить перекодировать числовое поле, получить новое поле или переклассифицировать значения номинального поля. Возможно также, вы захотите использовать в дальнейших исследованиях более современные средства визуализации. Для этого в меню Создать справа от браузера отчетов аудита можно создать любое число узлов, которые будут использоваться для преобразования или визуализации данных.

- Отсортируйте столбцы, щелкнув по нужному заголовку столбца, или измените их порядок посредством перетаскивания Поддерживается большинство стандартных операций вывода. Дополнительную информацию смотрите в разделе “Просмотр вывода” на стр. 286.
- Просмотрите для полей значения и диапазоны, щелкнув дважды по нужному полю в столбцах Измерение или Уникальные.
- При помощи панели инструментов или меню Правка можно вывести или скрыть метки значений, а также выбрать статистику, которую вы хотите вывести. Дополнительную информацию смотрите в разделе “Вывести статистики”.
- Проверьте значки хранения слева от имен полей. Хранение описывает способ хранения данных в поле. Например, в поле со значениями 1 и 0 хранятся целочисленные данные. В этом состоит отличие от шкалы измерений, которая описывает использование данных и не влияет на хранение. Дополнительную информацию смотрите в разделе “Задание форматирования и системы хранения для полей” на стр. 8.

Просмотр и генерирование диаграмм

Если никакое наложение не выбрано, на вкладке Аудит выводятся либо столбчатые диаграммы (для номинальных или флаговых полей), либо гистограммы (для непрерывных полей).

Для наложения номинального или флагового поля диаграммы раскрашиваются значениями цвета наложения.

Для наложения непрерывного поля вместо одномерных столбчатых диаграмм и гистограмм генерируются двумерные диаграммы рассеяния. В этом случае ось *x* отображается на поле наложения, что позволяет видеть одну и ту же шкалу на всех осях *x* при чтении таблицы сверху донизу.

- Для флаговых и номинальных полей установите указатель мыши на столбик, чтобы базовое значение или метка появились в подсказке.
- Для флаговых или номинальных полей при помощи панели инструментов переключите ориентацию диаграмм миниизображений с горизонтальной на вертикальную.
- Чтобы сгенерировать полноразмерную диаграмму из какого-либо миниизображения, щелкните дважды мышью по миниизображению или выберите **Вывод диаграммы** в меню Создать. *Примечание*: Если диаграмма миниизображения была создана на основе данных выборки, сгенерированная диаграмма будет содержать все наблюдения при условии, что поток исходных данных будет все еще открыт.

  Диаграмму можно сгенерировать, только если узел Аудит данных, создавший вывод, соединен с потоком.
- Чтобы сгенерировать совпадающий узел диаграммы, выберите одно или несколько полей на вкладке Аудит и выберите в меню Создать опцию **Узел диаграммы**. Полученный узел будет добавлен на холст потока, и его можно будет использовать для пересоздания диаграммы при каждом запуске потока.
- Если в наборе наложений более 100 значений, генерируется предупреждение, и наложение не включается.

Вывести статистики

В диалоговом окне Вывести статистики можно выбрать статистики, выводимые на вкладке Аудит. Исходные параметры задаются на узле Аудит данных. Дополнительную информацию смотрите в разделе “Вкладка Параметры узла Аудит данных” на стр. 298.

Минимум. Наименьшее значение числовой переменной.

Максимум. Наибольшее значение числовой переменной.

Sum. Сумма или итог для всех значений по всем наблюдениям, имеющим непропущенные значения.

Range . Разность между наибольшим и наименьшим значениями числовой переменной; максимум минус минимум.

Mean. Мера центральной тенденции. Арифметическое среднее; сумма, деленная на число наблюдений.

Стандартная ошибка среднего. Мера того, как сильно могут отличаться значения среднего от выборки к выборке, извлекаемых из одного и того же распределения. Можно применять для грубого сравнения наблюденного среднего с гипотетическим значением (то есть можно заключить, что два значения различаются, если отношение их разности к стандартному отклонению меньше -2 или больше +2).

стандартное отклонение. Мера разброса вокруг среднего, равная квадратному корню из дисперсии. Стандартное отклонение измеряется в тех же единицах, что и исходная переменная.

Дисперсия. Мера дисперсии относительно среднего значения. Равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньшее числа наблюдений. Дисперсия измеряется в единицах, которые равны квадратам единиц измерения самой переменной.

Асимметрия. Мера асимметрии распределения. Нормальное распределение симметрично, и для него асимметрия равна 0. Распределение со значимой положительной асимметрией имеет длинный хвост справа. Распределение со значимой отрицательной асимметрией имеет длинный хвост слева. В качестве грубого правила можно сказать, что значение асимметрии, более чем вдвое превышающее ее стандартную ошибку, указывает на наличие асимметрии распределения.

Стандартная ошибка асимметрии . Отношение асимметрии к ее стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение асимметрии указывает на длинный правый хвост (распределения); большое отрицательное значение - на длинный левый хвост.

Эксцесс. Мера сгруппированности наблюдений вокруг центральной точки. Для нормального распределения значение эксцесса равно 0. Положительный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы более плотно около центра и имеют более тонкие хвосты до экстремумов распределения, и более толстые хвосты в области экстремальных значений. Отрицательный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы менее плотно около центра и имеют более толстые хвосты до экстремумов распределения, и более тонкие хвосты в области экстремальных значений.

Стандартная ошибка эксцесса . Отношение эксцесса к его стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение эксцесса указывает, что хвосты распределения длиннее, чем у нормального; отрицательное значение эксцесса указывает на более короткие хвосты (как у равномерного распределения).

Уникальные. Оцениваются все эффекты одновременно, каждый эффект корректируется по всем остальным эффектам любого вида.

Допустимо. Допустимые наблюдения не содержат ни системных, ни пользовательских пропущенных значений. Имейте в виду, что пустые (не определенные) значения, пробельные значения, пробелы и пустые строки всегда обрабатываются как недопустимые значения.

Медиана. Значение, выше и ниже которого попадает по половине наблюдений, иначе 50-й процентиль. Если число наблюдений четно, медиана есть арифметическое среднее двух находящихся в середине значений, если выборку упорядочить по убыванию или по возрастанию. Медиана представляет собой меру центральной

тенденции, которая нечувствительна к выбросам, в отличие от среднего значения, которое могут исказить несколько экстремально больших или малых значений.

Мода. Чаще всего встречающееся значение. Если таких значений несколько, каждое из них является модой.

Имейте в виду, что по умолчанию медиана и мода подавляются для повышения производительности, но их можно выбрать на вкладке Параметры узла Аудит данных. Дополнительную информацию смотрите в разделе “Вкладка Параметры узла Аудит данных” на стр. 298.

Статистики для наложений

Если используется непрерывное поле наложения (числового диапазона), будут также доступны следующие статистики:

Ковариация. Ненормированная мера связи между двумя переменными, равная сумме попарных произведений отклонений, деленной на N-1.

Вкладка Качество браузера аудита данных

На вкладке Качество в браузере аудита данных выводятся результаты анализа качества данных и можно задать способы обработки выбросов, экстремумов и пропущенных значений.

Импутация пропущенных значений: В отчете аудита выводится процент полных записей для каждого поля наряду с числом допустимых, пустых и пробельных значений. Можно выбрать импутацию пропущенных значений для конкретных полей нужным вам образом, а затем сгенерировать надузел для применения этих преобразований.

1. В столбце **Импутировать пропущенные** задайте тип значений, которые вы хотите импутировать (если они есть). Можно выбрать импутацию пробельных значений или/и пустых значений либо задать пользовательское условие или выражение, выбирающее значение для импутации.

   Существует несколько типов пропущенных значений, распознаваемых IBM SPSS Modeler:

   - **Пустые или системные пропущенные значения.** Это те нестроковые значения, которые были оставлены в виде пробельных символов в базе данных или файле источника, а не были особо определены как "пропущенные" на узле источника или типа. Системные пропущенные значения выводятся как **$null$**. Имейте в виду, что пустые строки не рассматриваются как пустые значения (NULL) в IBM SPSS Modeler, хотя и могут обрабатываться как пустые значения некоторыми базами данных.
   - **Пустые строки и пробельные значения.** Значения пустых строк и пробельные значения (строки без видимых символов) обрабатываются не так, как пустые значения (NULL). Пустые строки обрабатываются для большинства целей как эквивалент пробельным значениям. Например, если выбрать опцию для обработки пробельных значений как пробельных символов на узле источника или типа, эта настройка будет также применена и к пустым строкам.
   - **Пустые или пользовательские пропущенные значения.** Существуют значения (например, `unknown`, `99` или `-1`), которые определяются на узле источника или узле типа явным образом как пропущенные значения. Дополнительно можно выбрать опцию обработки пустых значений (NULL) и пробельных значений как пробельных символов, что позволит пометить их флагами для специальной обработки и исключить из большинства вычислений. Например, функция `@BLANK` позволяет обрабатывать эти значения (также, как и пропущенные значения других типов) как пробельные символы.

2. В столбце **Метод** задайте метод, который вы хотите использовать.

   Для импутации пропущенных значений доступны следующие методы:

   Фиксированная. Подставляет фиксированное значение (среднее значение поля, среднюю точку диапазона либо задаваемую вами константу).

   Переменный. Подставляет случайное значение на основе нормального или равномерного распределения.

   Выражение. Позволяет задать пользовательское выражение. Например, можно заменить значения на глобальную переменную, созданную узлом задания глобальных значений.

Алгоритм. Подставляет значение, предсказанное моделью на основе алгоритма C&RT. Для каждого поля, импутированного этим методом, будет отдельная модель C&RT наряду с узлом заполнения, заменяющим пробелы и пустые значение на значение, предсказанное моделью. Затем узел заполнения будет использован для удаления сгенерированных моделью полей предсказания.

3. Чтобы сгенерировать надузел пропущенных значений, в меню выберите:

   Создать > Надузел пропущенных значений

   Откроется диалоговое окно Надузел пропущенных значений.

4. Выберите **Все поля** или **Только выбранные поля** и задайте размер выборки (по желанию). (Выборка задается в процентах; по умолчанию выбирается 10% всех записей.)
5. Нажмите кнопку **OK**, чтобы добавить сгенерированный надузел на холст потока.
6. Присоедините надузел к потоку, чтобы применить преобразования.

На надузле должным образом используется сочетание узлов слепка модели, заполнения и фильтра. Чтобы понять, как оно работает, можно отредактировать надузел и нажать кнопку **Увеличить**; можно также добавить, отредактировать или удалить конкретные узлы для надузла, чтобы точно настроить его поведение.

Обработка значений выбросов и экстремумов: Отчет аудита возвращает число выбросов и экстремумов для каждого поля на основе опций обнаружения, заданных на узле аудита данных. Дополнительную информацию смотрите в разделе “Вкладка Качество аудита данных” на стр. 299. Можно выбрать подавление, отбрасывание или аннулирование этих значений для конкретных полей нужным вам образом, а затем сгенерировать надузел для применения выбранных преобразований.

1. В столбце **Действие** задайте обработку выбросов и экстремумов для конкретных полей нужным вам образом.

   Для обработки выбросов и экстремумов доступны следующие действия:

   - **Принуждать.** Заменяет значения выбросов и экстремумов на ближайшее значение, которое не считается экстремальным. Например, если выброс определен как значение, превышающее по модулю три среднеквадратичных отклонения, все выбросы будут заменены наибольшим или наименьшим значением в этом диапазоне.
   - **Отклонить.** Отбрасывает записи со значениями выбросов или экстремумов для заданного поля.
   - **Аннулировать.** Заменяет выбросы и экстремумы на пустое или системное пропущенное значение.
   - **Подавлять выбросы / отбрасывать экстремумы.** Отбрасывает только значения экстремумов.
   - **Подавлять выбросы / аннулировать экстремумы.** Аннулирует только значения экстремумов.

2. Чтобы сгенерировать этот надузел, в меню выберите:

   Создать > Надузел выбросов и экстремумов

   Откроется диалоговое окно Надузел выбросов.

3. Выберите **Все поля** или **Только выбранные поля**, а затем нажмите кнопку **OK**, чтобы добавить сгенерированный надузел на холст потока.
4. Присоедините надузел к потоку, чтобы применить преобразования.

Дополнительно можно отредактировать надузле и увеличить масштаб для просмотра или внесения изменений. Значения на этом надузле отбрасываются, подавляются или аннулируются при помощи ряда узлов выбора и/или фильтра соответствующим образом.

Фильтрация полей при исследовании данных: В браузере аудита данных можно создать новый узел фильтра на основе результатов анализа качества при помощи опции Сгенерировать фильтр диалогового окна Качество.

Режим. Выберите нужную операцию для заданных полей: **Включить** или **Исключить**.

- **Выбранные поля.** Узел фильтра будет включать/исключать поля, выбранные на вкладке Качество. Например, можно отсортировать таблицу по столбцу **% заполнения**, выбрать, удерживая клавишу Shift, поля с наименьшим заполнением, а затем сгенерировать узел фильтра, исключающий эти поля.
- **В полях с процентом качества выше.** Узел фильтра будет включать/исключать поля с процентом заполнения записей выше заданного порога. Порог по умолчанию - 50%.

Фильтрация пустых полей или полей без типа

Имейте в виду. что после инстанциации значений данных поля без типа ил пустые поля исключаются из результатов аудита и из большей части остального вывода в IBM SPSS Modeler. Эти поля игнорируются для целей моделирования, но могут увеличивать в объеме или перегружать данные. Из браузера аудита данных можно также сгенерировать узел фильтра, удаляющий эти поля из потока.

1. Чтобы убедиться, что все поля включены в аудит, включая пустые поля или поля без типа, нажмите кнопку **Очистить все значения** в источнике восходящего потока или на узле типа либо задайте значения <*Передать*> для всех полей.
2. В браузере аудита данных отсортируйте столбец **% заполнения**, выберите поля с нулевыми допустимыми значениями (или каким-либо другим порогом) и при помощи меню Создать сгенерируйте узел фильтра, который может быть добавлен в поток.

Выбор записей при исследовании данных: В браузере аудита данных можно создать новый узел выбора на основе результатов анализа качества.

1. В браузере аудита данных выберите вкладку Качество.
2. Выберите в меню:

   Создать > Узел выбора пропущенных значений

   Откроется диалоговое окно Сгенерировать узел выбора.

Выбрать, если запись. Укажите , следует ли сохранять записи, если они **Допустимы** или **Недопустимы**.

Искать недопустимые значения:. Задайте, где проверять недопустимые значения.

- **Все поля.** Узел выбора будет проверять все поля на недопустимость значений.
- **В полях, выбранных в таблице.** Узел выбора будет проверять только поля, выбранные в текущий момент в выходной таблице Качество.
- **В полях с процентом качества выше.** Узел выбора будет проверять поля с процентом заполнения записей выше заданного порога. Порог по умолчанию - 50%.

Считать запись недопустимой, если недопустимое значение находится в. Задайте условие для идентификации записи как недопустимой.

- **В любом из вышеуказанных полей.** Узел выбора будет считать запись недопустимой, если *любое* из полей, указанных выше, содержит недопустимое значение для этой записи.
- **Во всех вышеуказанных полях.** Узел выбора будет считать запись недопустимой, если *все* поля, указанные выше, содержат недопустимые значения для этой записи.

Генерирование других узлов для подготовки данных

Ряд узлов, используемых при подготовке данных, можно сгенерировать непосредственно из браузера аудита данных, включая узлы переклассификации, разделения на интервалы и узлы извлечения. Например:

- Вы можете получить новое поле на основе значений *claimvalue* и *farmincome*, выбрав оба эти значения в отчете аудита и выбрав в меню Создать опцию**Извлечь**. Новый узел будет добавлен на холст потока.
- Подобным образом вы можете определить (на основе результатов аудита), что запись *farmincome* в интервалы на основе процентилей обеспечит более узконаправленный анализ. Чтобы сгенерировать узел разделения на интервалы, выберите в выводе на экран строку поля и в меню Создать выберите опцию **Разделение на интервалы**.

После добавления сгенерированного узла на холст потока его нужно присоединить к потоку и открыть, чтобы задать опции для выбранных полей.

Узел преобразования

Нормализация входных полей - важный шаг перед использованием традиционных методов скоринга, таких как регрессия, логистическая регрессия и дискриминантный анализ. Эти методы несут в себе предположения о нормальном распределении данных, не соответствующие действительности по отношению ко многим файлам необработанных данных. Один из подходов при работе с реальными данными предполагает применение к элементам необработанных данных преобразований, приближающих их распределение к нормальному. Кроме того, нормализованные поля можно легко сравнить с другими полями; так, например, прибыть и возраст - совершенно различные шкалы в файле необработанных данных, но после нормализации относительное влияние каждого из них легко поддается интерпретации.

Узел преобразования предоставляет средство просмотра вывода, позволяющее выполнить быструю визуальную оценку преобразований и определить оптимальное преобразование для использования. Сразу же можно увидеть, нормальное ли распределение у переменных, и при необходимости выбрать нужное преобразование и применить его. Можно выбрать сразу несколько полей и применить одно преобразование к каждому из них.

Выбрав для полей предпочитаемые преобразования, можно сгенерировать выполняющие их узлы извлечения и заполнения и присоединить эти узлы к потоку. Узел извлечения создает новые поля, тогда как узел заполнения преобразует существующие. Дополнительную информацию смотрите в разделе “Генерирование графиков” на стр. 307.

Вкладка Поля узла преобразования

На вкладке Поля можно указать, какие поля данных вы хотите использовать для просмотра возможных преобразований и их применения. Допускается преобразование только числовых полей. Нажмите кнопку выбора полей и выберите одно или несколько числовых полей в появившемся списке.

Вкладка Опции узла преобразования

На вкладке Опции можно задать тип преобразований, которые вы хотите подключить. Возможен выбор включить все доступные преобразования или выбрать их по отдельности.

В последнем случае можно также ввести номер для смещения данных для обратного и логарифмического преобразований. Эта операция полезна в ситуациях, где большая доля нулевых значений в данных может привести к смещению результатов по среднему и среднеквадратичному отклонению.

Например, у вас есть поле с именем *BALANCE*, содержащим несколько нулевых значений, и вы хотите применить для него обратное преобразование. Чтобы избежать нежелательного смещения, можно выбрать функцию **Обратная величина (1/x)**, а в поле **Использовать смещение данных** ввести 1. (Имейте в виду, что это смещение не связано с выполняемым функцией последовательности `@OFFSET` в IBM SPSS Modeler.)

Все формулы. Указывает, что все доступные преобразования должны быть вычислены и показаны в выводе.

Выбрать формулы. Позволяет выбрать различные преобразования для вычисления и представления в выводе.

- **Обратная величина (1/x).** Указывает, что в выводе должно быть показано обратное преобразование.
- **Логарифм (log n).** Указывает, что в выводе должно быть показано преобразование $\log_n$.
- **Логарифм (log 10).** Указывает, что в выводе должно быть показано преобразование $\log_{10}$.
- **Экспоненциальная.** Указывает, что в выводе должно быть показано экспоненциальное преобразование (e^x).
- **Кв. корень.** Указывает, что в выводе должно быть показано преобразование, извлекающее квадратный корень.

Вкладка Вывод узла преобразования

На вкладке Вывод можно задать формат и положение вывода. Можно выбрать вывод результатов на экран или отправку их в файл одного из стандартных типов. Дополнительную информацию смотрите в разделе “Вкладка Вывод узлов вывода” на стр. 290.

Средство просмотра вывода узла преобразования

Средство просмотра вывода позволяет просмотреть результаты выполнения узла преобразования. Средство просмотра - это мощный инструмент, выводящий на экран несколько преобразований для каждого поля в виде миниизображений преобразования, позволяя вам быстро сравнить поля. Опции его меню Файл можно использовать для операций сохранения, экспорта и вывода на печать. Дополнительную информацию смотрите в разделе “Просмотр вывода” на стр. 286.

Для каждого преобразования (иного чем Выбранное преобразование) выводится легенда ниже в формате:

```
Mean (Standard deviation)
```

Генерирование узлов для преобразований

Средство просмотра вывода предоставляет удобную отправную точку для подготовки данных. Например, можно нормализовать поле *AGE* (возраст) для возможности использования методов скоринга (таких как логистическая регрессия или дискриминантный анализ), предполагающих нормальное распределение. На основе исходных диаграмм и итожащих статистик можно решить преобразовать поле *AGE* в соответствии с конкретным распределением (например, логарифмическим). После выбора предпочитаемого преобразования можно сгенерировать узел извлечения со стандартным преобразованием, чтобы использовать его для скоринга.

Находясь в средстве просмотра, можно сгенерировать следующие узлы операций с полями:

- Произвести от
- Заполнитель

Узел извлечения создает новые поля с нужными преобразованиями, тогда как узел заполнения преобразует существующие поля. Эти узлы помещаются на холст в форме надузла.

Если вы выберете одно и то же преобразование для различных полей, узел извлечения или узел заполнения будет содержать формулы для преобразования этого типа для всех полей, к которым применяется это преобразование. Предположим, что вы выбрали поля и преобразования (показанные в следующей таблице), чтобы сгенерировать узел извлечения.

Таблица 39. Пример генерирования узла извлечения.

Поле	**Преобразование**
AGE	Текущее распределение
INCOME	Логарифмическое
OPEN_BAL	Обратное
BALANCE	Обратное

Надузел содержит следующие узлы:

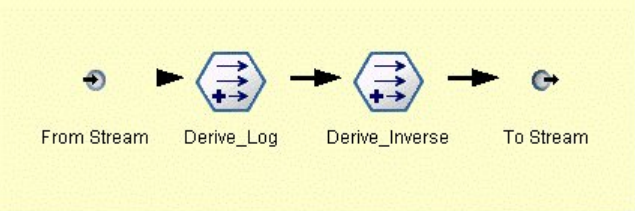


Рисунок 64. Надузел на холсте

В этом примере узел Derive_Log содержит формулу логарифмического преобразования для поля *INCOME*, а узел Derive_Inverse содержит формулы обратных преобразований для полей *OPEN_BAL* и *BALANCE*.

Чтобы сгенерировать узел

1. Для каждого поля в средстве просмотра вывода выберите нужное преобразование.
2. В меню Создать выберите нужную вам опцию **Узел извлечения** или **Узел заполнения**.

   Соответственно откроется диалоговое окно Сгенерировать узел извлечения или Сгенерировать узел заполнения.

   Выберите **Нестандартизованное преобразование** или **Стандартизированное преобразование**, как вам требуется. Вторая опция применяет к преобразованию *z*-оценку; *z*-оценка представляет значения в виде функции расстояния от среднего значения переменной в среднеквадратичных отклонениях. Например, если вы применяете логарифмическое преобразование к полю *AGE* и выбираете стандартизованное расстояние, заключительное уравнение для сгенерированного узла будет следующим:

   ```
   (log(AGE)-Mean)/SD
   ```

После того, как узел будет сгенерирован и появится на холсте потока:

1. Присоедините его к потоку.
2. Для надузла дополнительно щелкните дважды по узлу, чтобы просмотреть его содержимое.
3. Необязательно: щелкните дважды по узлу извлечения или заполнения, чтобы изменить опции для выбранных полей.

Генерирование графиков: Из гистограммы миниизображения в средстве просмотра вывода можно сгенерировать вывод полноразмерной гистограммы.

Чтобы сгенерировать диаграмму

1. Щелкните дважды кнопкой мыши по диаграмме миниизображения в средстве просмотра вывода.

   или

   Выберите диаграмму миниизображения в средстве просмотра вывода.
2. В меню Создать выберите **Вывод диаграммы**.

Гистограмма появится на экране с наложением кривой нормального распределения. Это наложение позволяет сравнить, насколько близко каждое доступное преобразование соответствует нормальному распределению.

Примечание: Диаграмму можно сгенерировать, только если узел преобразования, создавший вывод, соединен с потоком.

Другие операции: В менеджере вывода вы также можете:

- Отсортировать сетку вывода по столбцу Поле.
- Экспортировать вывод в файл HTML. Дополнительную информацию смотрите в разделе “Экспорт вывода” на стр. 288.

Узел статистики

Узел статистики предоставляет основную сводную информацию о числовых полях. Можно получить итожащие статистики для отдельных полей и корреляций между полями.

Вкладка Параметры узла статистики

Проверить. Выберите одно или несколько полей, для которых вам нужны отдельные итожащие статистики. Можно выбрать несколько полей.

Статистики. Выберите статистику для отчета. Доступны опции **Количество**, **Среднее**, **Сумма**, **Минимум**, **Максимум**, **Диапазон**, **Дисперсия**, **Среднеквадратичное отклонение**, **Среднекв. ошибка среднего**, **Медиана** и **Мода**.

Корреляции. Выберите одно или несколько полей, которые вы хотите коррелировать. Можно выбрать несколько полей. После выбора полей корреляции в выводе будет указываться корреляция между полем Проверить и полем (полями) корреляции.

Параметры корреляции. Можно задать опции для отображения на экране силы корреляции в выводе.

Параметры корреляции

IBM SPSS Modeler позволяет охарактеризовать корреляции описательными метками, помогая выделить важные взаимосвязи. **Корреляция** измеряет силу взаимосвязи между двумя непрерывными полями (числового диапазона). Она принимает значения между –1,0 и 1,0. Значения, близкие к +1,0, указывают на сильную положительную связь при условии, что высокие значения для одного поля связаны с высокими значениями для другого, а низкие значения связаны с низкими значениями. Значения, близкие к -1,0, указывают на сильную отрицательную связь при условии, что высокие значения для одного поля связаны с низкими значениями для другого и наоборот. Значения, близкие к 0,0, указывают на слабую связь при условии, что значения для двух полей почти независимы.

При помощи диалогового окна Параметры корреляции можно управлять выводом на экран меток корреляции, изменять пороги, определяющие категории, и изменять метки, используемые для каждого диапазона. Поскольку способ оценки значений корреляции в значительной мере зависит от предметной области, вы можете настроить диапазоны и метки в соответствии с конкретной ситуацией.

Показывать метки силы корреляции в выводе. Эта опция выбирается по умолчанию. Чтобы опустить описательные метки в выводе, отключите эту опцию.

Сила корреляции. Для определения и ранжирования метками силы корреляций предусмотрены две опции:

- **Определить силу корреляции по важности (1-p).** Ранжирует метками значения корреляции на основе важности, определяемой как 1 минус значимость или 1 минус вероятность, которой может быть объяснена разница в средних по отдельной случайности. Чем ближе значение к 1, тем больше вероятность, что два поля *не* независимы (или, другими словами, что между ними существует некоторая взаимосвязь). Ранжирование метками корреляций на основе важности в общем случае рекомендуется по абсолютному значению, поскольку оно объясняет изменчивость в данных; например, коэффициент 0,6 может быть высоко значимым в одном наборе данных и совсем не значимым в другом. По умолчанию значения важности от 0,0 до 0,9 ранжируются меткой *слабая*, от 0,9 до 0,95 0 - меткой *средняя* и от 0,95 до 1,0 - меткой *сильная*.
- **Определить силу корреляции по абсолютному значению.** Ранжирует метками корреляции на основе абсолютного значения коэффициента корреляции Пирсона, который изменяется от -1 до +1, как описано выше. Чем ближе абсолютное значение этой меры к 1, тем сильнее корреляция. По умолчанию корреляции от 0,0 до 0,3333 (по модулю) ранжируются меткой *слабая*, от 0,3333 до 0,6666 - меткой *средняя* и от 0,6666 до 1,0 - меткой *сильная*. Однако имейте в виду, что значимость отдельно взятого значения из одного набора данных трудно обобщить на другой набор; по этой причине в большинстве случаев рекомендуется определение корреляций на основе не абсолютного значения, а вероятности.

Браузер Statistics Output

Браузер статистики выводит результаты статистического анализа и позволяет выполнять операции, включая выбор полей, генерирование новых узлов на основе выбора, сохранение результатов и их вывод на печать. Обычные опции сохранения, экспорта и печати доступны в меню Файл, а обычные опции редактирования доступны в меню Правка. Дополнительную информацию смотрите в разделе “Просмотр вывода” на стр. 286.

При первом просмотре вывода статистики результаты развернуты. Чтобы скрыть результаты после их просмотра, сверните отдельные результаты, которые вы хотите скрыть, при помощи находящегося слева от них элемента управления расширением или сверните все результаты, нажав кнопку **Свернуть все**. Чтобы снова увидеть результаты после их сворачивания, выведите нужные результаты при помощи находящегося слева от них элемента управления расширением или выведите все результаты, нажав кнопку **Развернуть все**.

Вывод содержит раздел для каждого поля *Проверить*, где находится таблица затребованной статистики.

- **Частота.** Число записей с допустимыми для поля значениями.
- **Среднее значение.** Среднеарифметическое (среднее) для поля по всем записям.
- **Сумма.** Сумма значений для поля по всем записям.
- **Минимум.** Минимально допустимое значение для поля.
- **Максимум.** Максимально допустимое значение для поля.
- **Диапазон.** Разность между максимальным и минимальным значениями.
- **Дисперсия.** Мера изменчивости в значениях поля. Она вычисляется путем получения разницы между каждым значением и общим средним, возведения ее в квадрат, суммирования по всем значением и деления на число записей.
- **Среднеквадратичное отклонение.** Другая мера изменчивости в значениях поля, вычисляемая как квадратный корень дисперсии.
- **Среднеквадратичная ошибка среднего.** Мера изменчивости в оценке среднего значения поля, если среднее предполагается применять к новым данным.
- **Медиана.** "Срединное" значение для поля; то есть значение, отделяющее верхнюю половину данных от нижней (на основе значений поля).
- **Мода.** Наиболее распространенное отдельное значение в данных.

Корреляции. Если вы задали поля корреляции, вывод будет также содержать раздел с корреляцией Пирсона между полем Проверить и каждым полем корреляции и необязательными описательными метками для значений корреляции. Дополнительную информацию смотрите в разделе “Параметры корреляции” на стр. 308.

Меню Создать. Меню Создать содержит операции генерирования узлов.

- **Фильтр.** Генерирует узел фильтра для отфильтровывания полей, некоррелированных или слабо коррелированных с другими полями.

Генерирование узла фильтра на основе статистики

Узел фильтра, сгенерированный из браузера вывода статистики, будет фильтровать поля на основе их корреляций с другими полями. Работая, он упорядочивает корреляции по абсолютным значениям, принимает самые высокие корреляции (в соответствии с критерием, заданным в диалоговом окне Фильтр на основе статистики) и создает фильтр, который пропускает все поля, представленные в любой из этих высоких корреляций.

Режим. Решите, как выбирать корреляции. **Включать** - поля, представленные в указанных корреляциях, будут сохраняться. **Исключать** - поля будут фильтроваться.

Включать/исключать поля, представленные. Определите критерий для выбора корреляций.

- **Максимальное число корреляций.** Выбирает указанное число корреляций и включает/исключает поля, представленные в любой из этих корреляций.
- **Максимальный процент корреляций.** Выбирает указанный процент (*n%*) корреляций и включает/исключает поля, представленные в любой из этих корреляций.
- **В корреляциях выше, чем.** Выбирает корреляции по модулю выше указанного порога.

Узел средних

Узел средних выполняет сравнение средних значений между независимыми группами либо между парами связанных полей, чтобы проверить, не существует ли между ними значимая разница. Например, можно сравнить средние прибыли до и после рекламной кампании или сравнить прибыль от покупателей, не получивших рекламы, с прибылью от тех, кто ее получил.

Средние значения можно сравнить двумя различными способами в зависимости от данных:

- **Между группами в поле.** Чтобы сравнить независимые группы, выберите проверяемое поле и поле группировки. Например, можно исключить выборку "контрольных" покупателей при отправке рекламы и сравнить средние поступления от контрольной группы с поступлениями от всех остальных. В этом случае можно задать одно проверяемое поле, указывающее прибыль от каждого покупателя, с флаговым или номинальным полем, указывающим, получали ли они предложение. Эти выборки независимы в том, что каждая запись назначается в ту или иную группу, и отсутствует какой-либо способ связи конкретных элементов одной группы к конкретными элементами другой. Можно также задать номинальное поле с более чем двумя значениями, чтобы сравнить средние значения для нескольких групп. При вызове узел вычисляет критерий однофакторного дисперсионного анализа для выбранных полей. В тех случаях, когда существует только две группы полей, результаты однофакторного дисперсионного анализа фактически одинаковы с результатами проверки *t*-критерия для независимых выборок. Дополнительную информацию смотрите в разделе “Сравнение средних для независимых групп”.
- **Между парами полей.** При сравнении связанных полей группы должны быть некоторым образом объединены в пары, чтобы результаты были значимы. Например, можно сравнить средние поступления от одной и той же группы покупателей до и после проведения рекламной кампании либо сравнить нормы потребления услуг между парами муж-жена чтобы посмотреть различия между ними. Каждая запись содержит две отдельные, но связанные меры, которые можно сравнить по существу. При вызове узел вычисляет *t*-критерий для парных выборок для каждой выбранной пары полей. Дополнительную информацию смотрите в разделе “Сравнение средних между объединенными в пары полями”.

Сравнение средних для независимых групп

На узле средних выберите **Между группами в поле**, чтобы сравнить среднее для нескольких независимых групп.

Поле группировки. Выберите числовые флаговые или номинальные поля с несколькими отличительными значениями, подразделяющими записи на группы, которые вы хотите сравнить, например, на получивших и не получивших предложение. Независимо от числа проверяемых полей можно выбрать только одно поле группировки.

Проверяемые поля. Выберите одно или несколько числовых полей, содержащих показатели, которые вы хотите проверить. Для каждого выбранного вами поля будет проведена отдельная проверка. Например, можно проверить влияние данной рекламной кампании на использование, прибыль и отток.

Сравнение средних между объединенными в пары полями

Выберите на узле средних опцию **Между парами полей**, чтобы сравнить средние между отдельными полями. Чтобы результаты были значимы, эти поля должны быть некоторым образом связаны. Можно также выбрать несколько пар полей.

Первое поле. Выберите числовое поле, содержащее первый из показателей, которые вы хотите сравнить. При анализе до начала и после окончания процесса это должно быть поле "до начала процесса".

Второе поле. Выберите второе поле, которое вы хотите сравнить.

Добавить. Добавьте выбранную пару в список пар проверяемых полей.

Повторите выбор полей нужным вам образом, чтобы добавить в этот список несколько пар.

Параметры корреляции. Позволяет задать опции для ранжирования метками силы корреляции. Дополнительную информацию смотрите в разделе “Параметры корреляции” на стр. 308.

Опции узла средних

На вкладке Опции можно задать *p*-значения порогов, позволяющие пометить результаты как важные, пограничные или маловажные. Можно также отредактировать метку для каждого ранга. Важность измеряется в процентах и в целом может быть определена как 1 минус вероятность получения результата (например, разницы в средних между двумя полями) по степени крайности не ниже наблюденного результата отдельной случайности. Например, *p*-значение больше 0,95 указывает на 5% случаев, результат которых может быть объяснен отдельной случайностью.

Метки важности. Здесь можно отредактировать метки, используемые для каждой пары или группы полей в выводе. Метки по умолчанию - *важное*, *пограничное* и *маловажное*.

Значения отсечения. Задает порог для каждого ранга. Обычно *p*-значения больше 0,95 ранжируются как важные, а меньше 0,9 - как маловажные, но эти пороги можно настроить, как требуется.

Примечание: Показатели важности доступны на ряде узлов. Конкретные вычисления зависят от узла и типа используемых полей назначения и входных полей, но их значения все равно можно сравнить, поскольку все они измеряются в процентах.

Браузер вывода узла средних

Браузер вывода средних выводит данные комбинационной таблицы и позволяет выполнять стандартные операции, включая выбор и копирование по одной строке таблицы за раз, сортировку по любому столбцу, сохранение таблицы и вывод ее на печать. Дополнительную информацию смотрите в разделе “Просмотр вывода” на стр. 286.

Конкретная информация в этой таблице зависит от типа сравнения (сравнение групп в поле или отдельных полей).

Сортировать по. Позволяет отсортировать вывод по конкретному столбцу. Кнопка со стрелкой вверх или вниз изменяет направление сортировки. Можно также, щелкнув по заголовку любого столбца, выполнить сортировку по этому столбцу. (Чтобы измерить направление сортировки, щелкните по нему еще раз.)

Представление. Выбор опции **Простое** или **Расширенное** позволяет управлять уровнем подробностей вывода на экран. В расширенное представление включается вся информация из простого представления, но с дополнительными заданными подробностями.

Сравнение групп вывода средних в поле

При сравнении групп в поле имя поле группировки выводится над таблицей вывода, а средние значения и связанные статистики сообщаются отдельно для каждой группы. В эту таблицу включается отдельная строка для каждого проверяемого поля.

Выводятся следующие столбцы:

- **Поле.** Список имен выбранных проверяемых полей.
- **Средние по группе.** Выводит среднее для каждой категории поля группировки. Например, можно сравнить получивших специальное предложение (*Новая рекламная кампания*) с неполучившими (*Среднеквадратичное*). В расширенном предоставлении выводятся также среднеквадратичное отклонение, среднеквадратическая ошибка и количество.

- **Важность.** Выводит значение и метку важности. Дополнительную информацию смотрите в разделе “Опции узла средних” на стр. 311.

Расширенный вывод

В расширенном представлении выводятся следующие дополнительные столбцы:

- **F-критерий.** Этот критерий основан на отношении дисперсии между группами и дисперсии в каждой группе. Если средние для всех групп одинаковы, можно ожидать, что отношение *F* будет близким к 1, поскольку обе оценки относятся к той же дисперсии генеральной совокупности. Чем больше это отношение, тем выше дисперсия между группами и тем больше вероятность существования значимой разницы.
- **df.** Выводит число степеней свободы.

Сравнение пар полей вывода средних

При сравнении отдельных полей в таблицу вывода включается строка для каждой выбранной пары полей.

- **Первое поле/второе поле.** Выводит имя первого и второго поля в каждой паре. В расширенном предоставлении выводятся также среднеквадратичное отклонение, среднеквадратическая ошибка и количество.
- **Первое среднее/второе среднее.** Выводится среднее для каждого поля соответственно.
- **Корреляция.** Измеряется сила взаимосвязи между двумя непрерывными полями (числового диапазона). Значения, близкие к +1,0, указывают на сильную положительную связь, а значения, близкие к -1,0, указывают на сильную отрицательную связь. Дополнительную информацию смотрите в разделе “Параметры корреляции” на стр. 308.
- **Разность средних.** Выводится разница между двумя средними значениями полей.
- **Важность.** Выводится значение и метка важности. Дополнительную информацию смотрите в разделе “Опции узла средних” на стр. 311.

Расширенный вывод

Расширенный вывод добавляет следующие столбцы:

95% доверительный интервал. Нижняя и верхняя границы диапазона, в который вероятней всего попадет среднее значение true 95% всех возможных выборок этого размера из данной совокупности.

T-критерий. Статистика *t*-критерия вычисляется путем деления разности средних на ее среднеквадратическую ошибку. Чем больше у этой статистики значение по модулю, тем выше вероятность, что средние неодинаковы.

df. Выводит число степеней свободы для статистики.

Узел отчета

Узел отчета позволяет создавать форматированные отчеты, содержащие фиксированный текст, а также данные и другие выражения, полученные на основе данных. Формат отчета задается при помощи текстовых шаблонов, определяющих фиксированный текст и конструкции для вывода данных. Можно задать пользовательское форматирование, введя теги HTML в шаблоне и задав опции на вкладке Вывод. Значения данных и другой условный вывод включаются в отчет при помощи выражений CLEM в шаблоне.

Варианты замены узла отчета

Узел отчета чаще всего используется для составления списка записей или наблюдений из вывода потока, например, всех записей, соответствующих определенному условию. В этом смысле его можно считать менее структурированной альтернативой узлу таблицы.

- Если вам нужен отчет, возвращающий информацию о полях, или каких-либо других элементах, определенных в потоке, а не сами данные (такие как определения полей, заданные на узле типа), вместо него можно использовать сценарий.
- Для генерирования отчета, который содержит несколько объектов вывода (таких как собрание моделей, таблиц и диаграмм, сгенерированных одним или несколькими потоками) и который можно вывести в нескольких форматах (включая текстовый, HTML и Microsoft Word/Office), можно использовать проект IBM SPSS Modeler.
- Для генерирования списка имен полей без применения сценариев можно использовать узел таблицы, применив предварительно узел выборки, который отбросит все записи. Сгенерированную при этом таблицу без строк можно транспонировать при экспорте, чтобы получить список имен полей в одном столбце. (Для этого выберите опцию **Транспонировать данные** вкладки Вывод на узле таблицы.)

Вкладка Шаблон узла отчета

Создание шаблона. Для определения содержания отчета создается шаблон на вкладке Шаблон узла отчета. Этот шаблон состоит из текстовых строк, каждая из которых задает определенную информацию о содержимом отчета, а для указания области действия строк содержимого используются специальные строки тегов. В каждой строке содержимого перед ее отправкой в отчет оценивается выражение CLEM, заключенное в квадратные скобки ([]). Для строки в отчете существует три возможные области действия:

Фиксированная. Строки, не помеченные каким-либо образом, считаются фиксированными. Фиксированные строки копируются в отчет только один раз, после оценки выражений, которые они содержат. Например, строка

```
Это мой отчет, напечатанный [@TODAY]
```

скопирует в отчет одну строку, содержащую текст и текущую дату.

Глобальная (итерация ALL) Строки, находящиеся между специальными тегами #ALL и #, копируются в отчет один раз для каждой записи входных данных. Выражения CLEM (заключенные в квадратные скобки) оцениваются на основе текущей записи для каждой строки вывода. Например, строки

```
#ALL
Для записи [@INDEX] значение AGE - [AGE]
#
```

будет включать по одной строке для каждой записи, указывающей номер строки и возраст.

Чтобы сгенерировать список всех записей:

```
#ALL
[Возраст] [Пол] [Холестерин]  [АД]
#
```

Условная (итерация WHERE). Строки, находящиеся между специальными тегами #WHERE <*условие*> и #, копируются в отчет один раз для каждой записи, где выполняется (оценивается как true) заданное условие. Условие представляет собой выражение CLEM. (В условии WHERE квадратные скобки необязательны.) Например, строки

```
#WHERE [SEX = 'M']
Лицо мужского пола в записи номер [@INDEX] в возрасте [AGE].
#
```

будет записываться по одной строке для каждой записи с указанным для пола значением *M*. Полный отчет будет содержать фиксированные, глобальные и условные строки, определяемые посредством применения шаблона к входным данным.

Опции для вывода или сохранения результатов можно задать при помощи вкладки Вывод, общей для различных типов узлов вывода. Дополнительную информацию смотрите в разделе “Вкладка Вывод узлов вывода” на стр. 290.

Вывод данных в формате HTML или XML

Теги HTML или XML можно включать непосредственно в шаблон для написания отчетов в каком-либо из этих форматов. Например, следующий шаблон генерирует таблицу HTML.

```
Этот отчет написан в формате HTML.
Включаются только записи, где Возраст - более 60 лет.

<HTML>
  <TABLE border="2">
   <TR>
     <TD>Возраст</TD>
     <TD>АД</TD>
     <TD>Холестерин</TD>
     <TD>Препарат</TD>
   </TR>

  #WHERE Age > 60
    <TR>
       <TD>[Возраст]</TD>
       <TD>[АД]</TD>
       <TD>[Холестерин]</TD>
       <TD>[Препарат]</TD>
    </TR>
  #
  </TABLE>
</HTML>
```

Браузер вывода узла отчета

В браузере отчета выводится содержимое сгенерированного отчета. Обычные опции сохранения, экспорта и печати доступны в меню Файл, а обычные опции редактирования доступны в меню Правка. Дополнительную информацию смотрите в разделе “Просмотр вывода” на стр. 286.

Узел задания глобальных значений

Узел задания глобальных значений просматривает данные и вычисляет сводные значения, которые могут использоваться в выражениях CLEM. Например, при помощи узла задания глобальных значений можно вычислить статистики для поля *возраст*, а затем использовать общее среднее поля *возраст* в выражениях CLEM посредством вставки функции `@GLOBAL_MEAN(age)`.

Вкладка Параметры узла задания глобальных значений

Создаваемые глобальные. Выберите одно или несколько полей для которых вы хотите сделать доступными глобальные значения. Можно выбрать несколько полей. Для каждого поля задайте статистики для вычисления, чтобы обеспечить вывод нужных вам статистик в столбцах рядом с именем поля.

- **MEAN.** Среднеарифметическое (среднее) для поля по всем записям.
- **SUM.** Сумма значений для поля по всем записям.
- **MIN.** Минимально допустимое значение для поля.
- **MAX.** Максимально допустимое значение для поля.
- **SDEV.** Среднеквадратичное отклонение, представляющее собой меру изменчивости в значениях поля и вычисляемое как квадратный корень дисперсии.

Операции по умолчанию. Выбранные здесь опции будут использоваться при добавлении новых полей в список глобальных значений выше. Чтобы изменить набор статистик по умолчанию, выберите статистики или отмените их выбор нужным вам образом. Можно также, нажав кнопку **Применить**, применить операции по умолчанию ко всем полям в списке.

Примечание: Некоторые операции неприменимы к нечисловым полям (например, Сумма для поля даты/времени). Операции, которые нельзя использовать с выбранным полем, отключаются

Очистить все глобальные перед выполнением. Выберите эту опцию, чтобы удалить все глобальные значения перед вычислением новых значений. Если эта опция не выбрана, вновь вычисляемые значения будут заменять более значения, но глобальные значения, которые не вычисляются повторно, также останутся доступы.

Показать предварительный просмотр созданных глобальных после выполнения. Если выбрать эту опцию, вкладка Глобальные значения диалогового окна свойств потока появится после выполнения для вывода на экран вычисленных глобальных значений.

Узел подгонки имитации

Узел подгонки имитации подгоняет набор статистических распределений - кандидатов по каждому полю в данных. Каждая подгонка распределения по полю оценивается при помощи критерия качества подгонки. После выполнения узла подгонки имитации строится узел генерирования имитации или обновляется существующий. Каждому полю назначается лучшее подогнанное распределение. Затем узел генерирования имитации можно использовать для генерирования данных имитации по каждому полю.

Хотя узел подгонки имитации - конечный узел, он не добавляет модель на палитру сгенерированных моделей, не добавляет ни выходных полей, ни диаграмм на вкладку Вывод и не экспортирует данные.

Примечание: Если хронологические данные отрывочны (то есть много пропущенных значений), компоненту подгонки может оказаться затруднительно найти достаточно допустимых значений для подгонки распределений по данным. Если данных мало, нужно либо удалить малонаселенные поля перед подгонкой, если они не нужны, либо выполнить импутацию пропущенных значений. При помощи опций на вкладке **Качество** на узле аудита данных можно увидеть число заполненных записей, узнать о малонаселенных полях и выбрать метод импутации. Если число записей недостаточно для подгонки распределений, его можно увеличить при помощи узла балансировки.

Использование узла подгонки имитации для автоматического создания узла генерирования имитации

После первого выполнения узла подгонки имитации создается узел генерирования имитации со ссылкой обновления от узла подгонки имитации. Если узел подгонки имитации выполняется еще раз, то новый узел генерирования имитации создается только в случае, когда ссылка обновления была удалена. При помощи узла подгонки имитации можно также обновить связанный с ним узел генерирования имитации. Результат зависит от наличия одних и тех же полей на обоих узлах и незаблокированности этих полей на узле генерирования имитации. Дополнительную информацию смотрите в разделе “Узел Генерирование имитации” на стр. 49.

У узла подгонки имитации ссылка обновления может быть только на узел генерирования имитации. Чтобы задать ссылку на узел генерирования имитации, выполните следующие действия:

1. Щелкните правой кнопкой по узлу подгонки имитации.
2. В меню выберите **Определить связь обновления**.
3. Щелкните по узлу генерирования имитации, на который будет указывать задаваемая ссылка обновления.

Чтобы удалить ссылку обновления между узлом подгонки имитации и узлом генерирования имитации, щелкните правой кнопкой по ссылке обновления и выберите **Удалить ссылку**.

Подгонка распределения

Статистическое распределение - это теоретическая частота, с которой встречаются значения некоторой переменной. На узле подгонки имитации каждое поле данных сравнивается с рядом теоретических статистических распределений. Доступные для подгонки распределения описаны в разделе “Распределения” на стр. 59. У теоретического распределения есть параметры, которые настраиваются для наилучшего совпадения с с данными по показателю критерия согласия; это критерий Anderson-Darling или критерий Kolmogorov-Smirnov. В результатах подгонки распределения на узле подгонки имитации показывается, какие распределения подгонялись, каковы наилучшие оценки параметров для каждого распределения и насколько хорошо каждое распределение отвечает данным. Во время подгонки распределения также вычисляются корреляции между полями с числовым типом данных и сопряженности между полями с категориальным распределением. Результаты подгонки распределения используются при создании узла генерирования имитации.

Чтобы подгонять распределения под данные, сначала необходимо проверить количество пропущенные значения в первой 1000 записей. Если пропущенных значений слишком много, подгонка невозможна. В этом случае вам придется выбирать из следующих возможностей:

- Использовать узел, расположенный выше по потоку, чтобы удалить записи с пропущенными значениями.
- Использовать узел, расположенный выше по потоку, для импутации значений вместо пропущенных.

Подгонка распределения не исключает пользовательские пропущенные значения. Если ваши данные содержат пользовательские пропущенные значения, которые нужно исключить из подгонки распределения, замените их на системные пропущенные значения.

При подгонке распределения не учитывается роль поля. Например, поля **назначения** обрабатываются так же, как **входные** поля, поля **без роли**, поля **с обеими ролями**, поля **разделов**, **разбиения**, **частоты** и **ID**.

Во время подгонки распределения по-разному обрабатываются поля согласно типу хранения и типу измерения. Как обрабатываются поля во время подгонки распределения, описано в следующей таблице.

Таблица 40. Подгонка распределения согласно типу хранения и типу измерения полей

<table>
<tr><th rowspan="2">Тип хранения</th><th colspan="6">Шкала измерений</th></tr>
<tr><th>Количественная</th><th>Категориальная</th><th>Флаг</th><th>Номинальная</th><th>Порядковое</th><th>Без типа</th></tr>
<tr><td>Текстовое</td><td>Невозможна</td><td colspan="3">Подгоняются категориальные распределения, распределения игральной кости и фиксированные распределения</td><td rowspan="4">Поле игнорируется и не передается на узел генерирования имитации.</td></tr>
<tr><td>Целое число</td><td rowspan="3">Подгоняются все распределения. Вычисляются корреляции и сопряженности.</td><td colspan="2" rowspan="3">Подгоняется категориальное распределение. Корреляции не вычисляются.</td><td rowspan="3">Подгоняются биномиальное распределение, отрицательное биномиальное распределение и распределение Пуассона, и вычисляются корреляции.</td></tr>
<tr><td>Действительное число</td></tr>
<tr><td>Время</td></tr>
</table>

Таблица 40. Подгонка распределения согласно типу хранения и типу измерения полей (продолжение)

Тип хранения	**Шкала измерений**			
Дата				
Метка даты/времени				
Нет данных	Тип хранения узнается по данным.			

Поля с порядковым типом измерений обрабатываются как непрерывные поля и включаются в таблицу корреляций на узле генерирования имитации. Если нужно распределение, отличное от биномиального, измените тип измерения поля на непрерывный. Если вы ранее определили метки для всех значений порядкового поля, а теперь изменяете тип измерения на непрерывный, метки будут потеряны.

Во время подгонки распределения поля с одним значением обрабатываются также, как поля с несколькими значениями. Поля с типом хранения время, дата и отметка времени обрабатываются как числовые.

Подгонка распределений по полям разбиения

Если в ваших данных содержится поле разбиения и вы хотите провести отдельную подгонку распределения для каждой части, преобразуйте данные при помощи узла реструктуризации. Пользуясь узлом реструктуризации, сгенерируйте новое поле для каждого значения поля разбиения. Затем используйте реструктурированные данные для подгонки распределения на узле подгонки имитации.

Вкладка Параметры узла подгонки имитации

Имя узла источника. Имя сгенерированного (или обновленного) узла генерирования имитации можно генерировать автоматически, выбрав опцию **Автоматически**. Автоматически сгенерированное имя - это имя, заданное в узле Подгонка имитации, если в нем задано пользовательское имя, или Sim Gen, если пользовательское имя в нем не задано. Вы можете включить переключатель **Пользовательское** и задать имя в текстовом поле рядом. Если текстовое поле не изменено, пользовательское имя по умолчанию - Sim Gen.

Опции подгонки При помощи этих опций можно задать, каким образом распределения подгоняются под поля и как эти распределения оцениваются.

- **Количество наблюдений для выборки.** Задает число наблюдений, используемых при подгонке распределений по полям из набора данных. Выберите **Все наблюдения**, чтобы подгонять распределения по всем записям в данных. Если набор данных очень большой, есть смысл ограничить число наблюдений, используемых при подгонке распределений. Выберите **Ограничить первыми N наблюдениями**, чтобы использовать только первые N наблюдений. Нажимая на стрелки, задайте число используемых наблюдений. Другой вариант - при помощи узла, расположенного выше по потоку, выполните случайную выборку записей для подгонки распределений.
- **Критерии качества подгонки (только для непрерывных полей).** В случае непрерывных полей выберите критерий Anderson-Darling или критерий Kolmogorov-Smirnoff для оценки точности подгонки распределений по полям. Критерий Anderson-Darling выбран по умолчанию и рекомендуется, если нужно гарантировать оптимальную подгонку в области хвостов распределений. Оба статистических показателя вычисляются для каждого распределения-кандидата, но только выбранный показатель используется для сортировки распределений и выявления оптимально подогнанного распределения.
- **Число интервалов (только для эмпирических распределений).** В случае непрерывных полей эмпирическое распределение - это кумулятивная функция распределения хронологических данных. Она равна вероятности каждого значения или диапазона значений и вычисляется непосредственно из данных. Пользуясь стрелками, задайте количество интервалов при вычислении эмпирического распределения для непрерывных полей. По умолчанию задано значение 100, максимальное значение - 1000.

- **Поле веса (необязательно).** Если ваши данные содержат поле весов, щелкните по значку выбора поля и выберите поле весов в списке. Поле весов будет исключено из процесса подгонки распределений. Список содержит все поля набор данных с непрерывным типом измерений. Можно выбрать только одно поле весов.

Узел оценки имитации

Узел оценки имитации - это конечный узел, на котором оценивается заданное поле; для этого поля строится распределение и генерируются диаграммы распределений и корреляций. Прежде всего этот узел предназначен для оценки непрерывных полей. Таким образом, он дополняет диаграмму оценки, генерируемую узлом оценки и предназначенную для оценки дискретных полей. Еще одно отличие узла оценки имитации в том, что он оценивает одно предсказанное поле при нескольких итерациях, тогда как узел оценки оценивает несколько предсказанных полей с одной итерацией у каждого. Итерации генерируются, когда на узле генерирования имитации для какого-либо параметра распределения задается несколько значений. Дополнительную информацию смотрите в разделе “Итерации” на стр. 59.

Узел оценки имитации предназначен для обработки данных, полученных на узлах подгонки имитации и генерирования имитации. Однако этот узел можно использовать с любыми другими узлами. Между узлом генерирования имитации и узлом оценки имитации можно вставить любое число этапов промежуточной обработки.

Важное замечание: Для работы узла оценки имитации требуется не менее 1000 записей с допустимыми значениями поля назначения.

Вкладка Параметры узла оценки имитации

На вкладке Параметры узла оценки имитации можно задать роль каждого поля в наборе данных и настроить вывод, генерируемый при имитации.

Выбрать элемент. Здесь можно переключаться между тремя панелями узла оценки имитации: Поля, Функции плотности и Поля вывода.

Панель полей

Поле назначения. Это поле - обязательное. Щелкните по стрелке выпадающего списка, чтобы выбрать поле назначения вашего набора данных. Выбрать можно поле с типом измерений непрерывное, порядковое или номинальное, но не дата и не неопределенное.

Поле итераций (необязательно). Если в ваших данных есть поле итерации, показывающее принадлежность каждой записи к той или иной итерации, выберите его здесь. Тогда каждая итерация будет оцениваться независимо. Выбирать можно только поля с типом измерений непрерывное, порядковое или номинальное.

Входные данные уже отсортированы по итерации. Доступна, только если в поле **Поле итерации (необязательно)** выбрано поле итерации. Выбирайте эту опцию только в том случае, если уверены, что входные данные уже отсортированы по полю итерации, заданному в поле **Поле итерации (необязательно)**.

Максимальное число итераций для графика. Доступна, только если в поле **Поле итерации (необязательно)** выбрано поле итерации. Нажимая на стрелки, задайте число итераций на диаграмме. Задав это число, вы избежите невнятных диаграмм с чрезмерным числом итераций. Низший уровень, который можно задать для максимального числа итераций, - это 2; высший - 50. Для диаграмм максимальное число итераций в начале задается равным 10.

Входные поля для корреляционного торнадо. Диаграмма корреляционного торнадо - это полосчатая диаграмма, содержащая коэффициенты корреляции между заданным полем назначения и каждым из заданных входных полей. Щелкните по значку пипетки поля, чтобы выбрать выходные поля, которые будут включены в диаграмму торнадо, из списка доступных входных полей имитации. Выбирать можно только

входные поля с типом измерений непрерывное или порядковое. Номинальные входные поля, входные поля без типа и входные поля даты в этом списке недоступны, и такие поля выбирать нельзя.

Панель функций плотности

При помощи опций на этой панели можно настроить вывод функций плотности вероятности (probability density function, PDF) и кумулятивных функций распределения (cumulative distribution functions, CDF) для непрерывных полей назначения, а также вывод полосчатых диаграмм предсказанных значений для категориальных полей назначения.

Функции плотности. Функции плотности являются основными средствами проверки набора результатов имитации.

- **Функция плотности вероятности (Probability density function, PDF).** Выберите эту опцию, чтобы сгенерировать функцию плотности вероятности для поля назначения. Функция плотности вероятности показывает распределение значений поля назначения. При помощи функции плотности вероятности можно определить вероятность попадания поля назначения в тот или иной диапазон. Для категориальных целевых значений (целевые значения с количественной или порядковой шкалой измерения) создается столбчатая диаграмма, в которой показан процент наблюдений, которые относятся к каждой из категорий целевого значения.
- **Кумулятивная функции распределения (CDF).** Выберите эту опцию, чтобы сгенерировать кумулятивную функцию распределения для поля назначения. Кумулятивная функция распределения показывает вероятность того, что целевое значение меньше указанного значения либо равно ему. Она доступна только для количественных целевых значений.

Опорные линии (непрерывные). Эти опции доступны, если выбрана **Функция плотности вероятности (Probability density function, PDF)**, **Кумулятивная функции распределения (CDF)** или обе. При помощи этих опций можно добавлять различные неподвижные вертикальные опорные линии для функций плотности вероятности и кумулятивных функций распределения.

- **Среднее значение.** Выберите эту опцию, чтобы добавить опорную линию по среднему значению поля назначения.
- **Медиана.** Выберите эту опцию, чтобы добавить опорную линию по медиане поля назначения.
- **Среднеквадратичные отклонения.** Выберите эту опцию, чтобы добавить опорные линии, отстоящие на указанное число стандартных отклонений в плюс и минус от среднего значения поля назначения. При выборе этой опции становится доступно расположенное рядом поле **Число**. Нажимая на стрелки, задайте число стандартных отклонений. Минимальное число стандартных отклонений - 1, максимальное - 10. В начале число стандартных отклонений задано равным 3.
- **Процентили.** Выберите эту опцию, чтобы добавить опорные линии по двум процентилям распределения поля назначения. При выборе этой опции становятся доступны расположенные рядом текстовые поля **Нижняя** и **Верхняя**. Например, если ввести значение 90 в текстовое поле **Верхняя**, добавится опорная линия на 90%-й процентили поля назначения, то есть на таком значении, ниже которого попадает 90% наблюдений. Аналогичным образом, значение 10 в текстовое поле **Нижняя** представляет десятую процентиль поля назначения, то есть значение, ниже которого попадает 10% наблюдений.
- **Настраиваемые опорные линии.** Выберите эту опцию, чтобы добавить опорные линии по заданным значениям вдоль горизонтальной оси. При выборе этой опции становится доступна расположенная рядом таблица **Значения**. Каждый раз после ввода допустимого числа в таблицу **Значения** к ней добавляется новая пустая строка в конце. *Допустимое* число должно быть в диапазоне значений поля назначения

Примечание: Когда на одной диаграмме показано несколько функций плотности или распределения (для нескольких итераций), опорные линии, кроме пользовательских, применяются к каждой функции по отдельности.

Категориальное поле назначение (только для PDF). Эти опции доступны, только если выбрана **Функция плотности вероятности (Probability density function, PDF)** is selected.

- **Значения категории для отчета.** Для моделей с категориальными полями назначения результат модели - это предсказанные для каждой категории вероятности того, что поле назначения попадет в эту категорию. Категория с самой высокой вероятностью выбирается как предсказанная категория и используется для построения полосчатой диаграммы для функции плотности вероятности. Чтобы сгенерировать полосчатую диаграмму, выберите опцию **Предсказанная категория**. Чтобы сгенерировать гистограммы распределения предсказанных вероятностей для каждой категории поля назначения, выберите опцию **Предсказанные вероятности**. Кроме того, вы можете выбрать опцию **И то, и другое**, чтобы сгенерировать оба типа диаграмм.
- **Группирование для анализа чувствительности.** Если при имитации используются итерации анализа чувствительности, то для каждой итерации, заданной в анализе, генерируется независимое поле назначения (или предсказанное поле назначения из модели). Для каждого значения варьируемого параметра распределения используется своя итерация. При наличии итераций полосчатая диаграмма предсказанной категории категориального поля назначения выводится как кластеризованная полосчатая диаграмма, содержащая результаты всех итераций. Используйте опцию **Группировать категории** или **Группировать итерации**.

Панель Поля вывода

Значения процентилей распределений поля назначения. При помощи этих опций можно создать таблицу значений процентилей распределений поля назначения и задать процентили, показываемые на экране.

Создать таблицу значений процентилей. Для непрерывных полей назначения эта опция создает таблицу заданных процентилей распределений поля назначения. Для задания процентилей служат следующие опции:

- **Квартили.** Квартили - это 25, 50 и 75 процентили распределений поля назначения. Наблюдения разбиваются на четыре группы равного размера.
- **Интервалы.** Если нужно разбиение на заданное число равновеликих групп, и число групп отлично от 4, выберите опцию **Интервалы**. При выборе этой опции становится доступно расположенное рядом поле **Число**. Нажимая на стрелки, задайте число интервалов. Минимальное число интервалов - 2, максимальное - 100. В начале число интервалов задается равным 10.
- **Задать процентили.** Выберите опцию **Настраиваемые процентили**, чтобы задать процентили по отдельности, например, задать 99-ю процентиль. При выборе этой опции становится доступна расположенная рядом таблица **Значения**. Каждый раз после ввода допустимого числа от 1 до 100 в таблицу **Значения** к ней добавляется новая пустая строка в конце.

Вывод узла оценки имитации

После выполнения узла оценки имитации сгенерированная информация выводится в менеджер вывода. Результаты выполнения узла оценки имитации показаны в браузере вывода оценки имитации. Обычные опции сохранения, экспорта и печати доступны в меню **Файл**, а обычные опции редактирования доступны в меню **Правка**. Дополнительную информацию смотрите в разделе “Просмотр вывода” на стр. 286. Меню **Вид** доступно, только если выбрана одна из диаграмм. Оно недоступно для таблицы распределения и для информационных объектов вывода. В меню **Вид** можно выбрать **Режим редактирования**, чтобы изменить макет и внешний вид диаграммы, или **Режим изучения**, чтобы показать данные и значения, представленные на диаграмме. В Статическом режиме на диаграмме фиксируются в своих текущих положениях и перестают двигаться опорные линии и ползунки. Только в статическом режиме можно скопировать, экспортировать или напечатать диаграмму вместе с опорными линиями. Чтобы переключиться в этот режим, выберите **Статический режим** в меню **Вид**.

Окно браузера вывода для оценки имитации состоит из двух панелей. В левой части окна расположена панель навигации, на которой в виде миниизображений представлены диаграммы, сгенерированные при выполнении узла оценки имитации. При выборе миниизображения соответствующая диаграмма выводится на панели в правой части окна.

Панель навигации

На панели навигации браузера вывода находятся миниизображения диаграмм, сгенерированных при имитации. Миниизображения на панели навигации зависят от типа измерения поля назначения и опций, выбранных в диалоговом окне узла оценки имитации. Описание миниизображений дано в следующей таблице.

Таблица 41. Миниизображения на панели навигации

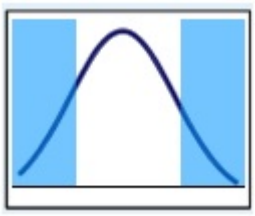
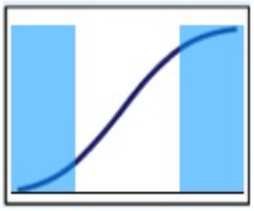
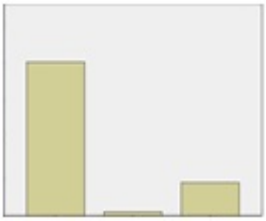
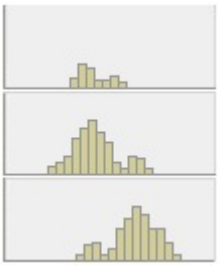
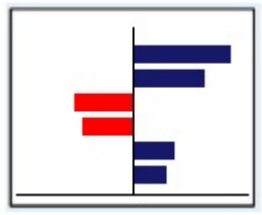
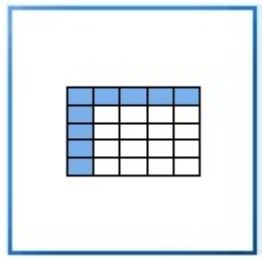
Миниизображение	Описание	Комментарии
	Функция плотности вероятности	Это миниизображение выводится, только если тип измерения поля назначения - непрерывный и на панели функций плотности в диалоговом окне узла оценки имитации выбрана **Функция плотности вероятности (probability density function, PDF)**. Если тип измерения поля назначения - категориальный, то это миниизображение не выводится.
	Кумулятивная функция распределения	Это миниизображение выводится, только если тип измерения поля назначения - непрерывный и на панели функций плотности в диалоговом окне узла оценки имитации выбрана **Кумулятивная функция распределения (CDF)**. Если тип измерения поля назначения - категориальный, то это миниизображение не выводится.
	Значения предсказанной категории	Это миниизображение выводится, только если тип измерения поля назначения - категориальный, и на панели функций плотности в диалоговом окне узла оценки имитации выбрана **Функция плотности вероятности (probability density function, PDF)**, и в области **Значения категории для отчета** выбрано **Предсказанная категория** или **Обе**. Если тип измерения поля назначения - непрерывный, то это миниизображение не выводится.
	Вероятности предсказанной категории	Это миниизображение выводится, только если тип измерения поля назначения - категориальный, и на панели функций плотности в диалоговом окне узла оценки имитации выбрана **Функция плотности вероятности (probability density function, PDF)**, и в области **Значения категории для отчета** выбрано **Предсказанные вероятности** или **Обе**. Если тип измерения поля назначения - непрерывный, то это миниизображение не выводится.

Таблица 41. Миниизображения на панели навигации (продолжение)

Миниизображение	Описание	Комментарии
	Диаграммы торнадо	Это миниизображение выводится, только если на панели полей в диалоговом окне узла оценки имитации выбрано одно или несколько полей в поле **Входные поля для корреляционного торнадо**.
	Таблица распределения	Это миниизображение выводится, только если тип измерения поля назначения - непрерывный и на панели Поля вывода в диалоговом окне узла оценки имитации включен переключатель **Создать таблицу значений процентилей**. Для этой диаграммы меню **Вид** недоступно. Если тип измерения поля назначения - категориальный, то это миниизображение не выводится.
	Информация	Это миниизображение выводится всегда. Меню **Вид** для этого объекта вывода недоступно.

Вывод диаграмм

Доступность тех или иных типов диаграмм вывода зависит от типа измерений поля назначения, от того, используется ли поле итерации, и от опций, выбранных в диалоговом окне узла оценки имитации. У ряда диаграмм, генерируемых из имитации, есть интерактивные возможности для настройки внешнего вида. Интерактивные возможности доступны, если выбрать **Опции диаграммы**. Все диаграммы имитаций являются визуализациями графической панели.

Диаграммы функций плотности вероятности для непрерывных целевых переменных. На этой диаграмме одновременно показаны вероятность и частота; шкала вероятности показана на левой вертикальной оси, а шкала частоты - на правой. Диаграмма содержит две вертикальные опорные линии с ползунками, которые делят диаграмму на отдельные области. В таблице под диаграммой указан процент распределения в каждой из этих областей. Если на диаграмме показано несколько функций плотности (из-за итераций), то для каждой из функций плотности дается своя строка соответствующих вероятностей в таблице; кроме того, таблица содержит дополнительный столбец с именем итерации и цветом соответствующей функции плотности. Итерации упорядочены в таблице по алфавиту по метке итерации. Если метка итерации недоступна, вместо нее используется значение итерации. Таблица недоступна для редактирования.

У каждой опорной линии есть ползунок (перевернутый треугольник), при помощи которого удобно перемещать линию. У каждого ползунка есть метка, показывающая текущую позицию. По умолчанию ползунки помещаются на 5-ю и 95-ю процентили распределения. Если есть несколько итераций, ползунки помещаются на 5-ю и 95-ю процентили первой итерации в таблице. При перемещении линий нельзя перевести одну линию за другую.

Ряд дополнительных возможностей становится доступен, если выбрать **Опции диаграммы**. В частности, можно явным образом задать положение ползунков, добавить неподвижные опорные линии и переключить вид диаграммы с непрерывной кривой на гистограмму. Дополнительную информацию смотрите в разделе “Параметры диаграммы” на стр. 324. Чтобы скопировать или экспортировать диаграмму, щелкните по ней правой кнопкой.

Кумулятивная функция плотности для непрерывных целевых переменных. Эта диаграмма содержит такие же две подвижные опорные линии и соответствующую таблицу, какие описаны для диаграммы функции плотности вероятности. Ползунки и таблица ведут себя так же, как для функции плотности вероятности при наличии нескольких итераций. Те же цвета, которые кодируют принадлежность каждой функции плотности определенной итерации, используются и для функций распределения.

На этой диаграмме также есть доступ к диалоговому окну Опции диаграммы, в котором можно явным образом задать положение ползунков, добавить неподвижные опорные линии и выбрать представление кумулятивной функции распределения - как возрастающей (по умолчанию) или убывающей. Дополнительную информацию смотрите в разделе “Параметры диаграммы” на стр. 324. Чтобы скопировать, экспортировать или отредактировать диаграмму, щелкните по ней правой кнопкой. Если выбрать **Редактировать**, диаграмма открывается во всплывающем окне редактора диаграмм.

Диаграмма предсказанных значений категории для категориальных полей назначения. В случае категориальных полей назначения полосчатая диаграмма содержит предсказанные значения. Предсказанные значения представлены как процент поля назначения, попадающий по прогнозу в данную категорию. В случае категориальных полей назначения с итерациями анализа чувствительности результаты для предсказанной категории назначения выводятся как кластеризованная полосчатая диаграмма, содержащая результаты для всех итераций. Диаграмма кластеризуется по категории или по итерации, в зависимости от опции, выбранной в области **Группирование для анализа чувствительности** на панели функций плотности в диалоговом окне узла оценки имитации. Чтобы скопировать, экспортировать или отредактировать диаграмму, щелкните по ней правой кнопкой. Если выбрать **Редактировать**, диаграмма открывается во всплывающем окне редактора диаграмм.

Диаграмма предсказанных вероятностей категории для категориальных полей назначения. Для категориальных полей назначения гистограмма содержит распределение предсказанных вероятностей для каждой категории поля назначения. Для категориальных полей назначения с итерациями анализа чувствительности гистограммы выводятся по категории или по итерации, в зависимости от опции, выбранной в области **Группирование для анализа чувствительности** на панели функций плотности в диалоговом окне узла оценки имитации. Если гистограммы сгруппированы по категории, то в выпадающем списке меток итераций можно выбрать нужную итерацию для вывода на экран. Кроме того, можно выбрать нужную итерацию для вывода на экран, щелкнув правой кнопкой по диаграмме и выбрав итерацию в подменю **Итерация**. Если гистограммы сгруппированы по итерации, то в выпадающем списке имен категорий можно выбрать нужную категорию для вывода на экран. Кроме того, можно выбрать нужную категорию для вывода на экран, щелкнув правой кнопкой по диаграмме и выбрав категорию в подменю **Категория**.

Эта диаграмма доступна только для подмножества моделей, причем на слепке модели должна быть выбрана опция генерировать все групповые вероятности. Например, для слепка логистической модели нужно выбрать **Добавить все вероятности**. Эту опцию поддерживают следующие слепки моделей:

- Логистическая, SVM, байесовская, нейронная сеть и метод k ближайших соседей
- Модели исследования в базе данных DB2/ISW для логистической регрессии, деревьев решений и Наивного Байеса

По умолчанию для этих слепков моделей опция генерировать все групповые вероятности не выбрана.

Диаграммы торнадо. Диаграмма торнадо - это полосчатая диаграмма, показывающая чувствительность поля назначения к каждому из заданных входных полей. Чувствительность измеряется корреляцией между полем назначения и тем или иным входным полем. Заголовок диаграммы содержит имя поля назначения. Каждая полоска на диаграмме представляет корреляцию между полем назначения и некоторым входным

полем. Имитированные входные поля на диаграмме - это входные поля, выбранные в поле **Входные поля для корреляционного торнадо** на панели полей в диалоговом окне узла оценки имитации. Каждая полоска помечена значением корреляции. Полоски отсортированы по абсолютному значению корреляции, от большего значения к меньшему. Если есть итерации, для каждой итерации генерируется отдельная диаграмма. У каждой диаграммы есть подзаголовок, содержащий имя итерации.

Таблица распределения. Эта таблица содержит значение поля назначения, ниже которого попадает заданный процент наблюдений. Таблица содержит по строке для каждой процентили, заданной на панели Поля вывода в диалоговом окне узла оценки имитации. Процентилями могут быть квартили, другое число равноотстоящих друг от друга процентилей или индивидуально заданные процентили. Таблица распределения содержит по столбцу для каждой итерации.

Информация. Этот раздел содержит общую сводку по полям и записям, используемым при оценке. Он также содержит входные поля и число записей, разбитых при каждой итерации.

Параметры диаграммы

В диалоговом окне Опции диаграмм можно настроить вывод активированных диаграмм функций плотности вероятности и кумулятивных функций распределения, сгенерированных из этой имитации.

Представление. Выпадающий список **Просмотр** относится только к диаграмме функции плотности вероятности. Служит для переключения между диаграммой с непрерывной кривой и гистограммой. Эта возможность недоступна, если на диаграмме показано несколько функций плотности (от нескольких итераций). Если функций плотности несколько, их можно просматривать только в виде непрерывных кривых.

Порядок. Выпадающий список **Порядок** относится только к диаграмме кумулятивной функции распределения. Оно указывает порядок вывода на экран функции: восходящий (по умолчанию) или убывающий. В представлении убывающей функции значение функции для каждой точки на горизонтальной оси - это вероятность того, что поле назначения лежит справа от этой точки.

Положения ползунка. Текстовое поле **Верхняя** содержит текущее положение правой опорной линии ползунка. Текстовое поле **Нижняя** содержит текущее положение левой опорной линии ползунка. Положение ползунков можно задавать явным образом, вводя значения в текстовые поля **Верхняя** и **Нижняя**. Значение в текстовом поле **Нижняя** должно быть строго меньше значения в текстовом поле **Верхняя**. Вы можете удалить левую опорную линию, нажав кнопку **-Бесконечность**, что эквивалентно сдвигу ползунка в минус бесконечность. После этого действия становится недоступно поле **Нижняя**. Вы можете удалить правую опорную линию, нажав кнопку **Бесконечность**, что эквивалентно сдвигу ползунка в бесконечность. После этого действия становится недоступно поле **Верхняя**. Вы не можете удалить обе опорных линии; если включить переключатель **-Бесконечность**, то переключатель **Бесконечность** станет недоступен, и наоборот.

Опорные линии. Вы можете добавлять различные неподвижные вертикальные опорные линии для функций плотности вероятности и кумулятивных функций распределения.

- **Среднее значение.** Вы можете добавить опорную линию по среднему значению поля назначения.
- **Медиана.** Вы можете добавить опорную линию по медиане поля назначения.
- **Среднеквадратичные отклонения.** Вы можете добавить опорные линии, отстоящие на указанное число стандартных отклонений в плюс и минус от среднего значения поля назначения. Указать число стандартных отклонений можно в текстовом поле рядом. Минимальное число стандартных отклонений - 1, максимальное - 10. В начале число стандартных отклонений задано равным 3.
- **Процентили.** Вы можете добавить одну или две опорные линии по заданным процентилям распределения поля назначения, введя значения в текстовые поля **Нижняя** и **Верхняя**. Например, если ввести значение 95 в текстовое поле **Верхняя**, будет показана опорная линия на 95-й процентили поля назначения, то есть на том уровне, ниже которого попадает 95% наблюдений. Аналогичным образом, если ввести значение 5 в текстовое поле **Нижняя**, будет показана опорная линия на 5-й процентили поля назначения, то есть на том

уровне, ниже которого попадает 5% наблюдений. Для текстового поля **Нижняя** минимальное значение процентили 0, максимальное 49. Для текстового поля **Верхняя** минимальное значение процентили 50, максимальное 100.

- **Настраиваемые позиции.** Можно добавить опорные линии в указанных значениях по горизонтальной оси. Настраиваемые опорные линии можно удалять, удаляя значения в таблице.

После нажатия кнопки **OK** все ползунки, метки над ползунками, опорные линии и таблица под диаграммой обновляются с учетом опций, выбранных в диалоговом окне Опции диаграммы. Нажмите кнопку **Отмена**, чтобы закрыть диалоговое окно без внесения изменений. Опорные линии можно удалить, если выключить соответствующие переключатели в диалоговом окне Опции диаграмм и нажать кнопку **OK**.

Примечание: Когда на одной диаграмме показано несколько функций плотности или распределения (как результат итераций при анализе чувствительности), опорные линии, кроме пользовательских, применяются к каждой функции по отдельности. Показаны только опорные линии для первой итерации. В метках опорных линий содержится метка итерации. Метка итерации создается выше по потоку, обычно на узле генерирования имитации. Если метка итерации недоступна, вместо нее используется значение итерации. Для кумулятивных функций распределения при нескольких итерациях недоступны опции **Среднее**, **Медиана**, **Стандартные отклонения** и **Процентили**.

Вспомогательные прикладные программы IBM SPSS Statistics

Если на вашем компьютере установлена и залицензирована совместимая версия IBM SPSS Statistics, IBM SPSS Modeler можно сконфигурировать для обработки данных с функциональной возможностью IBM SPSS Statistics при помощи узлов преобразования Statistics, модели Statistics, вывода Statistics или экспорта Statistics.

Информацию о совместимости с текущей версией IBM SPSS Modeler смотрите на корпоративном сайте поддержки по адресу *http://www.ibm.com/support*.

Чтобы сконфигурировать IBM SPSS Modeler для работы с IBM SPSS Statistics и другими прикладными программами, выберите:

Инструменты > Опции > Вспомогательные прикладные программы

IBM SPSS Statistics Interactive. Введите полный путь и имя команды (например: *C:\Program Files\IBM\SPSS\Statistics\<nn>\stats.exe*), которая должна использоваться при запуске IBM SPSS Statistics непосредственно для файла данных, сгенерированного узлом экспорта статистики. Дополнительную информацию смотрите в разделе “Узел Statistics Export” на стр. 360.

Подключение. Если сервер IBM SPSS Statistics находится на том же хосте, что и IBM SPSS Modeler Server, вы можете включить поддержку соединения между двумя указанными прикладными программами, что повысит эффективность, поскольку данные останутся на сервере при анализе. Выберите **Сервер**, чтобы включить опцию **Порт** ниже. Значение параметра по умолчанию - **Локальный**.

Порт. Задайте порт для сервера IBM SPSS Statistics.

Утилита определения положения IBM SPSS Statistics. Чтобы включить использование узлов Statistics Transform, Statistics Model и Statistics Output в IBM SPSS Modeler, необходимо, чтобы на компьютере, на котором выполняется поток, было установлено и лицензировано программное обеспечение IBM SPSS Statistics.

- Если программа IBM SPSS Modeler используется в локальном (автономном) режиме, лицензированная копия IBM SPSS Statistics должна быть установлена на локальном компьютере. Нажмите эту кнопку, чтобы задать положение локальной установки IBM SPSS Statistics, которую вы хотите использовать для лицензирования.

- Кроме того, при запуске в распределенном режиме на удаленном IBM SPSS Modeler Server также необходимо запустить служебную программу на хосте IBM SPSS Modeler Server для создания файла *statistics.ini*, который указывает IBM SPSS Statistics путь установки для IBM SPSS Modeler Server. Для этого в командной строке выберите каталог IBM SPSS Modeler Server *bin* и в Windows выполните команду:

  ```
  statisticsutility -location=<IBM SPSS Statistics_installation_path>/
  ```

  В UNIX выполните команду:

  ```
  ./statisticsutility -location=<IBM SPSS Statistics_installation_path>/bin
  ```

При отсутствии лицензированной копии IBM SPSS Statistics на локальном компьютере можно запустить узел Statistics File на сервере IBM SPSS Statistics, однако при попытке запустить другие узлы IBM SPSS Statistics будут возвращены сообщения об ошибке.

Комментарии

Если вы испытываете трудности с запуском узлов процедур IBM SPSS Statistics, просмотрите следующие советы:

- Если имена полей, используемые в IBM SPSS Modeler, длиннее восьми символов (для версий до IBM SPSS Statistics 12.0), длиннее 64 символов (для IBM SPSS Statistics 12.0 и последующих версий) или содержат недопустимые символы, их нужно переименовать или усечь перед передачей в IBM SPSS Statistics. Дополнительную информацию смотрите в разделе “Переименование или фильтрация полей для IBM SPSS Statistics” на стр. 361.
- Если компонент IBM SPSS Statistics устанавливался после IBM SPSS Modeler, может потребоваться задать положение IBM SPSS Statistics, как объяснялось выше.

Глава 7. Узлы экспорта

Обзор узлов экспорта

Узлы экспорта предоставляют механизм для экспорта данных в различных форматах для взаимодействия с другими используемыми программными инструментами.

Доступны следующие узлы экспорта:

Узел экспорта баз данных записывает данные в совместимый с ODBC источник реляционных данных. Чтобы произвести запись в источник данных ODBC, этот источник данных должен существовать и у вас должны быть разрешения записи для него.

Узел экспорта плоских файлов выводит данные в текстовом формате с разделителями. Он полезен для экспорта данных, которые может читать другое программное обеспечение анализа или электронных таблиц.

Узел Экспорт статистики выводит данные в формате IBM SPSS Statistics *.sav* или *.zsav*. Файлы *.sav* или *.zsav* могут быть прочитаны модулем IBM SPSS Statistics Base и другими продуктами. Это формат, используемый также для файлов кэша в IBM SPSS Modeler.

Узел экспорта Data Collection выводит данные в формате, используемом программным обеспечением изучения рынка Data Collection. Для использования этого узла должна быть установлена библиотека данных Data Collection.

Узел экспорта IBM Cognos BI экспортирует данные в формате, который могут прочесть базы данных Cognos BI.

Узел экспорта IBM Cognos TM1 экспортирует данные в формате, который могут прочесть базы данных Cognos TM1.

Узел экспорта SAS выводит данные в формате SAS, которые можно прочесть в программных пакетах SAS или SAS-совместимых. Доступно три формата файлов SAS: SAS для Windows/OS2, SAS для UNIX или SAS Версии 7/8.

Узел экспорта Excel выводит данные в формате файлов Microsoft Excel .xlsx. Дополнительно можно выбрать автоматический запуск Excel и открытие экспортированного файла при выполнении узла.

Узел экспорта XML выводит данные в файл в формате XML. Дополнительно вы можете создать узел источника XML, чтобы прочесть экспортированные данные обратно в поток.

Узел экспорта базы данных

При помощи узлов баз данных можно записать данные в совместимые с ODBC реляционные источники данных, которые объяснены в описании узла источника базы данных. Дополнительную информацию смотрите в разделе “Узел источника базы данных” на стр. 17.

Для записи данных в базу данных используйте следующие общие шаги:

1. Установите драйвер ODBC и сконфигурируйте источник данных для базы данных, которую вы хотите использовать.
2. На вкладке Экспорт узла базы данных задайте источник данных и таблицу, куда вы хотите записать данные. Можно создать новую таблицу или вставить данные в существующую.
3. Задайте дополнительные поля нужным вам образом.

Более подробно эти шаги описаны в следующих нескольких темах.

Вкладка Экспорт узла базы данных

Примечание: Некоторые из баз данных, в которые может выполняться экспорт, не поддерживают имен столбцов в таблицах длиннее 30 символов. Если появится сообщение об ошибке, указывающее, что в таблице есть неправильное имя столбца, сократите размер имени, чтобы в нем было меньше 30 символов.

Источник данных. Показывает выбранный источник данных. Введите имя или выберите его в выпадающем списке. Если вы не видите нужную базу данных в списке, выберите **Добавить новое соединение с базой данных** и найдите нужную базу данных в диалоговом окне Соединения с базами данных. Дополнительную информацию смотрите в разделе “Добавление соединения с базой данных” на стр. 19.

Имя таблицы. Введите имя таблицы, в которую вы хотите отправить данные. При выборе опции **Вставить в таблицу** можно будет выбрать существующую таблицу в базе данных, нажав кнопку **Выбрать**.

Создать таблицу. Выберите эту опцию, чтобы создать новую или переопределить существующую таблицу базы данных.

Вставить в таблицу. Выберите эту опцию, чтобы вставить данные в виде новых строк в существующую таблицу базы данных.

Объединить таблицу. Выберите эту опцию (при ее наличии), чтобы обновить выбранные столбцы базы данных значениями из соответствующих полей исходных данных. Выбор этой опции включает поддержку кнопки **Слить**, выводящей диалоговое окно, откуда поля исходных данных можно отобразить на столбцы базы данных.

Отбросить существующую таблицу. Выберите эту опцию, чтобы удалить любую существующую таблицу с таким же именем при создании новой таблицы.

Удалить существующие строки. Выберите эту опцию, чтобы удалить существующие строки из таблицы перед экспортом при операции вставки в таблицу.

Примечание: Если выбрана одна из двух опций выше, при вызове узла вы получите сообщение **Предупреждение о перезаписи**. Чтобы подавить предупреждения, выключите переключатель **Предупреждать, если узел перезаписывает таблицу базы данных** на вкладке Уведомления диалогового окна Пользовательские опции.

Размер строки по умолчанию. Поля, помеченные вами как поля без типа на узле типа восходящего потока, записываются в базу данных как строковые поля. Задайте размер строк, который должен использоваться для полей без типа.

Нажмите кнопку **Схема**, чтобы открыть диалоговое окно, где можно задать различные опции экспорта (для баз данных, поддерживающих эту возможность), задать для полей типы данных SQL и задать первичный ключ для целей индексации базы данных. Дополнительную информацию смотрите в разделе “Опции схемы экспорта базы данных” на стр. 330.

Нажмите кнопку **Индексы**, чтобы задать опции для индексации экспортированной таблицы для повышения производительности базы данных. Дополнительную информацию смотрите в разделе “Опции индексов экспорта базы данных” на стр. 332.

Нажмите кнопку **Дополнительно**, чтобы задать опции массовой загрузки и принятий базы данных. Дополнительную информацию смотрите в разделе “Дополнительные опции экспорта базы данных” на стр. 334.

Заключать в кавычки имена таблиц и столбцов. Выберите опции, используемые при отправке в базу данных оператора `CREATE TABLE`. Имена таблиц или столбцов с пробелами или нестандартными символами должны быть заключены в кавычки.

- **Когда требуется.** Выберите эту опцию, чтобы разрешить IBM SPSS Modeler автоматически определять, когда требуются кавычки, на индивидуальной основе.
- **Всегда.** Выберите эту опцию, чтобы всегда заключать имена таблиц и столбцов в кавычки.
- **Никогда.** Выберите эту опцию, чтобы отключить использование кавычек.

Сгенерировать узел импорта для этих данных. Выберите эту опцию, чтобы сгенерировать узел источника базы данных для данных, экспортируемых в заданный источник данных и таблицу. После выполнения этот узел будет добавлен на холст потока.

Опции слияния экспорта базы данных

В этом диалоговом окне можно отобразить поля из исходных данных на столбцы базы данных назначения. Там, где поле исходных данных отображается на столбец базы данных, значение столбца заменяется на значение исходных данных при обработке потока. Неотображенные исходные поля остаются в базе данных без изменений.

Отобразить поля. Здесь задается отображение между полями исходных данных и столбцами базы данных. Поля исходных данных с таким же именем, что и у столбцов в базе данных, отображаются автоматически.

- **Отобразить.** Отображает поле исходных данных, выбранное в списке полей слева от кнопки, на столбец базы данных, выбранный в списке справа. Можно отобразить сразу несколько полей, но в обоих списках должно быть выбрано одно и то же число записей.
- **Снять отображение.** Удаляет отображение для одного или нескольких столбцов базы данных. Эта кнопка активируется при выборе вами поля или столбца базы данных в таблице в правой части диалогового окна.
- **Добавить.** Добавляет одно или несколько полей исходных данных, выбранных в списке полей слева от кнопки, в список полей справа, готовых для отображения. Эта кнопка активируется при выборе вами поля в списке слева, если в списке справа поля с таким именем не существует. Нажатие этой кнопки отображает выбранное поле на новый столбец базы данных с таким же именем. Слово **<НОВОЕ>**, выводящееся после имени столбца базы данных, указывает, что это новое поле.

Объединить строки. Вы используете поле ключа, такое как *ID транзакции*, для слияния записей с одним и тем же значением в поле ключа. Это эквивалентно EQUI-объединению базы данных. Значения ключей должны представлять собой значения первичных ключей, то есть, они должны быть уникальными и не могут быть пустыми значениями.

- **Возможные ключи.** Возвращает список полей, найденных во всех входных источниках данных. Выберите одно или несколько полей из этого списка и при помощи кнопки со стрелкой добавьте их в качестве полей ключей для слияния записей. Любое поле отображения с соответствующим отображенным столбцом базы данных доступно в качестве ключа, за исключением того, что поля, добавленные в качестве новых столбцов базы данных (выводящихся со словом **<НОВОЕ>** после их имени), будут недоступны.
- **Ключи для слияния.** Возвращает список всех полей, используемых для слияния записей из входных источников данных на основе полей ключей. Для удаления ключа из списка выберите его и при помощи кнопки со стрелкой возвратите этот ключ в список Возможные ключи. При выборе нескольких полей ключей включается опция ниже.
- **Включать только записи, существующие в базе данных.** Выполняет частичное объединение; если запись находится в базе данных и потоке, отображенные поля будут обновлены.
- **Добавить записи к базе данных.** Выполняет внешнее объединение; все записи в потоке будут слиты (если такая же запись существует в базе данных) или добавлены (если соответствующая запись еще не существует в базе данных).

Чтобы отобразить поле исходных данных на новый столбец базы данных

1. Щелкните по имени исходного поля в списке слева в области **Отобразить поля**.
2. Нажмите кнопку **Добавить** для завершения отображения.

Чтобы отобразить поле исходных данных на существующий столбец базы данных

1. Щелкните по имени исходного поля в списке слева в области **Отобразить поля**.
2. Щелкните по имени столбца в области **Столбец базы данных** справа.
3. Нажмите кнопку **Отобразить**, чтобы завершить отображение.

Чтобы удалить отображение

1. В списке справа в области Поле щелкните по имени поля, отображение которого вы хотите удалить.
2. Нажмите кнопку **Снять отображение**.

Чтобы отменить выбор поля в любом из списков

Щелкните по имени поля, удерживая клавишу CTRL.

Опции схемы экспорта базы данных

В диалоговом окне Схема экспорта базы данных можно задать опции для экспорта базы данных (для баз данных, поддерживающих эти опции), задать для полей типы данных SQL, указать, какие поля будут первичными ключами, и настроить оператор `CREATE TABLE`, генерируемый после экспорта.

Это диалоговое окно состоит из нескольких частей:

- Раздел в верхней части (если он выводится) содержит опции для операции экспорта в базу данных, которая поддерживает эти опции. Если соединение с такой базой не установлено, этот раздел не выводится.
- В текстовом поле в средней части окна выводится шаблон, с помощью которого генерируется команда `CREATE TABLE`, для которой по умолчанию используется следующий формат:

  ```
  CREATE TABLE <имя-таблицы> <(столбцы таблицы)>
  ```

- Таблица в нижней части окна позволяет задать для каждого поля тип данных SQL и указать, какие поля будут первичными ключами, как описано ниже. Диалоговое окно автоматически генерирует значения параметров `<имя-таблицы>` и `<(столбцы таблицы)>` на основе спецификаций в этой таблице.

Задание опций экспорта базы данных

В этом разделе, если он выводится, можно задать ряд значений параметров для экспорта базы данных. Эта возможность поддерживается базами данных следующих типов.

- IBM InfoSphere Warehouse. Дополнительную информацию смотрите в разделе “Опции для IBM DB2 InfoSphere Warehouse”.
- SQL Server Enterprise edition и Developer edition. Дополнительную информацию смотрите в разделе “Опции для SQL Server” на стр. 332.
- Oracle Enterprise edition или Personal edition. Дополнительную информацию смотрите в разделе “Опции для Oracle” на стр. 332.

Настройка операторов CREATE TABLE

При помощи некоторых текстовых полей этого диалогового окна в оператор `CREATE TABLE` можно добавить добавочные опции для конкретных баз данных.

1. Включите переключатель **Настроить команду CREATE TABLE**, чтобы активировать текстовое окно.
2. Добавьте в оператор опции для конкретной базы данных. Обязательно сохраните текстовые параметры `<имя-таблицы>` и `(<столбцы таблицы>)`, поскольку они будут подставляться IBM SPSS Modeler в качестве фактического имени и определений столбцов таблицы.

Задание типов данных SQL

По умолчанию IBM SPSS Modeler позволяет серверу баз данных назначать типы данных SQL автоматически. Чтобы переопределить автоматический тип для поля, найдите строку, соответствующую полю, и выберите нужный тип вы выпадающем списке в столбце *Тип* таблицы схем. Можно также, удерживая нажатой клавишу Shift, выбрать несколько строк.

Для типов, принимающих аргумент длины, точности или шкалы, (`BINARY`, `VARBINARY`, `CHAR`, `VARCHAR`, `NUMERIC` и `NUMBER`) вы должны задать длину, а не разрешать серверу баз данных назначать ее автоматически. Например, задав для длины рациональное значение, такое как `VARCHAR(25)`, вы будете уверены, что соответствующий тип хранения в IBM SPSS Modeler будет перезаписан в соответствии с вашим намерением. Чтобы переопределить автоматическое назначение типа, в выпадающем списке Тип выберите **Задать** и замените определение типа соответствующим оператором определения типа SQL.

Проще всего это сделать, выбрав сначала тип, ближайший к нужному определению типа, а затем **Задать**, чтобы отредактировать определение. Например, чтобы задать тип данных SQL `VARCHAR(25)`, задайте сначала в выпадающем списке Тип **VARCHAR(length)**, а затем выберите **Задать** и замените текст length на значение 25.

Первичные ключи

Если у одного или нескольких столбцов в экспортированной таблице для каждой строки должно быть уникальное значение или сочетание значений, это можно указать, включив переключатель**Первичный ключ** для каждого применяемого поля. Большинство баз данных будут запрещать изменение таблицы способом, упраздняющим ограничение первичного ключа, и автоматически создавать индекс по первичному ключу, способствующий применению этого ограничения. (Необязательно: в диалоговом окне Индексы можно создать индексы для других полей. Дополнительную информацию смотрите в разделе “Опции индексов экспорта базы данных” на стр. 332.)

Опции для IBM DB2 InfoSphere Warehouse

Табличное пространство. Табличное пространство, которое должно использоваться для экспорта. Администраторы баз данных могут создать и сконфигурировать табличные пространства как многораздельные. Мы рекомендуем, выбрав одно или несколько табличных пространств (вместо табличного пространства по умолчанию), использовать их для экспорта базы данных.

Разделение данных по полю. Задает входное поле, которое будет использоваться для разделения.

Использовать сжатие. Эта опция, если она выбрана, создает таблицы для экспорта со сжатием (например, эквивалентного `CREATE TABLE MYTABLE(...) COMPRESS YES;` в SQL).

Опции для SQL Server

Использовать сжатие. Эта опция, если она выбрана, создает таблицы для экспорта со сжатием.

Сжатие для. Выберите уровень сжатия.

- **Строка.** Эта опция включает поддержку сжатия на уровне строк (например, эквивалентного `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW);` в SQL).
- **Страница.** Эта опция включает поддержку сжатия на уровне страниц (например, `CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);` в SQL).

Опции для Oracle

Параметры Oracle - Базовая опция

Использовать сжатие. Эта опция, если она выбрана, создает таблицы для экспорта со сжатием.

Сжатие для. Выберите уровень сжатия.

- **По умолчанию.** Эта опция включает поддержку сжатия по умолчанию (например, `CREATE TABLE MYTABLE(...) COMPRESS;` в SQL). В этом она действует также, как и опция **Базовый**.
- **Тип.** Эта опция включает поддержку базового метода сжатия (например, `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` в SQL).

Параметры Oracle - Дополнительная опция

Использовать сжатие. Эта опция, если она выбрана, создает таблицы для экспорта со сжатием.

Сжатие для. Выберите уровень сжатия.

- **По умолчанию.** Эта опция включает поддержку сжатия по умолчанию (например, `CREATE TABLE MYTABLE(...) COMPRESS;` в SQL). В этом она действует также, как и опция **Базовый**.
- **Тип.** Эта опция включает поддержку базового метода сжатия (например, `CREATE TABLE MYTABLE(...) COMPRESS BASIC;` в SQL).
- **OLTP.** Эта опция включает сжатие OLTP (например, `CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP;` в SQL).
- **Низкая/высокая для запросов.** (Только для серверов Exadata) Эта опция включает поддержку сжатия по столбцам для запросов, (например, `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW;` или `CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH;` в SQL). Сжатие для столбцов полезно в средах хранилищ данных; опция `HIGH` обеспечивает более высокую степень сжатия, чем `LOW`.
- **Низкая/высокая для архивов.** (Только для серверов Exadata) Эта опция включает поддержку сжатия по столбцам для архива, (например, `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW;` или `CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH;` в SQL). Сжатие для архивов полезно для сжатия данных, которые будут храниться длительные периоды времени; опция `HIGH` обеспечивает более высокую степень сжатия, чем `LOW`.

Опции индексов экспорта базы данных

Диалоговое окно Индексы позволяет автоматически создавать индексы на таблицах базы данных, экспортированных из IBM SPSS Modeler. Можно задать наборы полей, которые вы хотите включить, и настроить при необходимости команду `CREATE INDEX`.

Это диалоговое окно состоит из двух частей:

- В текстовом поле в верхней части выводится шаблон, с помощью которого можно сгенерировать одну или несколько команд CREATE INDEX, для которых по умолчанию используется следующий формат:

  ```
  CREATE INDEX <имя-индекса> ON <имя-таблицы>
  ```

- Таблица в нижней части этого диалогового окна позволяет добавить спецификации для каждого индекса, который вы захотите создать. Для каждого индекса задайте имя индекса и поля или столбцы для включения. Диалоговое окно автоматически генерирует значения параметров <имя-индекса> и <имя-таблицы> соответствующим образом.

  Например, сгенерированный SQL для одного индекса на полях *empid* и *deptid* мог бы выглядеть примерно так:

  ```
  CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)
  ```

  Можно добавить несколько строк, чтобы создать несколько индексов. Для каждой строки генерируется отдельная команда CREATE INDEX.

Настройка команды CREATE INDEX

Необязательно: вы можете настроить команду CREATE INDEX для всех индексов или только для конкретного индекса. Эта опция обеспечивает гибкость согласования конкретных требований или опций баз данных и применения настроек ко всем индексам или только к конкретным (при необходимости).

- Выберите **Настроить команду CREATE INDEX** в верхней части диалогового окна, чтобы изменить шаблон, используемый для всех индексов, которые будут добавлены впоследствии. Имейте в виду, что изменения не будут применены к индексам, уже добавленным в таблицу, автоматически.
- Выберите в таблице одну или несколько строк и нажмите кнопку **Изменить выбранные индексы** в верхней части диалогового окна, чтобы применить текущие настройки ко всем выбранным индексам.
- Включите переключатель **Настроить** в каждой строке, чтобы изменить шаблон команды только для этого индекса.

Имейте в виду, что значения параметров <имя-индекса> и <имя-таблицы> генерируются диалоговым окном автоматически на основе спецификаций таблицы и не могут быть отредактированы непосредственно.

КЛЮЧЕВОЕ СЛОВО BITMAP. Если используется база данных Oracle, можно настроить шаблон для создания битового индекса (bitmap) вместо стандартного следующим образом:

```
CREATE BITMAP INDEX <имя-индекса> ON <имя-таблицы>
```

Битовые индексы могут оказаться полезны для индексации столбцов с небольшим числом отличительных значений. Полученный SQL может выглядеть так:

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

ключевое слово UNIQUE. Большинство баз данных поддерживают ключевое слово UNIQUE в команде CREATE INDEX. Оно принудительно устанавливает ограничение уникальности подобное ограничению первичного ключа для базовой таблицы.

```
CREATE UNIQUE INDEX <имя-индекса> ON <имя-таблицы>
```

Имейте в виду, что для полей, фактически назначаемых в качестве первичных ключей, эта спецификация необязательна. Большинство баз данных автоматически создадут индекс для любых полей, задаваемых как поля первичного ключа в команде CREATE TABLE, поэтому создавать индексы на этих полях явным образом не обязательно. Дополнительную информацию смотрите в разделе “Опции схемы экспорта базы данных” на стр. 330.

Ключевое слово FILLFACTOR Некоторые физические параметры для индекса можно точно настроить. Например, SQL Server позволяет пользователям изменять размер индекса (после начального создания) ценой затрат на техобслуживание по мере внесения в таблицу последующих изменений.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

Другие комментарии

- Если индекс с указанным именем уже существует, создание индекса завершится неудачно. Все ошибки сначала будут обработаны как предупреждения, что разрешит создание последующих индексов, а затем обработаны повторно как сообщения об ошибках с записью в журнал сообщений после того, как будут предприняты все попытки создания индексов.
- Для оптимальной производительности индексы следует создавать после загрузки данных в таблицу. Индексы должны содержать по крайней мере один столбец.
- Перед выполнением обработки узла возможен предварительный просмотр сгенерированного SQL в журнале сообщений.
- Для временных таблиц, записываемых в базу данных, (то есть, при включенном кэшировании узлов) опции задания первичных ключей и индексов недоступны. Однако система может создавать индексы на временных таблицах соответствующим образом в зависимости от способа использования данных на узлах нисходящего потока. Например, если данные в кэше в дальнейшем будут объединены по столбцу *DEPT*, будет разумно индексировать таблицу в кэше по этому столбцу.

Индексы и оптимизация запросов

В некоторых системах управления базами данных, когда таблица базы данных уже создана, загружена и индексирована, требуется еще одна, дополнительная операция перед тем как оптимизатор сможет использовать индексы для ускорения выполнения запросов к этой новой таблице. Например, в Oracle оптимизатор запросов на основе стоимости требует, чтобы таблица была проанализирована перед тем, как ее индексы можно будет использовать в оптимизации запросов. Внутренний файл свойств ODBC для Oracle (невидимый для пользователей) содержит опцию для выполнения этой операции следующим образом:

```
# Определяется SQL, выполняемый после того, как таблица и все связанные с ней индексы
# будут созданы и заполнены
table_analysis_sql, 'ANALYZE TABLE <имя-таблицы> COMPUTE STATISTICS'
```

Этот шаг выполняется при каждом создании таблицы Oracle (независимо от того, определены ли первичные ключи или индексы). При необходимости файл свойств ODBC для дополнительных баз данных можно настроить аналогичным образом; обращайтесь за помощью в службу поддержки.

Дополнительные опции экспорта базы данных

При нажатии кнопки Дополнительные в диалоговом окне узла экспорта База данных открывается новое диалоговое окно, в котором можно задать технические подробности для экспорта результатов в базу данных.

Использовать пакетное принятие. Выберите эту опцию, чтобы отключить построчные принятия в базу данных.

Размер пакета. Задает число записей, отправляемых в базу данных перед принятием в память. Уменьшение этого числа увеличивает целостность данных ценой снижения скорости передачи. Возможно, вы захотите точно настроить это число, чтобы получить оптимальную производительность для используемой базы данных.

Опции InfoSphere Warehouse. Выводится, только если установлено соединение с базой данных InfoSphere Warehouse (IBM DB2 9.7 или новее). **Не записывать в журнал изменения** - позволяет избежать записи в журнал событий при создании таблиц и вставке данных.

Использовать массовую загрузку. Задает метод для массовой загрузки данных в базу данных непосредственно из IBM SPSS Modeler. Возможно, потребуются некоторые экспериментальные исследования по выбору опций массовой загрузки, подходящих для конкретного сценария.

- **Через ODBC.** Выберите эту опцию, чтобы использовать API ODBC для исключения многострочных вставок с большей интенсивностью, чем при обычном экспорте в базе данных. Выберите в опциях ниже связывание по строкам или по столбцам.
- **Через внешний загрузчик.** Выберите эту опцию для использования пользовательской программы массовой загрузки для данной конкретной базы данных. Выбор этой опции активирует ряд опций ниже.

Дополнительные опции ODBC. Эти опции доступны, только когда выбрана опция **Через ODBC**. Имейте в виду, что эта функциональная возможность может поддерживаться не всеми драйверами ODBC.

- **По строкам.** Выберите связывание по строкам, чтобы использовать вызов `SQLBulkOperations` для загрузки данных в базу данных. Связывание по строкам обычно повышает скорость по сравнению с использованием параметризованных вставок, где для вставки данных используется принцип запись за записью.
- **По столбцам.** Выберите эту опцию, чтобы использовать для загрузки данных в базу данных связывание по столбцам. Связывание по столбцам повышает производительность, благодаря связыванию каждого столбца базы данных (в параметризованном операторе `INSERT`) с массивом из *N* значений. Одно выполнение оператора `INSERT` приводит к вставке в базу данных *N* строк. Этот метод может резко повысить производительность.

Опции внешнего загрузчика. Когда задана опция **Через внешний загрузчик**, выводится ряд опций для экспорта набора данных в файл, а также для и задания и выполнения пользовательской программы загрузки для загрузки данных из указанного файла в базу данных. IBM SPSS Modeler может взаимодействовать с внешними загрузчиками для множества широко используемых систем баз данных. В программный продукт были включены несколько сценариев; они доступны вместе с технической документацией в подкаталоге *scripts*. Имейте в виду, что для использования этой функциональной возможности должен быть установлен Python 2.7 на том же компьютере, что и IBM SPSS Modeler или IBM SPSS Modeler Server, и в файле *options.cfg* должен быть задан параметр `python_exe_path`. Дополнительную информацию смотрите в разделе “Написание утилиты массовой загрузки” на стр. 336.

- **Использовать разделитель.** Указывает, какой символ-разделитель доложен использоваться в экспортированном файле. Выберите **Табулятор**, чтобы разделять данные символом табуляции, или **Пробел** для разделения данных пробелами. Выберите **Другой**, чтобы указать другой символ, например, запятую (,).
- **Задать файл данных.** Выберите эту опцию для ввода пути, который будет использоваться для файла данных, записываемого во время массовой загрузки. По умолчанию создается временный файл в каталоге temp на сервере.
- **Задать программу загрузчика.** Выберите эту опцию, чтобы задать программу массовой загрузки. По умолчанию программный продукт ищет в подкаталоге *scripts* установки IBM SPSS Modeler сценарий Python, выполняемый для данной базы данных. В программный продукт было включено несколько сценариев; они доступны наряду с технической документацией в подкаталоге *scripts*.
- **Сгенерировать журнал.** Включите эту опцию, чтобы сгенерировать файл журнала в заданном каталоге. Этот файл журнала будет содержать информацию об ошибках и окажется полезен, если операция массовой загрузки завершится неудачно.
- **Проверить размер таблицы.** Включите эту опцию для выполнения проверки таблицы, позволяющей убедиться, что рост размера таблицы соответствует числу строк, экспортированных из IBM SPSS Modeler.
- **Добавочные опции загрузчика.** Задает дополнительные аргументы к программе загрузчика. Для аргументов, содержащих пробелы, используйте двойные кавычки.

Двойные кавычки включаются в необязательные аргументы посредством добавления к ним в качестве эскейп-символа обратной дробной черты. Например, опция, заданная как `-comment "Это \"комментарий\""`, содержит флаг комментария `-comment` и сам комментарий, вывод которого будет обработан как `Это "комментарий"`.

Одиночную обратную дробную черту можно включить, добавив к ней в качестве эскейп-символа еще одну обратную дробную черту. Например, опция, заданная как `-specialdir "C:\\Сценарии тестирования\\"`, содержит флаг `-specialdir` и каталог, вывод которого будет обработан как `C:\Сценарии тестирования\`.

Написание утилиты массовой загрузки

Узел экспорта баз данных содержит опции для массовой загрузки в диалоговом окне Дополнительные опции. Программы массовой загрузки могут использоваться для загрузки данных из текстового файла в базу данных.

Опция **Использовать массовую загрузку - через внешний загрузчик** конфигурирует для IBM SPSS Modeler три действия:

- Создание всех необходимых таблиц базы данных.
- Экспорт данных в текстовый файл.
- Вызов программы массовой загрузки для загрузки данных из этого файла в таблицу базы данных.

Как правило, программа массовой загрузки - это не сама утилита загрузки баз данных (такая как утилита sqlldr Oracle), а небольшой сценарий или программа, составляющая правильные аргументы, создающая вспомогательные файлы для конкретных баз данных (например, файлы управления), а затем вызывающая утилиту загрузки базы данных. Информация в следующих разделах поможет отредактировать существующую утилиту массовой загрузки.

Вы можете также написать свою собственную программу массовой загрузки. Дополнительную информацию смотрите в разделе “Разработка программ массовой загрузки” на стр. 340. Обратите внимание на то, что эта возможность не входит в стандартное соглашение о технической поддержке, и вам нужно обратиться за помощью к представителю группы обслуживания IBM.

Сценарии для массовой загрузки

IBM SPSS Modeler поставляется с рядом программ массовой загрузки для различных баз данных, реализуемых при помощи сценариев Python. Если выполняется поток, содержащий узел экспорта баз данных с включенной опцией **Через внешний загрузчик**, IBM SPSS Modeler создает таблицу базы данных (если она требуется) через ODBC, экспортирует данные во временный файл временный файл на хосте, использующем IBM SPSS Modeler Server,после чего вызывает сценарий массовой загрузки. Этот сценарий, в свою очередь, выполняет утилиты, предоставляемые поставщиком СУБД, для выгрузки данных из временных файлов в базу данных.

Примечание: В установку IBM SPSS Modeler интерпретатор времени выполнения Python не входит, поэтому требуется отдельная установка Python. Дополнительную информацию смотрите в разделе “Дополнительные опции экспорта базы данных” на стр. 334.

Сценарии (в папке *\scripts* каталога установки IBM SPSS Modeler) предоставляются для базы данных, перечисленных в следующей таблице.

Таблица 42. Предоставляемые сценарии массовой загрузки.

База данных	**Имя сценария**	**Дополнительная информация**
IBM DB2	*db2_loader.py*	Дополнительную информацию смотрите в разделе “Массовая загрузка данных в базы данных IBM DB2” на стр. 337.
IBM Netezza	*netezza_loader.py*	Дополнительную информацию смотрите в разделе “Массовая загрузка данных в базы данных IBM Netezza” на стр. 337.
Oracle	*oracle_loader.py*	Дополнительную информацию смотрите в разделе “Массовая загрузка данных в базы данных Oracle” на стр. 338.
SQL Server	*mssql_loader.py*	Дополнительную информацию смотрите в разделе “Массовая загрузка данных в базы данных SQL Server” на стр. 339.
Teradata	*teradata_loader.py*	Дополнительную информацию смотрите в разделе “Массовая загрузка данных в базы данных Teradata” на стр. 339.

Массовая загрузка данных в базы данных IBM DB2

Следующие замечания могут помочь сконфигурировать массовую загрузку данных из IBM SPSS Modeler в базу данных IBM DB2 при помощи опции Внешний загрузчик в диалоговом окне Дополнительные опции экспорта баз данных.

Убедитесь, что установлена утилита процессора командной строки (CLP) DB2.

Сценарий *db2_loader.py* вызывает команду DB2 LOAD. Убедитесь, что процессор командной строки DB2 (*db2* в UNIX, *db2cmd* в Windows) установлен на сервере, где выполняется *db2_loader.py* (обычно это хост, использующий IBM SPSS Modeler Server).

Проверьте, совпадает ли алиас локальной базы данных с фактическим именем базы данных.

Алиас локальной базы данных DB2 - это имя, используемое клиентской программой DB2 для ссылки на базу данных в локальном или удаленном экземпляре DB2. Если алиас локальной базы данных отличается от имени удаленной базы данных, задайте добавочную опцию загрузчика:

```
-alias <алиас_локальной_базы_данных>
```

Например, удаленная база данных называется STARS на хосте GALAXY, а алиас локальной базы данных на хосте, использующем IBM SPSS Modeler Server - STARS_GALAXY. Используйте добавочную опцию загрузчика.

```
-alias STARS_GALAXY
```

Кодировка данных не символами ASCII

В случае массовой загрузки данных, формат которых отличен от ASCII, следует убедиться, что кодовая страница, доступная в разделе конфигурации файла *db2_loader.py*, правильно установлена в системе.

Пробельные строки

Пробельные строки экспортируются в базу данных как пустые значения (NULL).

Массовая загрузка данных в базы данных IBM Netezza

Следующие замечания могут помочь сконфигурировать массовую загрузку данных из IBM SPSS Modeler в базу данных IBM Netezza при помощи опции Внешний загрузчик в диалоговом окне Дополнительные опции экспорта баз данных.

Убедитесь, что установлена утилита Netezza nzload

Сценарий *netezza_loader.py* вызывает утилиту Netezza *nzload*. Убедитесь в правильности установки и конфигурации *nzload* на сервере, где должна выполняться утилита *netezza_loader.py*.

Экспорт данных формата, иного чем ASCII

Если при экспорте присутствуют данные не в формате ASCII, может потребоваться добавить `-encoding UTF8` в поле **Добавочные опции загрузки** в диалоговом окне Дополнительные опции экспорта баз данных. Это должно гарантировать правильность выгрузки данных формата, иного чем ASCII.

Данные форматов Date, Time и Timestamp

В свойствах потока задайте формат дат **ДД-ММ-ГГГГ** и формат времени **ЧЧ:ММ:СС**.

Пробельные строки

Пробельные строки экспортируются в базу данных как пустые значения (NULL).

Разный порядок столбцов в потоке и таблице назначения при вставке данных в существующую таблицу

Если порядок столбцов в потоке отличается от порядка в таблице назначения, значения данных будут вставлены в неправильные столбцы. При помощи узла переупорядочения полей убедитесь, что порядок столбцов в потоке совпадает с их порядком в таблице назначения. Дополнительную информацию смотрите в разделе “Узел переупорядочения полей” на стр. 173.

Слежение за ходом выполнения nzload

Если IBM SPSS Modeler запускается в локальном режиме, добавьте `-sts` в поле **Добавочные опции загрузки** в диалоговом окне Дополнительные опции экспорта баз данных, чтобы в командном окне, открываемом утилитой *nzload*, выводились сообщения о состоянии через каждые 10000 строк.

Массовая загрузка данных в базы данных Oracle

Следующие замечания могут помочь сконфигурировать массовую загрузку данных из IBM SPSS Modeler в базу данных Oracle при помощи опции Внешний загрузчик в диалоговом окне Дополнительные опции экспорта баз данных.

Убедитесь, что установлена утилита Oracle sqlldr

Сценарий *oracle_loader.py* вызывает утилиту Oracle *sqlldr*. Учтите, что *sqlldr* не включается в клиент Oracle автоматически. Убедитесь, что *sqlldr* установлена на сервере, где должна выполняться утилита *oracle_loader.py*.

Задайте SID или имя службы базы данных.

Если данные экспортируются на нелокальный сервер Oracle или если у сервера Oracle несколько баз данных, в поле **Добавочные опции загрузки** диалогового окна Дополнительные опции экспорта баз данных может потребоваться задать следующую строку, чтобы передать SID или имя службы:

```
-database <SID>
```

Редактирование раздела конфигурации в oracle_loader.py

В системах UNIX (и необязательно в Windows) отредактируйте раздел конфигурации в начале сценария *oracle_loader.py*. Здесь можно задать значения для переменных среды ORACLE_SID, NLS_LANG, TNS_ADMIN и ORACLE_HOME (если это уместно), а также полный путь утилиты *sqlldr*.

Данные форматов Date, Time и Timestamp

В свойствах потока обычно требуется задать формат дат **ГГГГ-ММ-ДД** и формат времени **ЧЧ:ММ:СС**.

Если потребуется использовать форматы дат и времени, отличающиеся от описанных выше, посмотрите документацию Oracle и отредактируйте файл сценария *oracle_loader.py*.

Кодировка данных не символами ASCII

В случае массовой загрузки данных, формат которых отличен от ASCII, следует убедиться, что в системе правильно задана переменная среды NLS_LANG. Ее читает читает утилита загрузки Oracle *sqlldr*. Например, правильное значение NLS_LANG для Shift-JIS в Windows: Japanese_Japan.JA16SJIS. Дополнительные подробности о NLS_LANG смотрите в документации Oracle.

Пробельные строки

Пробельные строки экспортируются в базу данных как пустые значения (NULL).

Массовая загрузка данных в базы данных SQL Server

Следующие замечания могут помочь сконфигурировать массовую загрузку данных из IBM SPSS Modeler в базу данных SQL Server при помощи опции Внешний загрузчик в диалоговом окне Дополнительные опции экспорта баз данных.

Убедитесь, что установлена утилита SQL Server bcp.exe

Сценарий *mssql_loader.py* вызывает утилиту SQL Server *bcp.exe*. Убедитесь, что *bcp.exe* установлена на сервере, где должна выполняться утилита *mssql_loader.py*.

Пробелы, используемые в качестве разделителя, не работают.

Избегайте выбора пробела в качестве разделителя в окне Дополнительные опции экспорта баз данных.

Рекомендуется опция Проверить размер таблицы

Мы рекомендуем включить опцию **Проверить размер таблицы** в диалоговом окне Дополнительные опции экспорта баз данных. Отказы в процессе массовой загрузки определяются не всегда, и эта опция, если она включена, выполняет дополнительную проверку правильности загруженного числа строк.

Пробельные строки

Пробельные строки экспортируются в базу данных как пустые значения (NULL).

Укажите полный путь к именованному экземпляру SQL Server

Может так случиться, что SPSS Modeler не сможет получить доступ к SQL Server из-за неполного определения имени хоста, и появится следующее сообщение об ошибке:

```
Обнаружена ошибка при выполнении утилиты массовой загрузки. В файле журнала
могут содержаться дополнительные подробности.
```

Чтобы исправить эту ошибку, добавьте следующую строку в двойных кавычках в поле **Дополнительные опции загрузчика**:

```
"-S mhreboot.spss.com\SQLEXPRESS"
```

Массовая загрузка данных в базы данных Teradata

Следующие замечания могут помочь сконфигурировать массовую загрузку данных из IBM SPSS Modeler в базу данных Teradata при помощи опции Внешний загрузчик в диалоговом окне Дополнительные опции экспорта баз данных.

Убедитесь, что установлена утилита Teradata fastload

Сценарий *teradata_loader.py* вызывает утилиту Teradata *fastload*. Убедитесь в правильности установки и конфигурации *fastload* на сервере, где должна запускаться утилита *teradata_loader.py*.

Массовая загрузка данных возможна только в пустые таблицы

В качестве назначений для массовой загрузки можно использовать только пустые таблицы. Если таблица назначения содержит перед массовой загрузкой какие-либо данные, операция завершится неудачно.

Данные форматов Date, Time и Timestamp

В свойствах потока задайте формат дат **ГГГГ-ММ-ДД** и формат времени **ЧЧ:ММ:СС**.

Пробельные строки

Пробельные строки экспортируются в базу данных как пустые значения (NULL).

ID процесса Teradata ID (tdpid)

По умолчанию *fastload* экспортирует данные в систему Teradata с `tdpid=dbc`. В большинстве случаев в файле HOSTS, связывающем `dbccop1` с IP-адресом сервера Teradata, будет существовать соответствующая запись. Для использования другого сервера задайте в поле **Добавочные опции загрузки** диалогового окна Дополнительные опции экспорта баз данных следующую строку, чтобы передать `tdpid` этого сервера:

```
-tdpid <id>
```

Пробелы в именах таблиц и столбцов

Если имена таблиц или столбцов содержат пробелы, операция массовой загрузки завершится неудачно. Если возможно, переименуйте имена таблиц и столбцов, чтобы удалить пробелы.

Разработка программ массовой загрузки

В этой теме объясняется, как разработать программу массовой загрузки, которую можно запускать с IBM SPSS Modeler для загрузки данных из текстового файла в базу данных. Обратите внимание на то, что эта возможность не входит в стандартное соглашение о технической поддержке, и вам нужно обратиться за помощью к представителю группы обслуживания IBM.

Использование Python для построения программ массовой загрузки

По умолчанию IBM SPSS Modeler выполняет поиск программы массовой загрузки по умолчанию на основе типа базы данных. Смотрите Табл. 42 на стр. 336.

Можно использовать сценарий*test_loader.py*, помогающий разрабатывать программы пакетной загрузки. Дополнительную информацию смотрите в разделе “Тестирование программ массовой загрузки” на стр. 342.

Объекты, передаваемые в программу массовой загрузки

IBM SPSS Modeler записывает два файла, передаваемые в программу массовой загрузки.

- **Файл данных.** Этот файл содержит данные для загрузки в текстовом формате.
- **Файл схемы.** Это файл XML, описывающий имена и типы столбцов и предоставляющий информацию о форматировании файла данных (например, какой символ используется в качестве разделителя полей).

Кроме того, IBM SPSS Modeler передает другую информацию, такую как имя таблицы, имя пользователя и пароль, в качестве аргументов при вызове программы массовой загрузки.

Примечание: Чтобы сообщить IBM SPSS Modeler об успешном завершении, программа массовой загрузки должна удалить файл схемы.

Аргументы, передаваемые в программу массовой загрузки

Аргументы, передаваемые в программу массовой загрузки, приведены в следующей таблице.

Таблица 43. Аргументы, передаваемые в программу массовой загрузки.

Аргумент	Описание
`schemafile`	Путь файла схемы.
`data file`	Путь файла данных.
`servername`	Имя сервера СУБД; может быть пустым.
`databasename`	Имя базы данных на сервере СУБД; может быть пустым.
`username`	Имя пользователя для входа в систему базы данных.
`password`	Пароль для входа в систему базы данных.
`tablename`	Имя таблицы для загрузки.
`ownername`	Имя владельца таблицы (другое название - имя схемы).

Таблица 43. Аргументы, передаваемые в программу массовой загрузки (продолжение).

Аргумент	Описание
logfilename	Имя файла журнала (если пусто, файл журнала не создается).
rowcount	Число строк в наборе данных.

Все опции, задаваемые в поле **Добавочные опции загрузчика** диалогового окна Дополнительные опции экспорта баз данных, передаются в программу массовой загрузки после этих стандартных аргументов.

Формат файла данных

Данные записываются в файл данных в текстовом формате с полями, разделяемыми друг от друга символом-разделителем, задаваемым в диалоговом окне Дополнительные опции экспорта баз данных. Вот пример возможного вывода файла данных со знаком табулятора в качестве разделителя.

```
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
```

Этот файл записывается в локальной кодировке, используемой IBM SPSS Modeler Server (илиIBM SPSS Modeler, если нет соединения с IBM SPSS Modeler Server). Управление форматированием в некоторой степени осуществляется при помощи параметров IBM SPSS Modeler.

Формат файла схемы

Файл схемы - это файл XML, описывающий файл данных. Вот пример файла, сопровождающего предыдущий файл данных:

```
<?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>
```

В следующих двух таблицах описаны атрибуты элементов файла схемы <table> и <column>.

Таблица 44. Атрибуты элемента <table>.

Атрибут	Описание
delimiter	Символ-разделитель полей (знак табулятора представлен \t).
commit_every	Интервал размеров пакета (как в диалоговом окне Дополнительные опции экспорта баз данных).
date_format	Формат, используемый для представления дат.
time_format	Формат, используемый для представления времени.
append_existing	true - если загружаемая таблица уже содержит данные; в противном случае - false.
delete_datafile	true - если программа массовой загрузки должна удалить файл данных по завершении загрузки.

Таблица 45. Атрибуты элемента <column>.

Атрибут	Описание
`name`	Имя столбца.
`encoded_name`	Имя столбца, преобразованное в ту же кодировку, что и файл данных, и представленное в выводе группой двузначных шестнадцатеричных чисел.
`type`	Типов данных столбца, представленный одним из следующих: `integer`, `real`, `char`, `time`, `date` или `datetime`.
`size`	Для типа данных `char`; максимальная ширина столбца в символах.

Тестирование программ массовой загрузки

Массовую загрузку можно проверить при помощи сценария тестирования *test_loader.py*, содержащегося в папке *\scripts* каталога установки IBM SPSS Modeler. Такая проверка полезна при разработке, отладке или устранении неисправностей программ или сценариев массовой загрузки для использования с IBM SPSS Modeler.

Для применения этого сценария тестирования выполните следующие действия:

1. Запустите сценарий *test_loader.py*, чтобы скопировать файлы схемы и данных в файлы *schema.xml* и *data.txt*, и создайте пакетный файл Windows (*test.bat*).
2. Отредактируйте файл *test.bat*, чтобы выбрать программу или сценарий массовой загрузки для тестирования.
3. Вызовите *test.bat* из командной оболочки, чтобы проверить выбранную программу или сценарий массовой загрузки.

Примечание: При выполнении *test.bat* данные не загружаются в базу данных фактически.

Узел экспорта плоских файлов

Узел экспорта плоских файлов позволяет записывать данные в текстовый файл с разделителями. Это полезно для экспорта данных, которые могут читать друге программы анализа и электронных таблиц.

Если ваши данные содержат геопространственную информацию, ее можно экспортировать как плоский файл, а если для использования в том же потоке сгенерирован узел источника Файл переменных, в новом узле источника сохранятся и все метаданные системы хранения, измерений и геопространственных данных. Однако в случае экспорта данных и последующего импорта их в другой поток необходимо предпринять некоторые дополнительные действия, чтобы задать геопространственные метаданные на новом узле источника. Более подробную информацию смотрите в теме “Узел файла переменных” на стр. 26.

Примечание: Запись файлов в старом формате кэширования невозможна, поскольку IBM SPSS Modeler больше не использует этот формат для кэширования файлов. Файлы кэша IBM SPSS Modeler теперь сохраняются в формате IBM SPSS Statistics *.sav*, запись в котором возможна при помощи узла экспорта Statistics. Более подробную информацию смотрите в теме “Узел Statistics Export” на стр. 360.

Вкладка Экспорт плоских файлов

Экспортировать файл. Задает имя файла. Введите имя файла или нажмите кнопку средства выбора файлов, чтобы найти положение нужного файла.

Режим записи. Если выбрать опцию **Перезаписывать**, все существующие данные в заданном файле будут перезаписываться. Если выбрать опцию **Добавлять**, вывод будет добавляться в конец существующего файла, с сохранением всех содержащихся в файле данных.

- **Включить имена полей.** Если эта опция включена, имена полей будут записываться в первую строку выходного файла. Эта опция доступна только для режима записи **Перезаписывать**.

Символ новой строки после каждой записи. Если эта опция включена, каждая запись будет вводиться в выходном файле в новой строке.

Разделитель полей. Задает символ для вставки между значениями полей в генерируемом текстовом файле. Есть опции **Запятая**, **Табулятор**, **Пробел** и **Другой**. При выборе опции **Другой** введите один или несколько нужных вам символов-разделителей в текстовом поле.

Кавычки для элементов. Задает тип кавычек, используемый для значений символических полей. Есть опции **Нет** (не заключать значения в кавычки), **Одинарные (')**, **Двойные (")** и **Другой**. При выборе опции **Другой** введите нужные вам символы кавычек в текстовом поле.

Кодировка. Задает используемый метод кодирования текста. Можно выбрать по умолчанию для системы, по умолчанию для потока или UTF-8.

- По умолчанию для системы задается на панели управления Windows или (при работе в распределенном режиме) на компьютере сервера.
- По умолчанию для потока задается в диалоговом окне Свойства потока.

Десятичный разделитель. Задает способ представления десятичных чисел в данных.

- **Как в потоке по умолчанию.** Будет использоваться десятичный разделитель, определяемый значением параметра по молчанию текущего потока. Обычно используется десятичный разделитель, определяемый локальными параметрами компьютера.
- **Точка (.).** В качестве разделителя десятичной части будет использоваться символ точки.
- **Запятая (,).** В качестве разделителя десятичной части будет использоваться символ запятой.

Сгенерировать узел импорта для этих данных. Выберите эту опцию для автоматического генерирования узла источника файлов переменных, который будет читать экспортированный файл данных. Дополнительную информацию смотрите в разделе “Узел файла переменных” на стр. 26.

Узел экспорта Data Collection

Узел экспорта Data Collection сохраняет данные в формате, используемом программой маркетинговых исследований Data Collection на основе Data Collection Data Model. Этот формат позволяет отличить данные наблюдений (фактические ответы на вопросы, собираемые во время опроса) от метаданных, описывающих способы сбора и организации данных наблюдений. В метаданные включается информация, такая как текст вопросов, имена и описания переменных, наборы множественных ответов, варианты перевода различных текстов и определения структуры данных наблюдений. Дополнительную информацию смотрите в разделе “Узел Data Collection” на стр. 32.

Файл метаданных. Задает имя файла определений вопросов (*.mdd*), где будут сохраняться экспортированные метаданные. Опросный лист создается на основе информации о типах полей. Например, номинальное поле (набора) можно представить как один вопрос с описанием поля, используемым в качестве текста вопроса, и отдельным переключателем для каждого определенного значения.

Объединить метаданные. Указывает, будут ли метаданные перезаписывать существующие версии или сливаться с существующими метаданными. Если выбрана опция слияния, при каждом запуске потока будет создаваться новая версия. Это позволяет отслеживать версии опросного листа по мере внесения в него изменений. Каждая версия может считаться снимком метаданных, используемых для сбора конкретного набора данных наблюдений.

Включить поддержку системных переменных. Задает, будут ли системные переменные включены в экспортированный файл *.mdd*. В их состав входят такие переменные, как *Respondent.Serial*, *Respondent.Origin* и *DataCollection.StartTime*.

Параметры данных наблюдений. Задает файл данных IBM SPSS Statistics (*.sav*) для экспорта данных наблюдений. Имейте в виду, что здесь применяются все ограничения на имена переменных и значений,

поэтому может потребоваться, например, перейти на вкладку Фильтр и применить в меню опций фильтра опцию "Переименовать для IBM SPSS Statistics", чтобы исправить в именах полей недопустимые символы.

Сгенерировать узел импорта для этих данных. Выберите эту опцию для автоматического генерирования узла источника Data Collection, который будет читать экспортированный файл данных.

Наборы множественных ответов. Все наборы множественных ответов, определяемые в потоке, будут автоматически сохраняться при экспорте файла. Наборы множественных ответов можно просмотреть и отредактировать с любого узла с вкладкой Фильтр. Дополнительную информацию смотрите в разделе “Редактирование наборов множественных ответов” на стр. 143.

Узел экспорта Analytic Server

Узел Экспорт Analytic Server позволяет записывать данные вашего анализа в существующий источник данных Analytic Server. Например, это могут быть текстовые файлы в распределенной файловой системе Hadoop (Hadoop Distributed File System, HDFS) или база данных.

Обычно поток с узлом экспорта Analytic Server также начинается с узлов источника Analytic Server, подставляется в Analytic Server и исполняется в HDFS. Другой вариант - поток с "локальными" источниками данных может заканчиваться узлом экспорта Analytic Server, чтобы закачать относительно небольшие наборы данных (не больше 100 тысяч записей) для использования с Analytic Server.

Источник данных. Выберите источник данных, содержащий данные, которые вы хотите использовать. Источник данных содержит файлы и метаданные, связанные с этим источником. Нажмите кнопку **Выбрать**, чтобы вывести список доступных источников данных. Дополнительную информацию смотрите в разделе “Выбор источника данных” на стр. 13.

Если нужно создать новый источник данных или изменить существующий, нажмите кнопку **Запустить редактор источников данных...**.

Режим. Выберите **Присоединить**, чтобы добавить новый источник данных, или **Перезаписать**, чтобы заменить содержимое источника данных.

Сгенерировать узел импорта для этих данных . Выберите эту опцию, чтобы сгенерировать узел источника для данных, экспортируемых в заданный источник данных. Этот узел будет добавлен на холст потока.

Узел экспорта IBM Cognos BI

Узел экспорта IBM Cognos BI позволяет экспортировать данные из потока IBM SPSS Modeler в Cognos BI в формате UTF-8. Таким образом, Cognos BI может использовать преобразованные или оцененные данные IBM SPSS Modeler. Например, при помощи Cognos BI Report Studio можно создать отчет на основе экспортированных данных, включая предсказания и значения доверительной вероятности. Затем этот отчет можно сохранить на сервере Cognos BI и распространить среди пользователей Cognos BI.

Примечание: Можно экспортировать только реляционные данные, но не данные OLAP.

Для экспорта данных в Cognos BI нужно задать следующие соединения:

- Подключение Cognos - соединение с сервером Cognos BI
- Подключение ODBC - соединение с сервером данных Cognos, используемым сервером Cognos BI

В подключении Cognos задается источник данных Cognos, который следует использовать. Этот источник данных должен использовать то же имя для входа в систему, что и источник данных ODBC.

Фактические данные потока экспортируются на сервер данных, а метаданные пакета - на сервер Cognos BI.

Как и на любом другом узле экспорта, здесь также можно при помощи вкладки Опубликовать диалогового окне узла опубликовать поток для внедрения при помощи IBM SPSS Modeler Solution Publisher.

Подключение Cognos

Здесь можно задать соединение с сервером Cognos BI, который вы хотите использовать для операции экспорта. Процедура предполагает экспорт метаданных в новый пакет на сервере Cognos BI, тогда как как данные потока экспортируются на сервер данных Cognos.

Подключение. Нажмите кнопку **Правка**, чтобы открыть диалоговое окно, в котором можно задать URL и другие подробности о сервере Cognos BI, на который вы хотите экспортировать данные. Если вы уже вошли в систему на сервер Cognos BI через IBM SPSS Modeler, то сможете также отредактировать и подробности текущего соединения. Дополнительную информацию смотрите в разделе “Подключения Cognos” на стр. 39.

Источник данных. Имя источника данных Cognos (обычно это база данных), куда будут экспортироваться данные. В выпадающем списке выводятся все источники данных Cognos, к которым можно обратиться из текущего сеанса соединения. Нажмите кнопку **Обновить**, чтобы обновить этот список.

Папка. Путь и имя папки сервера Cognos BI, где должен быть создан пакет экспорта.

Название пакета . Имя пакета, в заданной папке, где будут содержаться экспортированные данные. Это может быть новый пакет с одной темой запросов; экспорт в существующий пакет невозможен.

Режим. Задает предпочитаемый вами способ экспорта.

- **Опубликовать пакет сейчас.** (По умолчанию) Выполняет операцию экспорта сразу же после нажатия кнопки **Выполнить**.
- **Сценарий экспорта.** Создает сценарий XML, который можно запустить позднее (например, при помощи Framework Manager), чтобы выполнить операцию экспорта. Введите путь и имя файла для сценария в поле **Файл** либо задайте имя и положение файла сценария при помощи кнопки **Правка**.

Сгенерировать узел импорта для этих данных. Выберите эту опцию, чтобы сгенерировать узел источника для данных, экспортируемых в заданный источник данных и таблицу. При нажатии кнопки **Выполнить** этот узел будет добавлен на холст потока.

Подключение ODBC

Здесь задается соединение с сервером данных Cognos (то есть, с базой данных), куда будут экспортироваться данные потока.

Примечание: Обязательно убедитесь, что задаваемый здесь источник данных указывает на источник, заданный на панели **Подключения Cognos**. Нужно также убедиться, что источник данных соединения с Cognos использует то же имя входа в систему, что и источник данных ODBC.

Источник данных. Показывает выбранный источник данных. Введите имя или выберите его в выпадающем списке. Если вы не видите нужную базу данных в списке, выберите **Добавить новое соединение с базой данных** и найдите нужную базу данных в диалоговом окне Соединения с базами данных. Дополнительную информацию смотрите в разделе “Добавление соединения с базой данных” на стр. 19.

Имя таблицы. Введите имя таблицы, в которую вы хотите отправить данные. При выборе опции **Вставить в таблицу** можно будет выбрать существующую таблицу в базе данных, нажав кнопку **Выбрать**.

Создать таблицу. Выберите эту опцию, чтобы создать новую или переопределить существующую таблицу базы данных.

Вставить в таблицу. Выберите эту опцию, чтобы вставить данные в виде новых строк в существующую таблицу базы данных.

Объединить таблицу. Выберите эту опцию (при ее наличии), чтобы обновить выбранные столбцы базы данных значениями из соответствующих полей исходных данных. Выбор этой опции включает поддержку кнопки **Слить**, выводящей диалоговое окно, откуда поля исходных данных можно отобразить на столбцы базы данных.

Отбросить существующую таблицу. Выберите эту опцию, чтобы удалить любую существующую таблицу с таким же именем при создании новой таблицы.

Удалить существующие строки. Выберите эту опцию, чтобы удалить существующие строки из таблицы перед экспортом при операции вставки в таблицу.

Примечание: Если выбрана одна из двух опций выше, при вызове узла вы получите сообщение **Предупреждение о перезаписи**. Чтобы подавить предупреждения, выключите переключатель **Предупреждать, если узел перезаписывает таблицу базы данных** на вкладке Уведомления диалогового окна Пользовательские опции.

Размер строки по умолчанию. Поля, помеченные вами как поля без типа на узле типа восходящего потока, записываются в базу данных как строковые поля. Задайте размер строк, который должен использоваться для полей без типа.

Нажмите кнопку **Схема**, чтобы открыть диалоговое окно, где можно задать различные опции экспорта (для баз данных, поддерживающих эту возможность), задать для полей типы данных SQL и задать первичный ключ для целей индексации базы данных. Дополнительную информацию смотрите в разделе “Опции схемы экспорта базы данных” на стр. 330.

Нажмите кнопку **Индексы**, чтобы задать опции для индексации экспортированной таблицы для повышения производительности базы данных. Дополнительную информацию смотрите в разделе “Опции индексов экспорта базы данных” на стр. 332.

Нажмите кнопку **Дополнительно**, чтобы задать опции массовой загрузки и принятий базы данных. Дополнительную информацию смотрите в разделе “Дополнительные опции экспорта базы данных” на стр. 334.

Заключать в кавычки имена таблиц и столбцов. Выберите опции, используемые при отправке в базу данных оператора `CREATE TABLE`. Имена таблиц или столбцов с пробелами или нестандартными символами должны быть заключены в кавычки.

- **Когда требуется.** Выберите эту опцию, чтобы разрешить IBM SPSS Modeler автоматически определять, когда требуются кавычки, на индивидуальной основе.
- **Всегда.** Выберите эту опцию, чтобы всегда заключать имена таблиц и столбцов в кавычки.
- **Никогда.** Выберите эту опцию, чтобы отключить использование кавычек.

Сгенерировать узел импорта для этих данных. Выберите эту опцию, чтобы сгенерировать узел источника для данных, экспортируемых в заданный источник данных и таблицу. При нажатии кнопки **Выполнить** этот узел будет добавлен на холст потока.

Узел экспорта IBM Cognos TM1

Узел экспорта IBM Cognos BI позволяет экспортировать данные из потока IBM SPSS Modeler в Cognos TM1. Таким образом, Cognos BI может использовать преобразованные или оцененные данные IBM SPSS Modeler.

Примечание: Можно экспортировать только показатели, но не контекстные данные измерений; можно также добавить в куб новые элементы.

Для экспорта данных в Cognos BI нужно задать следующие соединения:

- Соединение с сервером Cognos TM1.

- Куб, куда будут экспортироваться данные.
- Отображение имен данных SPSS на эквивалентные показатели и измерения TM1.

Примечание: У пользователя TM1 должны быть следующие разрешения: привилегия записи для кубов, привилегия чтения для измерений и привилегия записи для элементов измерений.

Как и на любом другом узле экспорта, здесь также можно при помощи вкладки Опубликовать диалогового окне узла опубликовать поток для внедрения при помощи IBM SPSS Modeler Solution Publisher.

Примечание: Прежде чем использовать узлы источник TM1 или узлы экспорта в SPSS Modeler, необходимо верифицировать некоторые параметры в файле `tm1s.cfg`, то есть в файле конфигурации сервера TM1 в корневом каталоге сервера TM1.

- HTTPPortNumber - задать допустимый номер порта; обычно 1-65535.
- UseSSL - если задать для этого параметра значение *True*, в качестве транспортного протокола будет использоваться HTTPS. В этом случае необходимо импортировать сертификацию TM1 в SPSS Modeler Server JRE.

Соединение с кубом IBM Cognos TM1 для экспорта данных

Первый шаг экспорта данных в базу данных IBM Cognos TM1 - выбор соответствующего хоста администрирования TM1, связанного сервера и куба на вкладке **Соединение** диалогового окна IBM Cognos TM1.

Примечание: При экспорте в TM1 будут отбрасываться только фактические значения "null". Нулевые значения (0) будут экспортироваться как допустимые. Обратите внимание также на то, что только поля с типом хранения *строка* могут отображаться на измерения на вкладке Отображение. Перед экспортом в TM1 необходимо использовать клиент IBM SPSS Modeler для преобразования нестроковых типов данных в строку.

Хост администрирования Введите URL хоста администрирования, где установлен сервер TM1, с которым вы хотите соединиться. Хост администрирования определяется как один URL для всех серверов TM1. Из положения с этим URL можно обнаружить все серверы IBM Cognos TM1, установленные и запущенные в вашей среде, и связаться с ними.

Сервер TM1 Когда соединение с хостом администрирования будет установлено, выберите сервер, содержащий данные, которые вы хотите импортировать, и нажмите кнопку **Вход в систему**. Если прежде вы не соединялись с этим сервером, появится предложение ввести **Имя пользователя** и **Пароль**; другой вариант - найти ранее введенные сведения входа в систему, сохраненные как **Хранимые регистрационные данные**.

Выберите куб TM1 для экспорта Содержит имена кубов на сервере TM1, куда можно экспортировать данные.

Чтобы определить данные для экспорта, выберите куб и щелкните по стрелке вправо, чтобы перенести этот куб в поле **Экспортировать в куб**. Выбрав куб, используйте вкладку Отображение, чтобы отобразить измерения и меры TM1 на соответствующие поля SPSS или на фиксированное значение (операция *Выбор*).

Отображение данных IBM Cognos TM1 для экспорта

После выбора хоста администрирования TM1 и связанного сервера и куба TM1 используйте вкладку Отображение диалогового окна Экспорт IBM Cognos TM1 для отображения измерений и показателей TM1 на поля SPSS или для срезов измерений TM1 по фиксированным значениям.

Примечание: Только поля с типом хранения *строка* могут отображаться на измерения. Перед экспортом в TM1 необходимо использовать клиент IBM SPSS Modeler для преобразования нестроковых типов данных в строку.

Поля Список имен полей данных из файла данных SPSS, доступных для экспорта.

Измерения TM1 Показывает куб TM1, выбранный на вкладке Соединение вместе с его регулярными измерениями, измерением показателей и элементами выбранного измерения показателей. Выберите название измерения TM1 или показатель для отображения на поле данных SPSS.

На вкладке Отображение доступны следующие опции.

Выбор измерения показателя Из списка измерений для выбранного куба выберите одно, которое будет измерением показателя.

Если выбрать измерение, кроме измерения показателя, и нажать кнопку **Выбрать**, откроется диалоговое окно, показывающее конечные элементы выбранного измерения. Можно выбирать только конечные элементы. Выбранные элементы помечены буквой **S**.

Отобразить Отображает выбранное поле данных SPSS на выбранное измерение или показатель TM1 (регулярное измерение или конкретный показатель или элемент из измерения показателя). Отображенные поля помечаются буквой **M**.

Отмена отображения Отменяет отображение поля данных SPSS из выбранного измерения или показателя TM1. Обратите внимание на то, что одновременно можно отменить только одно отображение. Поле данных SPSS с отмененным отображением возвращается обратно в левый столбец.

Создать новый Создает новый показатель в измерении показателей TM1. Откроется диалоговое окно, в котором нужно ввести новое **Имя показателя TM1**. Эта опция доступна только для измерений показателей, но не для регулярных измерений.

Дополнительную информацию о TM1 смотрите в документации по IBM Cognos TM1 по адресу http://www-01.ibm.com/support/knowledgecenter/SS9RXT_10.2.2/com.ibm.swg.ba.cognos.ctm1.doc/welcome.html.

Узел экспорта SAS

Примечание: Эта возможность доступна в SPSS Modeler Professional и SPSS Modeler Premium.

Узел экспорта SAS позволяет записывать данные в формате SAS для их считывания в пакет SAS или совместимый с SAS программный пакет. Экспорт возможен в трех форматах файлов SAS: SAS for Windows/OS2, SAS for UNIX и SAS Версии 7/8.

Вкладка Экспорт узла экспорта SAS

Экспортировать файл. Задайте имя файла. Введите имя файла или нажмите кнопку средства выбора файлов, чтобы найти положение нужного файла.

Экспорт. Задайте формат файла экспорта. Есть опции: **SAS for Windows/OS2**, **SAS for UNIX** и**SAS Версии 7/8**.

Экспортировать имена полей. Выберите опции для экспорта имен и меток полей из IBM SPSS Modeler для использования с SAS.

- **Имена и метки переменных.** Выберите эту опцию для экспорта и имен, и меток полей IBM SPSS Modeler. Имена полей будут экспортироваться как имена переменных SAS, а метки - как метки переменных SAS.
- **Имена как метки переменных.** Выберите эту опцию для использования имен полей IBM SPSS Modeler в SAS в качестве меток переменных. IBM SPSS Modeler разрешает в именах полей символы, недопустимые в именах переменных SAS. Для предотвращения возможного создания недопустимых имен SAS выберите вместо этой опцию **Имена и метки переменных**.

Сгенерировать узел импорта для этих данных. Выберите эту опцию для автоматического генерирования узла источника SAS, который будет читать экспортированный файл данных. Дополнительную информацию смотрите в разделе “Узел источника SAS” на стр. 41.

Узел экспорта Excel

Узел экспорта Excel выводит данные в формате Microsoft Excel .xlsx. Другой вариант - можно выбрать автоматический запуск Excel и открытие экспортированного файла при вызове узла.

Вкладка Экспорт узла Excel

Имя файла. Введите имя файла или нажмите кнопку средства выбора файлов, чтобы найти положение нужного файла. Имя файла по умолчанию - *excelexp.xlsx*.

Тип файла. Поддерживается формат файлов Excel .xlsx.

Создать файл. Создает новый файл Excel.

Вставить в существующий файл. Содержимое заменяется, начиная с ячейки, обозначаемой полем **Запустить в ячейке**. Другие ячейки в электронной таблице остаются со своим исходным содержимым.

Включить имена полей. Укажите, следует ли включить имена полей в первую строку рабочей таблицы.

Запустить в ячейке. Положение ячейки, используемое для первой записи экспорта (или первого имени поля, если включен переключатель **Включать имена полей**). Заполнение данными будет происходить вправо и вниз от этой начальной ячейки.

Выбрать рабочую таблицу. Задает рабочую таблицу, в которую вы хотите экспортировать данные. Идентификация рабочей таблиц возможна по индексу или имени.

- **По индексам.** Если вы создаете новый файл, задайте число от 0 до 9, идентифицирующее рабочую таблицу, в которую вы хотите экспортировать данные, начиная с 0 для первой рабочей таблицы, 1 для второй рабочей таблицы и так далее. Значения 10 и выше можно использовать, только если рабочая таблица уже существует в этой позиции
- **По именам.** Если вы создаете новый файл, задайте имя, используемое для рабочей таблицы. Если выполняется операция вставки в существующий файл, данные будут вставляться в рабочую таблицу, если она существует, в противном случае будет создана новая рабочая таблица с этим именем.

Запустить Excel. Задает, будет ли Excel автоматически запускаться для экспортируемого файла при вызове узла. Имейте в виду, что при работе в распределенном режиме на IBM SPSS Modeler Server, вывод сохраняется в файловую систему сервера, а Excel запускается на клиенте Client с копией экспортируемого файла.

Сгенерировать узел импорта для этих данных. Выберите эту опцию для автоматического генерирования узла источника Excel, который будет читать экспортированный файл данных. Дополнительную информацию смотрите в разделе “Узел источника Excel” на стр. 42.

Узел экспорта XML

Узел экспорта XML позволяет выводить данные в формате XML, используя кодировку UTF-8. Дополнительно можно создать узел источника XML для обратного чтения данных экспорта в поток.

Файл экспорта XML . Полный путь и имя файла XML, в который вы хотите экспортировать данные.

Использовать схему XML. Включите этот переключатель, если вы хотите использовать схему или DTD для управления структурой экспортируемых данных. При этом активируется кнопка **Отобразить**, описанная ниже.

Если вы не используете схему или DTD, по умолчанию используется следующая структура для экспортируемых данных:

```
<records>
     <record>
          <имя_поля1>значение</имя_поля1>
          <имя_поля2>значение</имя_поля2>
      :
          <имя_поляN>значение</имя_поляN>
     </record>
     <record>
      :
      :
     </record>
      :
      :
</records>
```

Пробелы в имени поля заменяются на нижние подчеркивания; например, "Мое поле" приобретает вид <Мое_поле>.

Отобразить. Если вы выбрали использование схемы XML, нажатие этой кнопки открывает диалоговое окно, в котором можно задать, какая часть структуры XML должна использоваться как начало каждой новой записи. Дополнительную информацию смотрите в разделе “Опции записей отображения XML”.

Отображенные поля. Обозначает количество полей, которые были отображены.

Сгенерировать узел импорта для этих данных. Выберите эту опцию для автоматического генерирования узла источника XML, который будет читать экспортированный файл данных обратно в поток. Дополнительную информацию смотрите в разделе “Узел источника XML” на стр. 43.

Запись данных XML

Когда задан элемент XML, значение поля размещается в теге элемента:

```
<элемент>значение</элемент>
```

Когда отображается атрибут, значение поля размещается как значение для атрибута:

```
<элемент атрибут="значение">
```

Если поле отображается на элемент над элементом <записи>, это поле записывается только однажды и будет постоянным для всех записей. Значение для этого элемента будет взято из первой записи.

Если должно быть записано значение null, это делается заданием пустого содержимого. Для элементов это:

```
<элемент></элемент>
```

Для атрибутов это:

```
<элемент атрибут="">
```

Опции записей отображения XML

Вкладка Записи позволяет задать, какая часть структуры XML будет использоваться как начало каждой новой записи. Чтобы выполнить правильное отображение на схему, необходимо задать разделитель записи.

Структура XML. Иерархическое дерево, показывающее структуру схемы XML, заданную на предыдущем экране.

Записи (выражение XPath) . Чтобы задать разделитель записей, выберите элемент в структуре XML и нажмите кнопку с правой стрелкой. Всякий раз, когда этот элемент встречается в данных источника, в выходном файле создается новая запись.

Примечание: Если вы выбираете в структуре XML корневой элемент, может быть записана только одна запись, а все остальные записи пропускаются.

Опции полей отображения XML

Вкладка Поля используется для отображения полей в наборе данных на элементы или атрибуты в структуре XML, когда используется файл схемы.

Имена полей, которые совпадают с именем элемента или атрибута, отображаются автоматически, если имя элемента или атрибута уникально. Таким образом, если есть и элемент, и атрибут с именем `поле1`, автоматическое отображение не происходит. Если есть только один элемент в структуре с именем `поле1`, поле с этим именем в потоке отображается автоматически.

Поля. Список полей в модели. Выберите одно или несколько полей как исходную часть отображения. Вы можете использовать кнопки под списком для выбора всех полей или всех полей с конкретным уровнем измерения.

Структура XML. Выберите элемент в структуре XML как назначение отображения. Для создания отображения нажмите кнопку Отобразить. Затем отображение выводится на экран. Количество полей, отображенных таким способом, выводится внизу этого списка.

Для удаления отображения выберите элемент в списке структуры XML и нажмите кнопку **Отменить отображение**.

Вывод на экран атрибутов . Выводит или скрывает атрибуты, если такие есть, элементов XML в структуре XML.

Предварительный просмотр XML

На вкладке Предварительный просмотр нажмите кнопку **Обновить**, чтобы предварительно просмотреть XML, который будет записан.

Если отображение неправильное, вернитесь на вкладку Записи или Поля для исправления ошибок и снова нажмите кнопку **Обновить**, чтобы увидеть результаты.

Глава 8. Узлы IBM SPSS Statistics

Узлы IBM SPSS Statistics - Обзор

В дополнение к IBM SPSS Modeleri его возможностей исследования данных IBM SPSS Statistics предоставляет возможность выполнения дополнительного статистического анализа и управления данными.

Если установлена совместимая залицензированная копия IBM SPSS Statistics, с ней можно соединиться из IBM SPSS Modeler и выполнять комплексную многошаговую обработку и анализ данных, иначе не поддерживаемые IBM SPSS Modeler. Для опытных пользователей предусмотрена также опция для дальнейшего видоизменения анализа при помощи синтаксиса команд. Информацию о совместимости версий смотрите в замечаниях по выпуску.

Узлы IBM SPSS Statistics, если они доступны, выводятся в специально отведенной области палитры узлов.

Примечание: Мы рекомендуем инициировать данные на узле типа перед использованием узлов преобразования, модели или вывода IBM SPSS Statistics. Это также является требованием при использовании синтаксиса команды AUTORECODE.

палитра IBM SPSS Statistics содержит следующие узлы:

Узел Файл статистики читает данные в формате файлов *.sav*, используемом IBM SPSS Statistics, а также файлы кэша, сохраненные IBM SPSS Modeler, которые также используют этот формат.

Узел Преобразование статистики запускает разнообразные команды синтаксиса IBM SPSS Statistics для источников данных в IBM SPSS Modeler. Этому узлу требуется лицензированная копия IBM SPSS Statistics.

Узел Статистическая модель позволяет проанализировать свои данные и работать с ними, запустив процедуры IBM SPSS Statistics, создающие PMML. Этому узлу требуется лицензированная копия IBM SPSS Statistics.

Узел Вывод статистики позволяет вызвать процедуру IBM SPSS Statistics для анализа ваших данных IBM SPSS Modeler. Доступны разнообразные аналитические процедуры IBM SPSS Statistics. Этому узлу требуется лицензированная копия IBM SPSS Statistics.

Узел Экспорт статистики выводит данные в формате IBM SPSS Statistics *.sav* или *.zsav*. Файлы *.sav* или *.zsav* могут быть прочитаны модулем IBM SPSS Statistics Base и другими продуктами. Это формат, используемый также для файлов кэша в IBM SPSS Modeler.

Примечание: Если копия SPSS Statistics залицензирована только для одного пользователя, при вызове потока с несколькими ветвями, каждая из которых содержит узел SPSS Statistics, можно получить сообщение об

ошибке лицензирования. Такое происходит, если сеанс SPSS Statistics для одной ветви не будет завершен до попытки запуска сеанса для другой ветви. По возможности переделайте поток так, чтобы несколько ветвей с узлами SPSS Statistics не выполнялись параллельно.

Узел Statistics File

Узел Statistics File можно использовать для чтения данных непосредственно из файла IBM SPSS Statistics (*.sav*или *.zsav*). Этот формат теперь используется для замены файла кэша из более ранних версий IBM SPSS Modeler. Если вы хотите импортировать сохраненный файл кэша, следует использовать узел IBM SPSS Statistics File.

Файл импорта. Задайте имя файла. Можно ввести имя файла или нажать кнопку с многоточием (**...**), чтобы выбрать файл. Путь файла появится, как только вы выберете файл.

Файл зашифрован с помощью пароля. Включите этот переключатель, если известно, что файл зашифрован паролем; вам предложат ввести **Пароль**. Если файл защищен паролем, а вы не ввели пароль, появится предупреждение при попытке перейти на другую вкладку, обновить данные, предварительно просмотреть содержимое узла или выполнить содержащий данный узел поток.

Примечание: Защищенные паролем файлы можно открыть только в IBM SPSS Modeler версии 16 или новее.

Имена переменных. Выберите способ обработки имен и меток переменных при импорте из файла IBM SPSS Statistics *.sav* или *.zsav*. Метаданные, выбираемые вами здесь для включения, сохраняются в течение всей вашей работы в IBM SPSS Modeler и могут быть снова экспортированы для использования в IBM SPSS Statistics.

- **Читать имена и метки.** Выберите эту опцию для чтения имен и меток переменных в IBM SPSS Modeler. По умолчанию эта опция включена, а имена переменных вводятся на узле типа. Метки могут выводиться на диаграммах, в браузерах моделей и выводе других типов, в зависимости от опций, заданных в диалоговом окне свойств потока. По умолчанию используется вывод меток на экран, указанный в выводе.
- **Читать метки как имена.** Выберите эту опцию для чтения описательных меток переменных из файла IBM SPSS Statistics *.sav* или *.zsav* вместо кратких имен полей и для использования этих меток в качестве имен переменных в IBM SPSS Modeler.

Значения. Выберите способ обработки значений и меток при импорте из файла IBM SPSS Statistics *.sav* или *.zsav*. Метаданные, выбираемые вами здесь для включения, сохраняются в течение всей вашей работы в IBM SPSS Modeler и могут быть снова экспортированы для использования в IBM SPSS Statistics.

- **Читать данные и метки.** Выберите эту опцию для считывания фактических значений и значений меток в IBM SPSS Modeler. По умолчанию эта опция включена, а собственно значения вводятся на узле типа. Метки значений могут выводиться на диаграммах Построителя выражений, в браузерах моделей и выводе других типов, в зависимости от опций, заданных в диалоговом окне свойств потока.
- **Читать метки как данные.** Выберите при желании использовать метки значений из файла *.sav* или *.zsav* вместо числовых или символьных кодов, используемых для представления значений. Например, при выборе этой опции для данных, значения которых 1 и 2 фактически представляют соответственно *М* и *Ж*, поле будет преобразовано в строковое, а *М* и *Ж* будут импортированы как фактические значения.

  Перед выбором этой опции важно рассмотреть использование пропущенных значений в данных IBM SPSS Statistics. Например, если числовое поле использует метки только для пропущенных значений (0 = *Нет ответа*, –99 = *Неизвестно*), то при выборе опции выше будут импортированы только метки значений *Нет ответа* и *Неизвестно* и поле будет преобразовано в строковое. В таких случаях вы должны импортировать сами значения и задать пропущенные значения на узле типа.

Использовать информацию формата полей для определения хранения. Если этот переключатель не включен, значения полей, сформатированные в файле *.sav* как целочисленные (то есть полей, заданных в представлении Переменные в IBM SPSS Statistics как как F*n*.0), импортируются с использованием хранения целочисленного типа. Все остальные значения полей, кроме строковых, импортируются как действительные числа.

Если этот переключатель включен (по умолчанию), все значения полей, кроме строковых, импортируются как действительные числа (сформатированы ли они в файле *.sav* как целочисленные или нет).

Наборы множественных ответов. Все наборы множественных ответов, определяемые в файлеIBM SPSS Statistics, будут автоматически сохраняться при импорте файла. Наборы множественных ответов можно просмотреть и отредактировать с любого узла с вкладкой Фильтр. Дополнительную информацию смотрите в разделе “Редактирование наборов множественных ответов” на стр. 143.

Узел Statistics Transform

Узел Statistics Transform позволяет выполнять преобразования данных при помощи командного синтаксиса IBM SPSS Statistics. Это делает возможным выполнить ряд преобразований, не поддерживаемых IBM SPSS Modeler, и допускает автоматизацию сложных многошаговых преобразований, включая создание ряда полей с одного узла. Этот узел во всем подобен узлу Statistics Output за исключением того, что данные возвращаются в IBM SPSS Modeler для дальнейшего анализа, между тем как на узле вывода данные возвращаются в виде затребованных объектов вывода, таких как диаграммы или таблицы.

Для использования этого узла на вашем компьютере должна быть установлена и залицензирована совместимая версия IBM SPSS Statistics. Дополнительную информацию смотрите в разделе “Вспомогательные прикладные программы IBM SPSS Statistics” на стр. 325. Информацию о совместимости смотрите в замечаниях по выпуску.

При необходимости на вкладке Фильтр можно отфильтровать поля или переименовать их, чтобы они соответствовали стандартам именования IBM SPSS Statistics. Дополнительную информацию смотрите в разделе “Переименование или фильтрация полей для IBM SPSS Statistics” на стр. 361.

Руководство по синтаксису. Подробности о конкретных процедурах IBM SPSS Statistics смотрите в *Руководстве по синтаксису команд IBM SPSS Statistics*, включенному в вашу копию программного пакета IBM SPSS Statistics. Чтобы просмотреть это руководство с вкладки Синтаксис, выберите опцию **Редактор синтаксиса** и нажмите кнопку Запустить справку по синтаксису IBM SPSS Statistics.

Примечание: Этот узел поддерживает не весь синтаксис IBM SPSS Statistics. Дополнительную информацию смотрите в разделе “Разрешаемый синтаксис” на стр. 356.

Узел Statistics Transforme - Вкладка Синтаксис

Опция диалогового окна IBM SPSS Statistics

Если синтаксис IBM SPSS Statistics для процедуры вам незнаком, проще всего создать синтаксис в IBM SPSS Modeler, выбрав опцию **Диалоговое окно IBM SPSS Statistics**, выбрав диалоговое окно для процедуры, заполнив его и нажав кнопку OK. При этом синтаксис будет вставлен во вкладку Синтаксис узла IBM SPSS Statistics, используемого вами в IBM SPSS Modeler. Затем можно запустить поток, чтобы получить вывод из процедуры.

Опция Редактор синтаксиса IBM SPSS Statistics

Проверить. После ввода команд синтаксиса в верхней части диалогового окна эта кнопка позволяет проверить введенные записи. Весь обнаруженный неправильный синтаксис выводится в нижней части диалогового окна.

Чтобы при поверке вами синтаксиса процесс проверки гарантированно не занимал слишком много времени, вместо проверки всего набора данных проверяется допустимость записей на представительной выборке данных.

Разрешаемый синтаксис

При наличии большого объема унаследованного синтаксиса из IBM SPSS Statistics или при достаточном знании возможностей подготовки данных IBM SPSS Statistics узел Statistics Transform можно использовать для выполнения многих из существующих у вас преобразований. В качестве основной линии узел позволяет преобразовывать данные предсказуемыми способами (например, посредством выполнения в циклах команд или путем изменения, добавления, фильтрации или выбора данных.

Примеры команд, которые можно выполнить:

- Вычислить случайные числа, соответствующие биномиальному распределению:
  `COMPUTE newvar = RV.BINOM(10000,0.1)`
- Перекодировать переменную в новую переменную:
  `RECODE Age (Lowest thru 30=1) (30 thru 50=2) (50 thru Highest=3) INTO AgeRecoded`
- Заменить пропущенные значения:

`RMV Age_1=SMEAN(Age)`

Синтаксис IBM SPSS Statistics, поддерживаемый узлом Statistics Transform, приведен в таблице ниже.

Команда

ADD VALUE LABELS

APPLY DICTIONARY

AUTORECODE

BREAK

Компакт-диск

CLEAR MODEL PROGRAMS

CLEAR TIME PROGRAM

CLEAR TRANSFORMATIONS

COMPUTE

COUNT

CREATE

DATE

DEFINE-!ENDDEFINE

DELETE VARIABLES

DO IF

DO REPEAT

ELSE

ELSE IF

END CASE

END FILE

END IF

END INPUT PROGRAM

END LOOP

END REPEAT

EXECUTE

FILE HANDLE

FILE LABEL

Команда

FILE TYPE-END FILE TYPE

FILTER

FORMATS

IF

INCLUDE

INPUT PROGRAM-END INPUT PROGRAM

INSERT

LEAVE

LOOP-END LOOP

MATRIX-END MATRIX

MISSING VALUES

N OF CASES

NUMERIC

PERMISSIONS

PRESERVE

РАНГ

RECODE

RENAME VARIABLES

RESTORE

RMV

SAMPLE

SELECT IF

SET

SORT CASES

STRING

SUBTITLE

TEMPORARY

TITLE

UPDATE

V2C

VALIDATEDATA

VALUE LABELS

VARIABLE ATTRIBUTE

VARSTOCASES

VECTOR

Узел Statistics Model

Узел Statistics Model позволяет анализировать данные и работать с ними, вызывая процедуру IBM SPSS Statistics, генерирующую PMML. Создаваемые вами слепки моделей можно затем использовать обычным образом в потоках IBM SPSS Modeler для скоринга и других задач.

Для использования этого узла на вашем компьютере должна быть установлена и залицензирована совместимая версия IBM SPSS Statistics. Дополнительную информацию смотрите в разделе “Вспомогательные прикладные программы IBM SPSS Statistics” на стр. 325. Информацию о совместимости смотрите в замечаниях по выпуску.

Состав доступных аналитических процедур IBM SPSS Statistics зависит от типа вашей лицензии.

Узел Statistics Model - Вкладка Модель

Имя модели Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Выбрать диалоговое окно. Нажмите эту кнопку, чтобы вывести список доступных процедур IBM SPSS Statistics, которые можно выбрать и запустить. Этот список содержит только процедуры, генерирующие PMML, на которые у вас есть лицензия, и не содержит процедуры, написанные пользователями.

1. Щелкните по нужной процедуре; откроется соответствующее диалоговое окно IBM SPSS Statistics.
2. В диалоговом окне IBM SPSS Statistics введите подробности для процедуры.
3. Нажмите кнопку **OK** для возврата на узел Statistics Model; на вкладке Модель появится синтаксис IBM SPSS Statistics.
4. Чтобы вернуться в диалоговое окно IBM SPSS Statistics в любое время, например, чтобы видоизменить запрос, нажмите кнопку вывода диалогового окна IBM SPSS Statistics справа от кнопки выбора процедур.

Узел Statistics Model - Сводка слепков моделей

При вызове узла Statistics Model он выполняет связанную процедуру IBM SPSS Statistics и создает слепок модели, который можно использовать в потоках IBM SPSS Modeler для скоринга.

На вкладке Сводка слепка модели выводится информация о полях, параметрах построения и процессе оценки модели. Результаты возвращаются в трех представлениях, которые можно развернуть и свернуть, щелкнув по конкретным позициям.

Кнопка **Просмотреть модель** выводит результаты в видоизмененной форме из средства просмотра вывода IBM SPSS Statistics. Дополнительную информацию об этом средстве просмотра смотрите в документации IBM SPSS Statistics.

Обычные опции экспорта и печати доступны в меню Файл. Дополнительную информацию смотрите в разделе “Просмотр вывода” на стр. 286.

Узел Statistics Output

Узел Statistics Output позволяет вызвать процедуру IBM SPSS Statistics для анализа исследуемых данных IBM SPSS Modeler. Результаты модно просмотреть в окне браузера или сохранить их в формате файлов вывода IBM SPSS Statistics. В IBM SPSS Modeler доступен широкий набор аналитических процедур IBM SPSS Statistics.

Для использования этого узла на вашем компьютере должна быть установлена и залицензирована совместимая версия IBM SPSS Statistics. Дополнительную информацию смотрите в разделе “Вспомогательные прикладные программы IBM SPSS Statistics” на стр. 325. Информацию о совместимости смотрите в замечаниях по выпуску.

При необходимости на вкладке Фильтр можно отфильтровать поля или переименовать их, чтобы они соответствовали стандартам именования IBM SPSS Statistics. Дополнительную информацию смотрите в разделе “Переименование или фильтрация полей для IBM SPSS Statistics” на стр. 361.

Руководство по синтаксису. Подробности о конкретных процедурах IBM SPSS Statistics смотрите в *Руководстве по синтаксису команд IBM SPSS Statistics*, включенному в вашу копию программного пакета IBM

SPSS Statistics. Чтобы просмотреть это руководство с вкладки Синтаксис, выберите опцию **Редактор синтаксиса** и нажмите кнопку Запустить справку по синтаксису IBM SPSS Statistics.

Узел Statistics Transforme - Вкладка Синтаксис

Эта вкладка используется для создания синтаксиса для процедуры IBM SPSS Statistics, которую вы хотите использовать для анализа данных. Синтаксис состоит из двух частей: **оператора** и связанных с ним **опций**. Оператор задает метод анализа или операцию для выполнения и используемые поля. Опции задают все остальное, включая статистику для вывода, сохраняемые полученные поля и так далее.

Опция диалогового окна IBM SPSS Statistics

Если синтаксис IBM SPSS Statistics для процедуры вам незнаком, проще всего создать синтаксис в IBM SPSS Modeler, выбрав опцию **Диалоговое окно IBM SPSS Statistics**, выбрав диалоговое окно для процедуры, заполнив его и нажав кнопку OK. При этом синтаксис будет вставлен во вкладку Синтаксис узла IBM SPSS Statistics, используемого вами в IBM SPSS Modeler. Затем можно запустить поток, чтобы получить вывод из процедуры.

Дополнительно можно сгенерировать узел источника Statistics File для импорта полученных данных. Эта опция полезна, например, если процедура записывает поля, такие как оценки, в активный набор данных отображения вывода на экран.

Чтобы создать синтаксис:

1. Нажмите кнопку **Выбрать диалоговое окно**.
2. Выберите одну из следующих опций:
   - **Анализ.** Возвращает содержание меню Анализ IBM SPSS Statistics; выберите процедуру, которую вы хотите использовать.
   - **Другое.** Эта опция, если она выводится, возвращает список диалоговых окон, созданных Конструктором настраиваемых диалоговых окон в IBM SPSS Statistics, также как и всех остальных диалоговых окон IBM SPSS Statistics, которые не выводятся в меню Анализ и для которых у вас есть лицензия. Если таковые диалоговые окна отсутствуют, эта опция не выводится.

Примечание: Диалоговые окна Автоматическая подготовка данных не выводятся.

Если у вас есть пользовательское диалоговое окно IBM SPSS Statistics, создающее новые поля, эти поля использовать в IBM SPSS Modeler будет невозможно, поскольку узел Statistics Output является терминальным.

Необязательно: включите переключатель **Сгенерировать узел импорта для полученных данных**, чтобы создать узел источника Statistics File, который можно использовать для импорта полученных данных в другой поток. Этот узел будет помещен на холст потока, с с данными, содержащимися в файле *.sav*, заданном в поле **Файл** (значение по умолчанию - каталог установки IBM SPSS Modeler).

Опция Редактор синтаксиса

Чтобы сохранить синтаксис, созданный для часто используемой процедуры:

1. Нажмите кнопку Опции файла (первую кнопку на панели инструментов).
2. В меню выберите **Сохранить** или **Сохранить как**.
3. Сохраните этот файл как файл *.sps*.

Чтобы использовать созданные ранее файлы синтаксиса, заменив текущее содержания редактора синтаксиса (если оно есть):

1. Нажмите кнопку Опции файла (первую кнопку на панели инструментов).
2. В меню выберите **Открыть**.

3. Выберите файл *.sps*, чтобы вставить его содержимое на вкладку Вывод узла вывода.

Чтобы вставить сохраненный ране синтаксис без замены текущего содержания:

1. Нажмите кнопку Опции файла (первую кнопку на панели инструментов).
2. В меню выберите **Вставить**.
3. Выберите файл *.sps*, чтобы вставить его содержимое в узел вывода туда, куда указывает указатель мыши.

Необязательно: включите переключатель **Сгенерировать узел импорта для полученных данных**, чтобы создать узел источника Statistics File, который можно использовать для импорта полученных данных в другой поток. Этот узел будет помещен на холст потока, с с данными, содержащимися в файле *.sav*, заданном в поле **Файл** (значение по умолчанию - каталог установки IBM SPSS Modeler).

При нажатии кнопки **Выполнить** результаты появятся в средстве просмотра вывода IBM SPSS Statistics. Дополнительную информацию об этом средстве просмотра смотрите в документации IBM SPSS Statistics.

Узел Statistics Output - Вкладка Вывод

На вкладке Вывод можно задать формат и положение вывода. Можно выбрать вывод результатов на экран или отправку их в файл одного из доступных типов.

Имя вывода. Задает имя вывода, генерируемого при вызове узла. Опция **Авто** выбирает имя на основе узла, генерирующего вывод. Дополнительно можно выбрать **Пользовательское**, чтобы задать другое имя.

Вывод на экран (по умолчанию)**.** Создает объект вывода для просмотра в оперативном режиме. Этот объект вывода появляется на вкладке Выходные поля окна менеджера при вызове узла вывода.

Вывод в файл. Сохраняет вывод в файле при вызове узла. При выборе этой опции введите имя файла в поле **Имя файла** (или перейдите в каталог и задайте имя файла при помощи кнопки средства выбора файлов) и выберите тип файла.

Тип файла. Выберите тип файла, в который вы хотите отправить вывод.

- **Документ HTML (*.html).** Записывает вывод в формате HTML.
- **Файл средства просмотра IBM SPSS Statistics (*.spv).** Записывает вывод в формате, который может быть прочитан средством просмотра вывода IBM SPSS Statistics.
- **Файл веб-отчетов IBM SPSS Statistics (*.spw).** Записывает вывод в формате веб-отчетов IBM SPSS Statistics, который можно опубликовать в репозитории IBM SPSS Collaboration and Deployment Services и в дальнейшем просмотреть в Web-браузере. Дополнительную информацию смотрите в разделе “Опубликовать в Web” на стр. 287.

Примечание: Если выбрать опцию **Вывод на экран**, директива OMS IBM SPSS Statistics `VIEWER=NO` действовать не будут; кроме того, API сценариев (модуль *Basic* и *Python SpssClient*) будут недоступны в IBM SPSS Modeler.

Узел Statistics Export

Узел Statistics Export позволяет экспортировать данные в формате IBM SPSS Statistics *.sav*. Файлы IBM SPSS Statistics *.sav* могут быть прочитаны модулем IBM SPSS Statistics Base и другими модулями. Формат этих файлов используется также для файлов кэша IBM SPSS Modeler.

При отображении имен полей IBM SPSS Modeler на имена переменных IBM SPSS Statistics иногда могут возникать ошибки, поскольку имена переменных IBM SPSS Statistics ограничены 64 символами и не могут содержать определенные символы, такие как пробелы, знаки доллара и дефисы (–). Эти ограничения можно скорректировать двумя способами:

- Щелкнув по вкладке Фильтр, можно переименовать имена полей в соответствии с требованиями к именам переменных IBM SPSS Statistics. Дополнительную информацию смотрите в разделе “Переименование или фильтрация полей для IBM SPSS Statistics”.
- Выбрать экспорт и имен, и меток полей из IBM SPSS Modeler.

Примечание: IBM SPSS Modeler записывает файлы *.sav* в формате Unicode UTF-8. IBM SPSS Statistics поддерживает файлы в формате Unicode UTF-8, только начиная с выпуска 16.0. Для предотвращения возможности повреждения данных файлы *.sav* в кодировке Unicode не следует использовать в выпусках IBM SPSS Statistics до выпуска 16.0. Дополнительную информацию смотрите в справке IBM SPSS Statistics.

Наборы множественных ответов. Все наборы множественных ответов, определяемые в потоке, будут автоматически сохраняться при экспорте файла. Наборы множественных ответов можно просмотреть и отредактировать с любого узла с вкладкой Фильтр. Дополнительную информацию смотрите в разделе “Редактирование наборов множественных ответов” на стр. 143.

Узел Statistics Export - Вкладка Экспорт

Экспортировать файл. Задает имя файла. Введите имя файла или нажмите кнопку средства выбора файлов, чтобы найти положение нужного файла.

Тип файла. Выберите, надо ли сохранить файл в обычном (*.sav*) или в сжатом (*.zsav*) формате.

Зашифровать файл с помощью пароля. Для защиты файла паролем включите переключатель; в отдельном диалоговом окне вам предложат ввести и подтвердить **Пароль**.

Примечание: Защищенные паролем файлы можно открыть только в IBM SPSS Modeler версии 16 или новее, а также в IBM SPSS Statistics версии 21 или новее.

Экспортировать имена полей. Задает способ обработки имен и меток переменных при экспорте из IBM SPSS Modeler в файл IBM SPSS Statistics *.sav* или *.zsav*.

- **Имена и метки переменных.** Выберите эту опцию для экспорта и имен, и меток полей IBM SPSS Modeler. Имена полей будут экспортироваться как имена переменных IBM SPSS Statistics SAS, а метки - как метки переменных IBM SPSS Statistics.
- **Имена как метки переменных.** Выберите эту опцию для использования имен полей IBM SPSS Modeler в IBM SPSS Statistics в качестве меток переменных. IBM SPSS Modeler разрешает в именах полей символы, недопустимые в именах переменных IBM SPSS Statistics. Для предотвращения возможного создания недопустимых имен IBM SPSS Statistics выберите вместо этой опцию **Имена и метки переменных** или скорректируйте имена полей при помощи вкладки Фильтр.

Запуск программы. Если на вашем компьютере установлена программа IBM SPSS Statistics, вы можете выбрать эту опцию для вызова прикладной программы непосредственно для сохраненного файла данных. Опции запуска прикладной программы должны быть заданы в диалоговом окне Вспомогательные прикладные программы. Дополнительную информацию смотрите в разделе “Вспомогательные прикладные программы IBM SPSS Statistics” на стр. 325. Чтобы просто создать файл IBM SPSS Statistics *.sav* или *.zsav* без открытия внешней программы, выключите эту опцию.

Сгенерировать узел импорта для этих данных. Выберите эту опцию для автоматического генерирования узла Statistics File, который будет читать экспортированный файл данных. Дополнительную информацию смотрите в разделе “Узел Statistics File” на стр. 354.

Переименование или фильтрация полей для IBM SPSS Statistics

Перед экспортом или внедрением данных из IBM SPSS Modeler во внешние прикладные программы, такие как IBM SPSS Statistics, может потребоваться переименовать или скорректировать имена полей. Диалоговые окна Statistics Transform, Statistics Output и Statistics Export содержат вкладку Фильтр, упрощающую этот процесс.

Основное описание вкладки фильтр приведено в другом месте. Дополнительную информацию смотрите в разделе “Задание опций фильтрации” на стр. 142. В этой теме даются советы по считыванию данных в IBM SPSS Statistics.

Чтобы скорректировать имена полей в соответствии с соглашениями об именованииIBM SPSS Statistics:

1. На вкладке Фильтр нажмите кнопку панели инструментов Меню опций фильтра (первую кнопку на панели инструментов).
2. Выберите Переименовать для IBM SPSS Statistics.
3. В диалоговом окне Переименовать для IBM SPSS Statistics можно выбрать замену недопустимых символов в именах файлов либо на **символ решетки (#)**, либо на **символ подчеркивания (_)**.

Переименовать наборы множественных ответов . Выберите эту опцию, если хотите скорректировать имена наборов множественных ответов, которые можно импортировать в IBM SPSS Modeler при помощи узла источника Statistics File. Наборы множественных ответов используются для записи данных, содержащих несколько значений для каждого наблюдения, таких как ответы на опросы.

Глава 9. надузлы

Обзор надузлов

Одна из причин, по которым визуальный интерфейс программирования IBM SPSS Modeler так легко освоить, заключается в том, что у каждого узла - четко определенная функция. Однако для сложной обработки может потребоваться длинная последовательность узлов. В конечном счете, холст потока может быть загроможден, и станет трудно отслеживать диаграммы потока. Чтобы избежать загромождения холста длинным и сложным потоком:

- Последовательность обработки можно разбить на несколько потоков, перетекающих один в другой. Например, первый поток создает файл данных, который второй поток использует в качестве входного. Второй поток создает файл, который третий поток использует в качестве входного, и так далее. Этими несколькими потоками можно управлять, сохраняя их в **проекте**. Проект обеспечивает организацию нескольких потоков и их вывода. Однако файл проекта содержит только ссылку на помещенные в него объекты, и придется все равно управлять несколькими файлами потоков.
- Работая со сложными процессами потоков, можно создать **Надузел** в качестве более простой альтернативы.

Надузлы группируют несколько узлов в один узел, инкапсулируя сегменты потока данных. Благодаря этому, исследователь данных получает многочисленные преимущества:

- Более аккуратные и управляемые потоки.
- Возможность объединения узлов в один надузел для конкретного бизнеса.
- Возможность экспорта надузлов в библиотеки для повторного использования в нескольких проектах исследования данных.

Типы надузлов

В потоке данных надузлы представлены значком в виде звезды. Для этого значка используется затенение, показывающее тип надузла и направление потока (к узлу или от узла).

Надузлы бывают трех типов:

- Надузлы источников
- Надузлы процессов
- Терминальные надузлы

Надузлы источников

Надузлы источника содержат источник данных точно также, как и обычный узел источника, и могут использоваться всюду, где может использоваться обычный узел этого типа. Левая сторона надузла источника затенена, указывая на то, что слева он "закрыт" и что данные должны проходить вниз по потоку *От* надузла.

У надузла источника есть только одна точка соединения, справа, показывающая, что данные покидают надузел и поступают в поток.

Надузлы процессов

Надузлы процессов содержат только узлы процессов, и они не затенены, что указывает на возможность прохождения данных и *К* надузлу, и *От* надузла этого типа.

У надузлов есть точки соединения и слева, и справа, показывающие, что данные прибывают на надузел и покидают его, поступая обратно в поток. Хотя надузлы и могут содержать дополнительные сегменты потока и даже добавочные потоки, обе точки соединения должны проходить через один путь, соединяющий точки *Из потока* и *В поток*.

Примечание: Надузлы процессов иногда называют также *надузлами преобразования*.

Терминальные надузлы

Терминальные надузлы содержат один или несколько терминальных узлов (графиков, таблиц и так далее), и их можно использовать тем же способом, что и терминальный узел. Правая сторона терминального надузла затенена, указывая на то, что справа он "закрыт" и что данные могут только поступать *К* терминальному надузлу.

У терминального надузла есть только одна точка соединения, слева, показывающая, что данные прибывают на надузел и прекращают движение на этом надузле.

Терминальные надузлы могут также содержать сценарии, используемые для задания порядка вызова всех терминальных улов в данном надузле. Дополнительную информацию смотрите в разделе “Надузлы и сценарии” на стр. 369.

Создание надузлов

Создание надузлов "уплотняет" поток данных, благодаря инкапсулированию нескольких узлов в одни узел. После загрузки потока на холст или его создания предлагается несколько способов создать надузел.

Множественный выбор

Чтобы создать надузел простейшим способом, выберите все узлы, которые вы хотите инкапсулировать:

1. При помощи мыши выберите несколько узлов на холсте потока. Можно также использовать клавишу Shift+щелчок мыши, чтобы выбрать поток или сегмент потока. *Примечание*: Узлы, которые вы выбираете, должны быть из непрерывного или из разветвленного потока. Нельзя выбрать узлы, не являющиеся смежными или не соединенные каким-либо способом между собой.
2. Затем инкапсулируйте выбранные узлы, одним из трех способов:
   - Щелкните на панели инструментов по значку надузла.
   - Щелкните правой кнопкой мыши по надузлу и в контекстном меню выберите:

     Создать надузел > Из выбранного
   - В меню Надузел выберите:

     Создать надузел > Из выбранного

Все эти опции инкапсулируют узлы в надузел, который будет оттенен в соответствии его типом (источника, процесса или терминальный) на основе содержимого.

Одиночный выбор

Можно также создать надузел, выбрав один узел и определив при помощи опций меню начало и завершение надузла либо инкапсулировав все вниз по потоку от выбранного узла.

1. Щелкните по узлу, определяющему начало инкапсуляции.
2. В меню Надузел выберите:

   Создать надузел > Отсюда

Надузлы можно также создавать более интерактивно, выбирая для инкапсуляции узлов начало и завершение сегмента потока.

1. Щелкните по первому или последнему узлу, который вы хотите инкапсулировать в надузел.

2. В меню Надузел выберите:

   Создать надузел > Выбрать...

3. Другой способ - воспользуйтесь опциями контекстного меню, щелкнув по нужному узлу правой кнопкой мыши.
4. Указатель превратится в значок надузла, указывающий, что вы должны выбрать другую точку в потоке. Переместитесь либо вверх, либо вниз по потоку к "другой конечной точке" сегмента надузла и щелкните по узлу. При этом действии все узлы в промежутке будут заменены значком Надузел в виде звезды.

Примечание: Узлы, которые вы выбираете, должны быть из непрерывного или из разветвленного потока. Нельзя выбрать узлы, не являющиеся смежными или не соединенные каким-либо способом между собой.

Вложение надузлов

Надузлы могут быть вложенными в другие надузлы. Одни и те же правила для каждого типа надузлов (источника, процесса и терминального) применяются и к вложенным надузлам. Например, у надузла процесса с вложением должен быть непрерывный поток данных через все вложенные надузлы, чтобы он оставался надузлом процесса. Если один из вложенных узлов будет терминальным, данные больше не будут проходить через эту иерархию.

Терминальные надузлы и надузлы источников могут содержать вложенные надузлы других типов, но к ним применяются одни и те же базовые правила создания надузлов.

Блокировка надузлов

Создав надузел, его можно заблокировать паролем для предотвращения внесения в него исправлений. Например, это можно сделать, если вы создаете потоки или их сегменты в виде шаблонов фиксированных значений для других пользователей в вашей организации, у которых меньше опыта конфигурирования запросных систем IBM SPSS Modeler.

Если надузел заблокирован, пользователи все равно смогут ввести на вкладке Параметры значения для любых параметров, которые были определены, а заблокированный надузел можно будет вызвать без ввода пароля.

Примечание: Блокировку и разблокировку нельзя выполнять при помощи сценариев.

Блокировка и разблокировка надузла

Предупреждение: Утерянные пароли восстановлению не подлежат.

Надузел можно заблокировать и разблокировать с любой из трех вкладок.

1. Нажмите кнопку **Блокировать узел**.
2. Введите и подтвердите пароль.
3. Щелкните по **OK**.

На защищенный паролем надузел на холсте потока указывает маленький символ замка вверху слева от значка Надузел.

Разблокируйте надузел

1. Чтобы удалить защиту паролем на постоянной основе, нажмите кнопку **Разблокировать узел**; появится приглашение ввести пароль.
2. Введите пароль и нажмите кнопку **OK**; Надузел больше не будет защищен паролем, и символ замка перестанет выводиться рядом со значком в потоке.

Редактирование заблокированного надузла

При попытке задать параметры или увеличить масштаб для вывода на экран заблокированного надузла появляется приглашение ввести пароль.

Введите пароль и щелкните по **OK**.

Теперь вы можете редактировать определения параметров и изменять масштаб сколько угодно часто, пока не будет закрыт поток, в котором находится надузел.

Имейте в виду, что эта операция не удаляет защиту паролем, а только позволяет получить доступ к работе с надузлом. Дополнительную информацию смотрите в разделе “Блокировка и разблокировка надузла” на стр. 365.

Редактирование надузлов

Создав надузел, его можно проверить более детально, увеличив масштаб надузла; если надузел заблокирован, появится приглашение ввести пароль. Дополнительную информацию смотрите в разделе “Редактирование заблокированного надузла”.

Для просмотра содержимого надузла можно использовать значок увеличения масштаба на панели инструментов IBM SPSS Modeler или следующий метод:

1. Щелкните правой кнопкой мыши по надузлу.
2. В контекстном меню выберите **Увеличить**.

Содержимое выбранного надузла появится в слегка отличающиеся среде IBM SPSS Modeler с соединителями, показывающими прохождение данных через поток или сегмент потока. На этом уровне холста потока представлены несколько задач, которые можно выполнить:

- Изменить тип надузла (источника, процесса или терминальный).
- Создать параметры или отредактировать значения параметра. Параметры используются в сценариях и в выражениях CLEM.
- Задать опции кэширования для надузла и его подузлов.
- Создать или видоизменить сценарий надузла (только для терминальных надузлов).

Изменение типов надузлов

При некоторых обстоятельствах полезно изменить тип надузла. Эта опция доступна только в режиме увеличенного масштаба и применима у надузлу только на этом уровне. В следующей таблице описаны три типа надузлов.

Таблица 46. Типы надузлов.

Тип надузла	Описание
Надузел источника	Одно исходящее соединение
Надузел процесса	Два соединения: одно входящее и одно исходящее
Терминальный надузел	Одно входящее соединение

Чтобы изменить тип надузла:

1. Убедитесь, что находитесь в режиме увеличенного масштаба.
2. В меню Надузел выберите **Тип надузла**, затем выберите тип.

Аннотирование и переименование надузлов

Для надузла можно изменить имя вывода в потоке, а также написать аннотации, используемые в проекте или отчете. Чтобы обратиться к этим свойствам:

- Щелкните правой кнопкой мыши по надузлу (режим уменьшенного масштаба) и выберите **Переименовать и аннотировать**.
- Другой вариант - выберите **Переименовать и аннотировать** в меню Надузел. Эта опция доступна в режимах и увеличенного, и уменьшенного масштаба.

В обоих случаях откроется диалоговое окно с выбранной вкладкой Аннотации. При помощи опций этой вкладки задайте пользовательское имя, выводимое на холсте потока, и введите текст, относящийся к операциям надузла.

Использование комментариев с надузлами

При создании надузла из содержащего комментарий узла или слепка комментарий нужно включить в выбор для создания надузла, если вы хотите, чтобы этот комментарий появился в надузле. Если опустить комментарий в выборе, при создании надузла комментарий останется в потоке.

При открытии надузла с включенными в него комментариями они будут восстановлены там, где они были до его создания.

При открытии надузла с включенными в него объектами с комментариями, но не включенными комментариями, в прежнем положении будут восстановлены объекты, но комментарии повторно прикреплены не будут.

Параметры надузла

В IBM SPSS Modeler предусмотрена возможность задания пользовательских переменных, таких как *Minvalue*, значения которых могут быть определены для использования в сценариях или выражениях CLEM. Эти переменные называются **параметрами**. Можно задать параметры для потоков, сеансов и надузлов. Все параметры, заданные для надузла, будут доступны при построении выражений CLEM на этом надузле и всех вложенных узлах. Параметры, заданные для вложенных надузлов, для их родительского надузла недоступны.

Создание и задание параметров для надузлов выполняется в два шага:

1. Для надузла определяются параметры.
2. Затем для каждого параметра надузла задается значение.

После этого заданные параметры можно использовать в выражениях CLEM для всех инкапсулированных узлов.

Определение параметров надузла

Параметры для надузла можно определить как в режиме уменьшенного, так и в режиме увеличенного масштаба. Заданные параметры применяются ко всем инкапсулированных узлам. Для определения параметров надузла сначала нужно получить доступ к вкладке Параметры диалогового окна Надузел. Используйте для открытия этого диалогового окна один из следующих способов:

- Щелкните дважды кнопкой мыши по надузлу в потоке.
- В меню Надузел выберите **Задать параметры**.
- Другой вариант - выберите **Задать параметры** в контекстом меню, находясь в режиме увеличения масштаба.

После открытия этого диалогового окна вы увидите вкладку Параметры со всеми ранее определенными параметрами.

Чтобы задать новый параметр:

Нажмите кнопку **Задать параметры**, открывающую указанное диалоговое окно.

Имя. Здесь перечислены имена параметров. Введя имя в этом поле, можно создать новый параметр. Например, чтобы создать параметр для минимальной температуры, можно ввести `minvalue`. Не включайте префикс `$P-`, обозначающий параметр в выражениях CLEM. Это имя используется также для вывода в Построителе выражений CLEM.

Длинное имя. Выводит описательное имя для каждого созданного параметра.

Хранение. Выберите тип хранения файла из списка. Хранение указывает способ хранения значений данных в параметре. Например, при работе со значениями, содержащими начальные нули, которые вы хотите сохранить (например, 008), в качестве типа хранения следует выбрать **Строка**. В противном случае нули будут убраны из значения. Доступно хранение типов строка, целое, действительное число, время, дата и отметка времени. Для параметров дат следует учитывать, что значения нужно задавать при помощи нотации стандарта ISO, как показано в следующей теме.

Значение. Выводит текущее значение для каждого созданного параметра. Настройте параметр нужным вам образом. Имейте в виду, что для параметров дат значения должны задаваться в нотации стандарта ISO (то есть: `ГГГГ-ММ-ДД`). Даты, заданные в других форматах, не принимаются.

Тип (необязательно). Если планируется внедрение потока во внешнюю прикладную программу, выберите в списке шкалу измерений. В противном случае рекомендуется оставить столбец *Тип* как есть. Если вы хотите задать для параметра ограничения значений, такие как верхняя и нижняя границы для числового диапазона, выберите в списке **Задать**.

Имейте в виду, что полное имя, опции хранения и типа можно задать для параметров только через пользовательский интерфейс. С помощью сценариев эти опции задать нельзя.

Для перемещения выбранного параметра вверх или вниз по списку доступных параметров используйте кнопки со стрелками в правой части окна. Для удаления выбранного параметра используйте кнопку удаления (с меткой *X*).

Задание значений для параметров надузла

После определения параметров для надузла можно задать значения при использовании параметров в выражении или сценарии CLEM.

Чтобы задать параметры надузла:

1. Щелкните дважды кнопкой мыши по значку Надузел, чтобы открыть диалоговое окно Надузел.
2. Другой вариант - в меню Надузел выберите **Задать параметры**.
3. Щелкните по вкладке **Параметры**. *Примечание*: Поля в этом диалоговом окне - это поля, определяемые нажатием кнопки **Задать параметры** на указанной вкладке.
4. Введите в текстовом поле значение для каждого параметра, который вы создали. Например, можно задать значение параметра *minvalue* для конкретного исследуемого порога. Этот параметр можно затем использовать в многочисленных операциях, таких как выбор записей выше или ниже этого порога для дальнейшего исследования.

Использование параметров надузла для обращения к свойствам узла

Параметры надузла можно также использовать для определения свойств узлов (другое название - **параметры слота**) для инкапсулированных узлов. Предположим, вы хотите указать, чтобы надузел обучал инкапсулированный узел нейросети в течение определенного времени при помощи случайной выборки доступных данных. При помощи этих параметров можно задать значения для продолжительности периода и процента выборки.

Предположим, ваш надузел примера содержит узел выборки с именем *Sample* и узел нейросети с именем *Train*. При помощи диалоговых окон узлов можно задать параметр узла выборки **Sample** как **Random %**, а параметр узла нейросети **Stop on** как **Time**. Задав эти опции можно обратиться к свойствам узла с

параметрами и задать конкретные значения для надузла. В диалоговом окне Надузел нажмите кнопку **Задать параметры** и создайте параметры, указанные в следующей таблице.

Таблица 47. Создаваемые параметры

Параметр	Значение	Длинное имя
Train.time	5	Время на обучение (в минутах)
Sample.random	10	Процент случайной выборки

Примечание: В именах параметров, таких как *Sample.random*, используется верный синтаксис для именования свойства узла, где *Sample* - это имя узла, а *random* - его свойство.

Задав эти параметры, можно легко изменить значения для свойств двух узлов (выборки и нейросети), не открывая повторно каждое диалоговое окно. Вместо этого просто выберите опцию **Задать параметры** в меню надузла, чтобы открыть вкладку Параметры диалогового окна Надузел, где можно задать для **Random %** и **Time** новые значения. В частности, это полезно при исследовании данных во время многократных итераций построения модели.

Надузлы и кэширование

На надузле можно кэшировать все узлы, кроме терминальных. Кэшированием можно управлять, щелкнув правой кнопкой мыши по узлу и выбрав одну из нескольких опций в контекстном меню Кэш. Эта опция меню доступна и вне надузла, и для узлов, инкапсулированных в надузел.

Для кэшей надузла соблюдается несколько правил:

- Если на любом из узлов, инкапсулированных в надузел, включено кэширование, оно также будет включено и на надузле.
- Отключение кэша на надузле отключает кэш для *всех* инкапсулированных узлов.
- Включение кэширования на надузле фактически включает кэш на последнем кэшируемом подузле. Другими словами, если последний подузел - узел выбора, кэш будет включен для этого узла выбора. Если последний подузел - терминальный (где кэширование не разрешено), кэширование будет включено для следующего поддерживающего кэширование узла в восходящем потоке.
- При задании кэшей для подузлов надузла все операции в восходящем потоке узла в кэше, такие как добавление или редактирование узлов, выгружаются из кэшей.

Надузлы и сценарии

На языке сценариев SPSS Modeler можно написать простые программы, преобразующие и вызывающие элементы содержимого терминального надузла. Например, можно задать порядок обработки сложного потока. Предположим, если надузел содержит узел задания глобальных значений, который требуется вызвать перед узлом графика, можно создать сценарий, вызывающий сначала узел задания глобальных значений. Значения, вычисляемые этим узлом, такие как среднее или среднеквадратичное отклонение, можно затем использовать при вызове узла графика.

Вкладка Сценарий диалогового окна Надузел доступна только для терминальных надузлов.

Чтобы открыть диалоговое окно Сценарии для терминального надузла:

- Щелкните правой кнопкой мыши по холсту надузла и выберите **Сценарий надузла**.
- Другой вариант - в режиме и увеличенного, и уменьшенного масштаба можно выбрать опцию **Сценарий надузла** в меню Надузел.

Примечание: Сценарии надузлов выполняются с потоком и надузлом, только если в диалоговом окне выбрана опция **Выполнить этот сценарий**.

Конкретные опции для написания сценария и его использования в SPSS Modeler обсуждаются в *Руководстве по сценариям и автоматизации*, доступном в виде файла PDF в составе скачиваемого продукта.

Сохранение и загрузка надузлов

Одно из преимуществ надузлов состоит в том, что их можно сохранять и повторно использовать в других потоках. При сохранении и загрузке надузлов имейте в виду, что для них используется расширение *.slb*.

Чтобы сохранить надузел:

1. Увеличьте масштаб надузла.
2. В меню Надузел выберите **Сохранить надузел**.
3. В диалоговом окне задайте имя файла и каталог.
4. Выберите, следует ли добавить сохраненный надузел в текущий проект.
5. Нажмите кнопку **Сохранить** .

Чтобы загрузить надузел:

1. В меню Вставка в окне IBM SPSS Modeler выберите **Надузел**.
2. Выберите надузел (*.slb*) в текущем каталоге или найдите другой.
3. Нажмите кнопку **Загрузить**.

Примечание: У всех параметров импортированных улов будут значения по умолчанию. Чтобы их изменить, щелкните дважды по надузлу на холсте потока.

Уведомления

Эта информация относится к продуктам и сервису, предлагаемым в США. Этот материал может быть доступен от IBM на других языках. Однако для его получения может понадобиться приобрести продукт или версию продукта на нужном языке.

IBM может не предоставлять в других странах продукты, услуги и аппаратные средства, описанные в данном документе. За информацией о продуктах и услугах, предоставляемых в вашей стране, обращайтесь к местному представителю IBM. Ссылки на продукты, программы или услуги IBM не означают и не предполагают, что можно использовать только указанные продукты, программы или услуги IBM. Разрешается использовать любые функционально эквивалентные продукты, программы или услуги, если при этом не нарушаются права IBM на интеллектуальную собственность. Однако ответственность за оценку и проверку работы любого продукта, программы или сервиса, не произведенного корпорацией IBM, лежит на пользователе.

IBM может располагать патентами или рассматриваемыми заявками на патенты, относящимися к предмету данного документа. Предъявление данного документа не предоставляет какую-либо лицензию на эти патенты. Вы можете послать письменный запрос о лицензии по адресу:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

По поводу лицензий, связанных с использованием наборов двухбайтных символов (DBCS), обращайтесь в отдел интеллектуальной собственности IBM в вашей стране или направьте запрос в письменной форме по адресу:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION ПРЕДСТАВЛЯЕТ ДАННУЮ ПУБЛИКАЦИЮ "КАК ЕСТЬ", БЕЗ КАКИХ-ЛИБО ГАРАНТИЙ, КАК ЯВНЫХ, ТАК И ПОДРАЗУМЕВАЕМЫХ, ВКЛЮЧАЯ, НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ, ПРЕДПОЛАГАЕМЫЕ ГАРАНТИИ СОБЛЮДЕНИЯ ЧЬИХ-ЛИБО АВТОРСКИХ ПРАВ, ВОЗМОЖНОСТИ КОММЕРЧЕСКОГО ИСПОЛЬЗОВАНИЯ ИЛИ ПРИГОДНОСТИ ДЛЯ КАКИХ-ЛИБО ЦЕЛЕЙ И СООТВЕТСТВИЯ ОПРЕДЕЛЕННОЙ ЦЕЛИ. В некоторых странах для ряда сделок не допускается отказ от явных или предполагаемых гарантий; в таком случае данное положение к вам не относится.

Эта информация может содержать технические неточности и типографские ошибки. В представленную здесь информацию периодически вносятся изменения; эти изменения будут включаться в новые издания данной публикации. Фирма IBM может в любое время без уведомления вносить изменения и усовершенствования в продукты и программы, описанные в этой публикации.

Любые ссылки в этой публикации на сайты, не принадлежащие IBM, приведены только для удобства и никоим образом не означают их поддержки. Материалы на этих сайтах не входят в число материалов по данному продукту IBM, и весь риск пользования этими сайтами несете вы сами.

Любую предоставленную вами информацию IBM может использовать или распространять любым способом, какой сочтет нужным, не беря на себя никаких обязательств по отношению к вам.

Если обладателю лицензии на данную программу понадобятся сведения о возможности: (i) обмена данными между независимо разработанными программами и другими программами (включая данную) и (ii) совместного использования таких данных, он может обратиться по адресу:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Такая информация может быть доступна при соответствующих условиях и соглашениях, включая в некоторых случаях взимание платы.

Описанную в данном документе лицензионную программу и все прилагаемые к ней лицензированные материалы IBM предоставляет на основе положений Соглашения между IBM и Заказчиком, Международного Соглашения о Лицензиях на Программы IBM или любого эквивалентного соглашения между IBM и заказчиком.

Данные производительности и примеры клиентов представлены только для иллюстрации. Фактическая производительность зависит от конкретной конфигурации и условий работы.

Информация о продуктах других компаний (не IBM) получена от поставщиков этих продуктов, из их опубликованных объявлений или из иных общедоступных источников. IBM не производила тестирование этих продуктов и никак не может подтвердить информацию о их точности работы и совместимости, а также прочие заявления относительно продуктов других компаний (не IBM). Вопросы о возможностях продуктов других компаний (не IBM) следует направлять поставщикам этих продуктов.

Все утверждения о будущих планах и намерениях IBM могут быть изменены или отменены без уведомлений, и описывают исключительно цели фирмы.

Эти сведения содержат примеры данных и отчетов, используемых в повседневных деловых операциях. Чтобы проиллюстрировать их настолько полно, насколько это возможно, данные примеры включают имена индивидуумов, названия компаний, брендов и продуктов. Все эти имена являются вымышленными и любое их сходство с реальными именами и адресами предприятий является случайным.

Товарные знаки

IBM, логотип IBM, и ibm.com являются товарными знаками или зарегистрированными товарными знаками компании International Business Machines Corp., зарегистрированными во многих странах мира. Прочие наименования продуктов и услуг могут быть товарными знаками, принадлежащими IBM или другим компаниям. Текущий список товарных знаков IBM смотрите на веб-сайте "Copyright and trademark information" (Информация об авторских правах и товарных знаках) по адресу www.ibm.com/legal/copytrade.shtml.

Adobe, логотип Adobe, PostScript и логотип PostScript являются либо зарегистрированными товарными знаками, либо товарными знаками корпорации Adobe Systems в Соединенных Штатах и/или других странах.

Intel, логотип Intel, Intel Inside, логотип Intel Inside, Intel Centrino, логотип Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium и Pentium являются товарными знаками или зарегистрированными товарными знаками компании Intel или ее дочерних компаний в Соединенных Штатах и других странах.

Linux является зарегистрированным товарным знаком Linus Torvalds в Соединенных Штатах и других странах.

Microsoft, Windows, Windows NT и логотип Windows являются товарными знаками корпорации Microsoft в Соединенных Штатах и других странах.

UNIX является зарегистрированным товарным знаком The Open Group в Соединенных Штатах и других странах.

Java и все основанные на Java товарные знаки и логотипы - товарные знаки или зарегистрированные товарные знаки Oracle и/или его филиалов.

Правила и условия для документации продукта

Разрешения для использования этих публикаций предоставляются на следующих условиях.

Применимость

Данные правила и условия являются дополнением к правилам использования для сайта IBM.

Персональное использование

Вы можете воспроизводить эти публикации для персонального некоммерческого использования при условии сохранения всех замечаний о правах собственности. Вам запрещается распространять эти публикации, полностью или по частям, демонстрировать их или создавать из них производные продукты без явного на то согласия от IBM.

Коммерческое использование

Вам предоставляется право воспроизводить эти публикации исключительно в пределах своего предприятия при условии, что будут воспроизведены все замечания об авторских правах. За пределами вашего предприятия вам запрещается распространять эти публикации, полностью или по частям, демонстрировать их или создавать из них производные продукты без явного на то согласия от IBM.

Права

За исключением прав, явным образом предоставляемых настоящим разрешением, никаких иных разрешений, лицензий и прав, ни явных, ни подразумеваемых, в отношении публикаций и любой содержащейся в них информации, данных, программ или иной интеллектуальной собственности, не предоставляется.

IBM оставляет за собой право отозвать разрешения, предоставленные этим документом, если, по мнению IBM, использование публикаций наносит ущерб IBM или, как это установлено IBM, вышеприведенные инструкции не соблюдаются должным образом.

Запрещается загружать, экспортировать или реэкспортировать эту информацию, если при этом не будут полностью соблюдаться все применимые законы и постановления, включая все законы и постановления США, касающиеся экспорта.

IBM НЕ ДАЕТ НИКАКИХ ГАРАНТИЙ ОТНОСИТЕЛЬНО СОДЕРЖАНИЯ ЭТИХ ПУБЛИКАЦИЙ. ПУБЛИКАЦИИ ПРЕДСТАВЛЯЮТСЯ "КАК ЕСТЬ", БЕЗ КАКИХ-ЛИБО ГАРАНТИЙ, ЯВНЫХ ИЛИ ПОДРАЗУМЕВАЕМЫХ, ВКЛЮЧАЯ (НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ) ПРЕДПОЛАГАЕМЫЕ ГАРАНТИИ ОТСУТСТВИЯ НАРУШЕНИЙ, КОММЕРЧЕСКОЙ ПРИГОДНОСТИ ИЛИ СООТВЕТСТВИЯ ОПРЕДЕЛЕННОЙ ЦЕЛИ.

Глоссарий

К

Ковариация. Ненормированная мера связи между двумя переменными, равная сумме попарных произведений отклонений, деленной на N-1.

K

Эксцесс. Мера сгруппированности наблюдений вокруг центральной точки. Для нормального распределения значение эксцесса равно 0. Положительный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы более плотно около центра и имеют более тонкие хвосты до экстремумов распределения, и более толстые хвосты в области экстремальных значений. Отрицательный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы менее плотно около центра и имеют более толстые хвосты до экстремумов распределения, и более тонкие хвосты в области экстремальных значений.

H

Максимум. Наибольшее значение числовой переменной.

Mean. Мера центральной тенденции. Арифметическое среднее; сумма, деленная на число наблюдений.

Медиана. Значение, выше и ниже которого попадает по половине наблюдений, иначе 50-й процентиль. Если число наблюдений четно, медиана есть арифметическое среднее двух находящихся в середине значений, если выборку упорядочить по убыванию или по возрастанию. Медиана представляет собой меру центральной тенденции, которая нечувствительна к выбросам, в отличие от среднего значения, которое могут исказить несколько экстремально больших или малых значений.

Минимум. Наименьшее значение числовой переменной.

Мода. Чаще всего встречающееся значение. Если таких значений несколько, каждое из них является модой.

R

Range . Разность между наибольшим и наименьшим значениями числовой переменной; максимум минус минимум.

Ю

Асимметрия. Мера асимметрии распределения. Нормальное распределение симметрично, и для него асимметрия равна 0. Распределение со значимой положительной асимметрией имеет длинный хвост справа. Распределение со значимой отрицательной асимметрией имеет длинный хвост слева. В качестве грубого правила можно сказать, что значение асимметрии, более чем вдвое превышающее ее стандартную ошибку, указывает на наличие асимметрии распределения.

стандартное отклонение. Мера разброса вокруг среднего, равная квадратному корню из дисперсии. Стандартное отклонение измеряется в тех же единицах, что и исходная переменная.

Стандартное отклонение. Мера дисперсии вокруг среднего, выраженная в тех же единицах измерения, что и наблюдения. Равна корню квадратному из дисперсии. При нормальном распределении 68% наблюдений укладываются в одно стандартное отклонение от среднего, и 95% - в два стандартных отклонения. Если, например, средний возраст равен 45 годам со стандартным отклонением 10, то 95% наблюдений должны оказаться между 25 и 65 годами при нормальном распределении.

Стандартная ошибка. Мера того, насколько значение статистики критерия меняется от выборки к выборке. Это стандартное отклонение выборочного распределения статистики. Например, стандартная ошибка среднего - это стандартное отклонение выборочных средних.

Стандартная ошибка эксцесса . Отношение эксцесса к его стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение эксцесса указывает, что хвосты распределения длиннее, чем у нормального; отрицательное значение эксцесса указывает на более короткие хвосты (как у равномерного распределения).

Стандартная ошибка среднего. Мера того, как сильно могут отличаться значения среднего от выборки к выборке, извлекаемых из одного и того же распределения. Можно применять для грубого сравнения наблюденного среднего с гипотетическим значением (то есть можно заключить, что два значения различаются, если отношение их разности к стандартному отклонению меньше -2 или больше +2).

Стандартная ошибка асимметрии . Отношение асимметрии к ее стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение асимметрии указывает на длинный правый хвост (распределения); большое отрицательное значение - на длинный левый хвост.

Sum. Сумма или итог для всех значений по всем наблюдениям, имеющим непропущенные значения.

U

Уникальные. Оцениваются все эффекты одновременно, каждый эффект корректируется по всем остальным эффектам любого вида.

З

Допустимо. Допустимые наблюдения не содержат ни системных, ни пользовательских пропущенных значений.

Дисперсия. Мера дисперсии относительно среднего значения. Равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньшее числа наблюдений. Дисперсия измеряется в единицах, которые равны квадратам единиц измерения самой переменной.

Индекс

Спец. символы

C

E

F

H

I

O

P

R

S

T

A

Б

B

Г

Д

Е

З

И

К

Л

М

Н

C

T

У

Ф

Х

Ц

Ч

Ш

Э

Я

Напечатано в Дании