

***STATISTICA* Data Mining**

Передовые технологии
анализа данных



data analysis • data mining • quality control • web-based analytics

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Bulgaria: StatSoft Bulgaria Ltd.
Czech Rep.: StatSoft Czech Rep. s.r.o.
China: StatSoft China

France: StatSoft France
Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
India: StatSoft India Pvt. Ltd.
Israel: StatSoft Israel Ltd.

Italy: StatSoft Italia srl
Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Norway: StatSoft Norway AS

Poland: StatSoft Polska Sp. z o.o.
Portugal: StatSoft Ibérica Lda
Russia: StatSoft Russia
Spain: StatSoft Ibérica Lda

S. Africa: StatSoft S. Africa (Pty) Ltd.
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

STATISTICA Data Miner – Добыча Данных

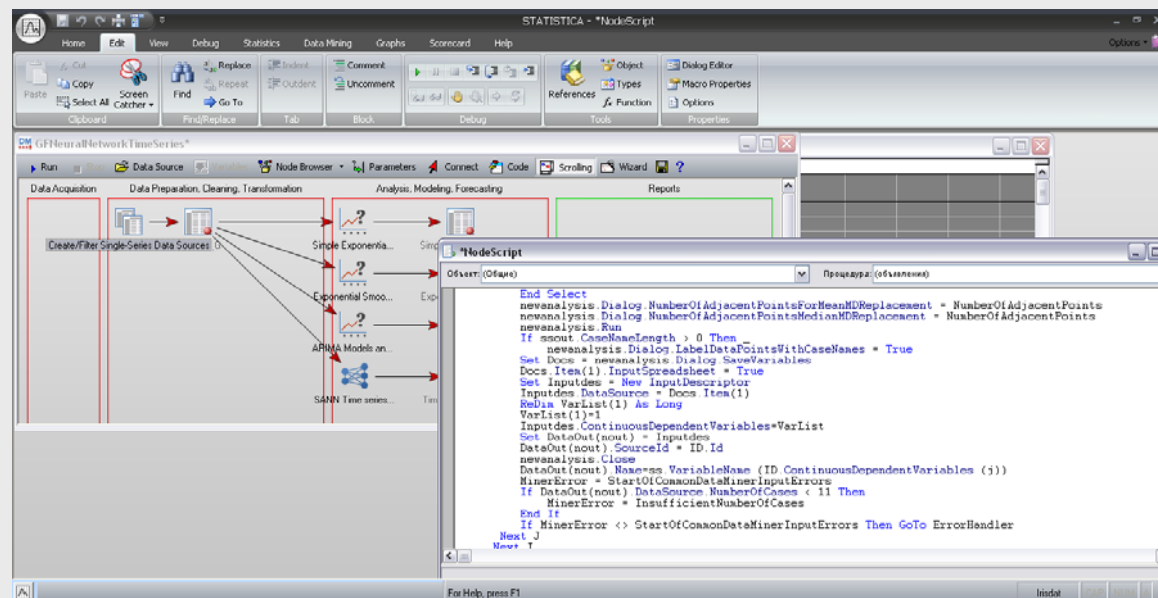
- Наиболее полный пакет методов Data Mining на рынке программного обеспечения
- Большой набор готовых решений
- Удобный пользовательский интерфейс, полностью интегрированный с MS Office
- Мощные средства разведочного анализа

STATISTICA Data Miner – Добыча Данных

- Полностью оптимизированный пакет для работы с огромным объемом информации
- Гибкий механизм управления
- Многозадачность системы
- Чрезвычайно быстрое и эффективное прогнозирование и классификация новых значений

STATISTICA Data Miner – Добыча Данных

- Неограниченные возможности автоматизации
- Поддержка пользовательских приложений:
 - Visual Basic (является встроенным языком)
 - Java
 - C/C++
 - PMML



Открытая COM архитектура

- COM – объект экспонирует определенные методы, позволяющие вступать с ним во взаимодействие
- COM – интерфейс – это средство, с помощью которого пользователь объекта получает доступ к его функциональным
- Различные среды разработки компьютерных программ, поддерживающие COM – интерфейс, позволяют создавать мощные вычислительные программы, использующие методы анализа *STATISTICA* и *STATISTICA Data Miner*

Примеры применения Добычи данных

■ Телекоммуникации

Базы данных содержат порядка несколько десятков тысяч записей и для нахождения связей, для выделения целевых групп необходимо использовать Добычу Данных

■ Маркетинг

В супермаркете имеется огромный товарный ряд, и управляющему важно знать взаимосвязи между продажами, выделением однородных групп товаров и прогнозирование продаж

Примеры применения Добычи данных


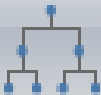
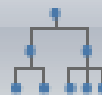




















- Геологоразведка

Нахождение месторождений полезных ископаемых, например газа и нефти, по результатам проб, географических особенностей местности

- Прогноз погоды

Для надежного прогнозирования погоды приходится анализировать большие базы данных, содержащие данные о температуре, ветре, осадках и других климатических явлениях

Линейка инструментов Data Miner

 Data Miner Recipes Recipes	       C&RT CHAID I-Trees Boosted Trees Random Forests MARS Splines Trees/Partitioning	 Neural Networks  Machine Learning  GAM Learning	 IC Analysis  Optimal Binning  Cluster... Clustering/Grouping
 Text Mining  Web Crawling Text Mining	 Association Rules  Link Analysis Rule Extraction	 Rapid Deployment  Goodness of Fit Deployment	 Workspaces ▾  Optimizations ▾  Feature Selection Tools

В пакете *Statistica Data Miner* предлагается исчерпывающий набор процедур анализов и методов визуализации данных

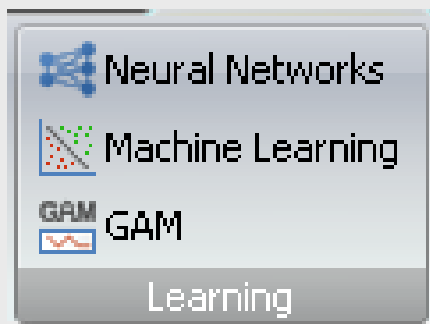
Линейка инструментов Data Miner



Мастер добычи данных

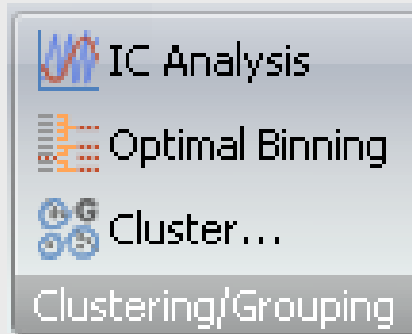


Trees/Partitioning – Деревья классификации



Learning – Обучение

Линейка инструментов Data Miner



Clustering/Grouping –
Кластеризация/Группировка



Text Mining – Добыча текста

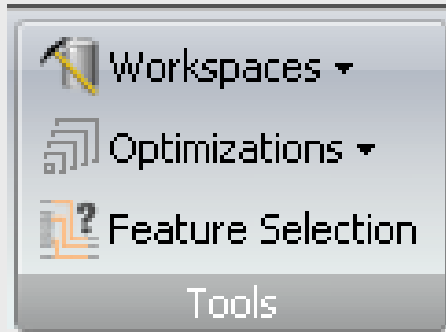


Rule Extraction – Поиск связей

Линейка инструментов Data Miner



Deployment – Внедрение



Tools – Инструменты

Линейка инструментов Data Miner

- Interactive Drill-Down Explorer - Интерактивный углубленный разведчик
- Feature Selection and Variable Filtering (for very large data sets) - Специальная выборка и фильтрация данных (для больших объемов данных)
- Association Rules - Правила ассоциации
- Generalized EM & k-Means Cluster Analysis - Обобщенный метод максимума среднего и K средних кластерного анализа

Линейка инструментов Data Miner

- Generalized Additive Models (GAM) - Обобщенные аддитивная модели (GAM)
- General Classification and Regression Trees (GTrees) - Обобщенные деревья классификации и регрессии (GTrees)
- General CHAID (Chi-square Automatic Interaction Detection) Models - Обобщенные CHAID модели (Хи-квадрат, автоматическое обнаружение взаимодействия)
- Interactive Classification and Regression Trees - Интерактивная деревья классификации и регрессии

Линейка инструментов Data Miner

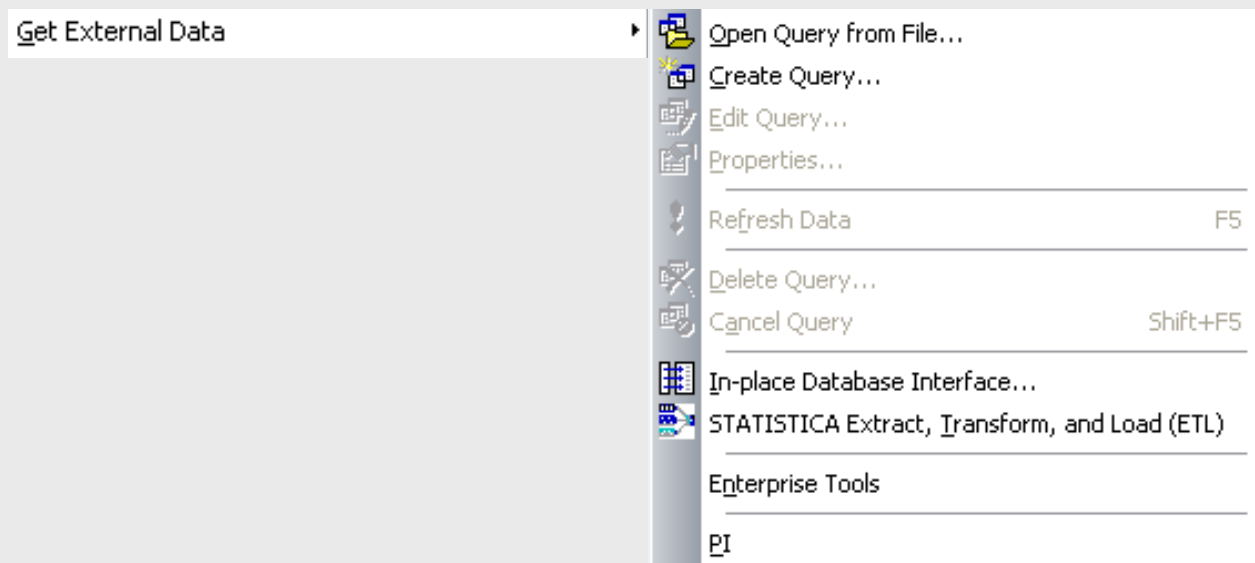
- Boosted Trees – Растущие деревья
- Multivariate Adaptive Regression Splines (Mar Splines) - Многомерные адаптивные регрессионные сплайны (Mar Splines)
- Goodness of Fit Computations – Критерии согласия
- Rapid Deployment of Predictive Models - Быстрые прогнозирующие модели (для большого числа наблюдаемых значений)

Соединение *STATISTICA* с базами данных

- Система *STATISTICA* предоставляет гибкие возможности для взаимодействия с базами данных различной архитектуры
- Ключевым понятием является **запрос *STATISTICA***
- *STATISTICA Query* используется для легкого доступа к данным целого ряда баз данных с помощью технологий Microsoft OLE DB (Object Linking and Embedding Database)

Создание запроса. Шаг 1

- Выбираем в *STATISTICA* Create Query – Создать Запрос

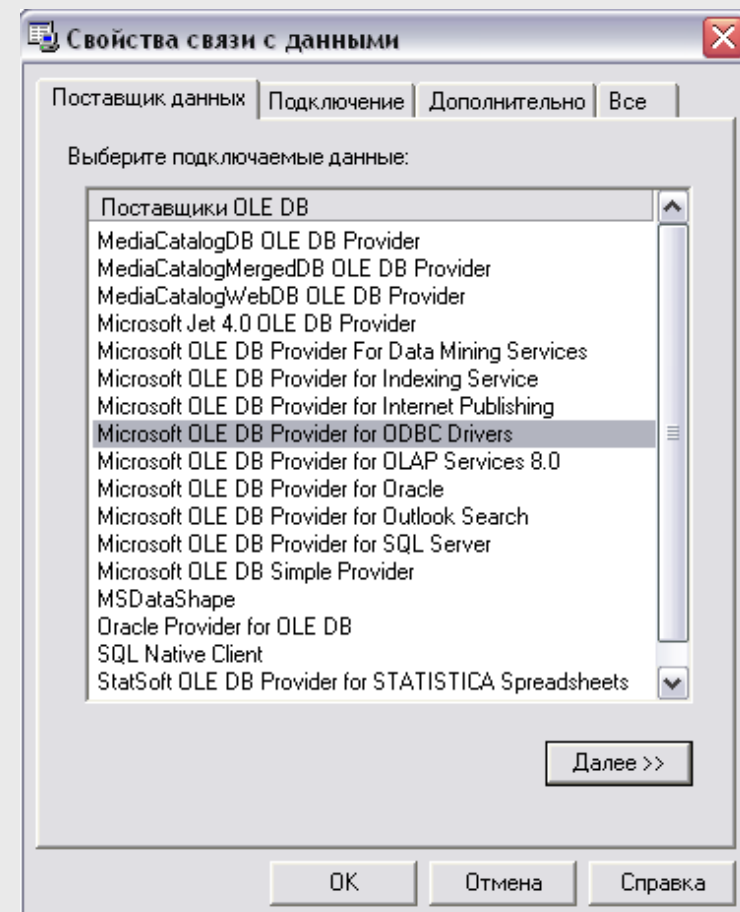


- Попадаем в рабочее пространство *STATISTICA* Query, где мы можем выбрать существующее подключение к БД или создать новое подключение

Создание запроса. Шаг 2

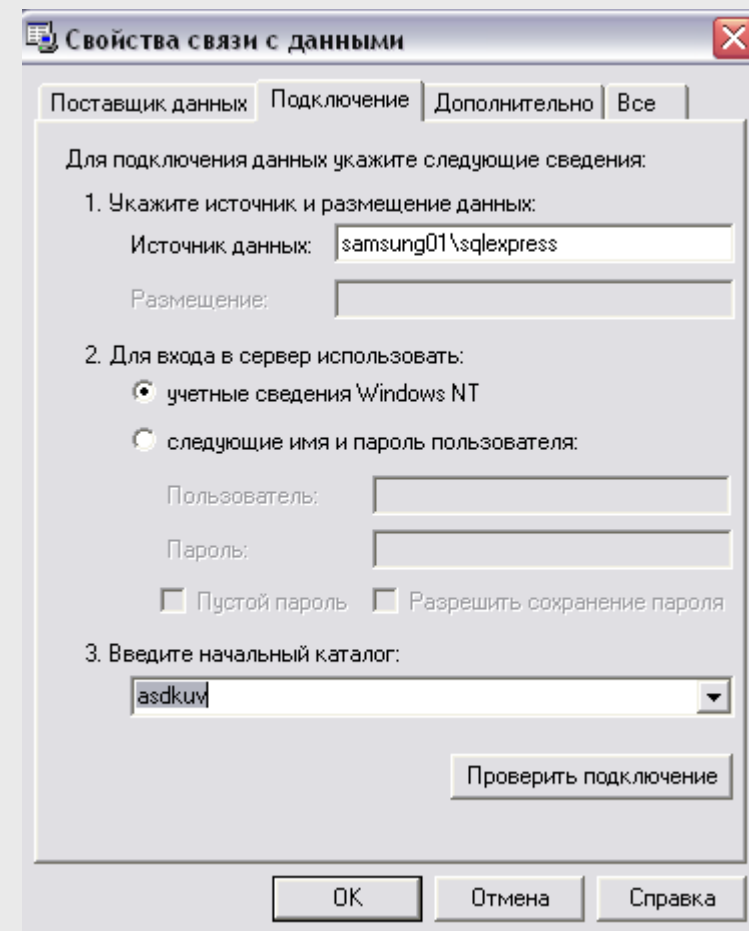
- *Поставщик данных* используется для выбора подходящего поставщика OLE DB для тех данных, к которым осуществляется доступ

- В списке отражаются все обнаруженные на жестком диске поставщики OLE DB



Создание запроса. Шаг 2

- Переключаемся на вкладку *Подключение*, соответствующую выбранному поставщику OLE DB
- Присутствуют только те свойства соединения, которые необходимы для поставщика OLE DB
- Если вход на сервер ограничен, то необходимо ввести имя пользователя и пароль, необходимые для подключения к источнику данных



Свойства связи с данными

Поставщик данных | Подключение | Дополнительно | Все

Для подключения данных укажите следующие сведения:

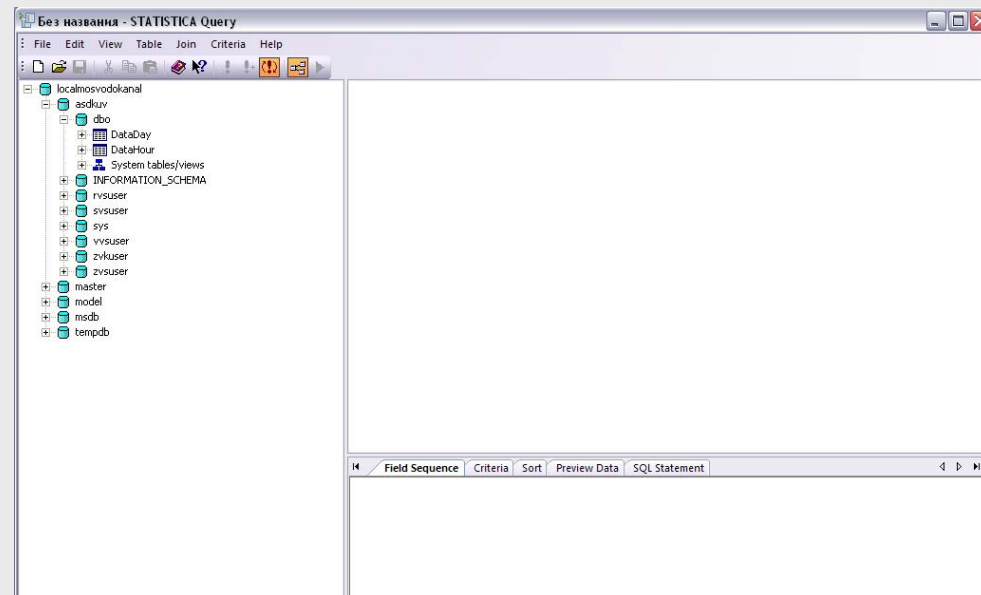
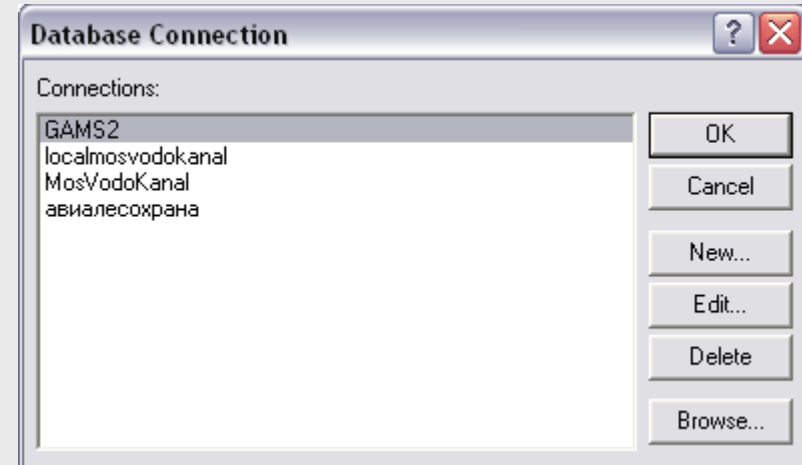
1. Укажите источник и размещение данных:
Источник данных: samsung01\sqlexpress
Размещение:
2. Для входа в сервер использовать:
 учетные сведения Windows NT
 следующие имя и пароль пользователя:
Пользователь:
Пароль:
 Пустой пароль Разрешить сохранение пароля
3. Введите начальный каталог:

Проверить подключение

OK Отмена Справка

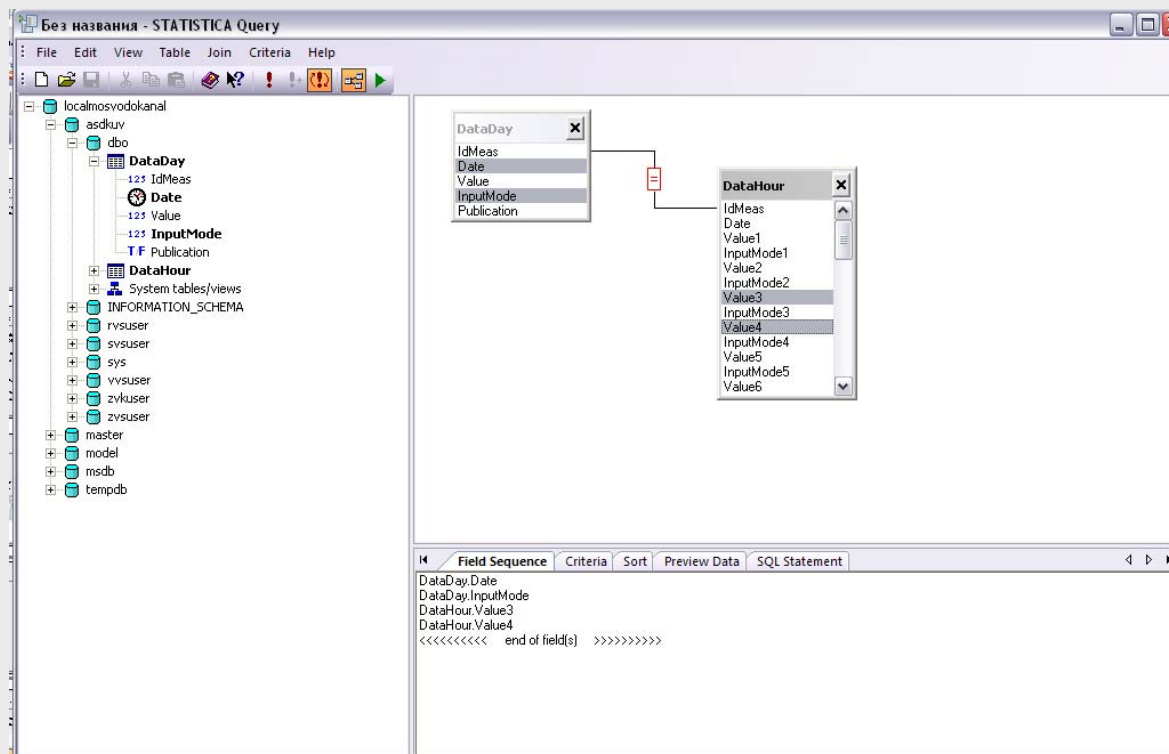
Создание запроса. Шаг 3

- Выбираем БД, к которой хотим подключиться
- Запрос *STATISTICA* предлагает два режима работы: Графический и Текстовый.
- Графический режим - это самый простой и удобный режим в *STATISTICA*



Создание запроса. Шаг 3

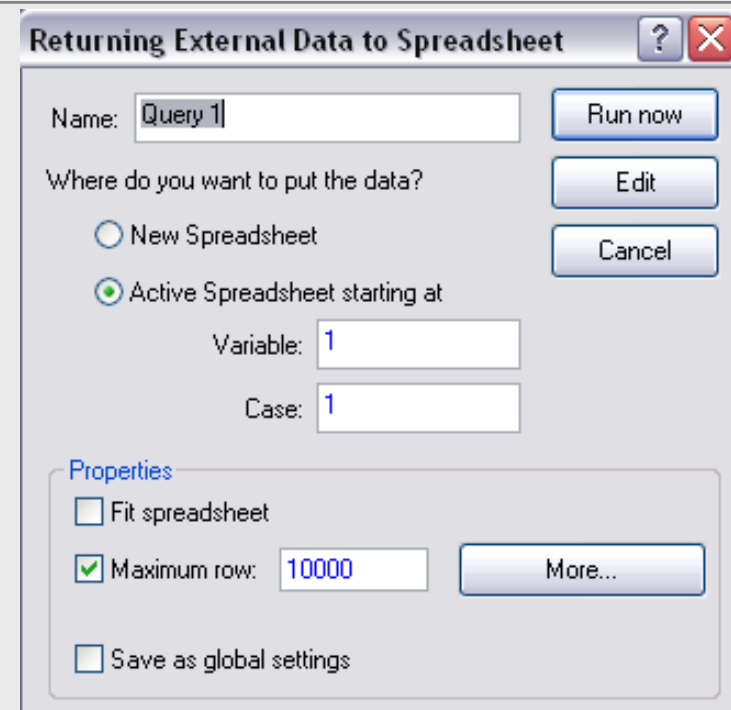
- Легкость и простота создания запроса
- Добавление критерия или связи



Для ручной правки SQL запроса нужно переключиться в Текстовый режим ввода запросов в *STATISTICA*

Создание запроса. Шаг 4

- Экспорт внешних данных в таблицу
- Размещение данных в таблицы
- Опция Максимальное число строк направлена против случайного создания запроса, когда в результате получается очень большое число Записей
- Дополнительные опции для редактирования экспорта данных

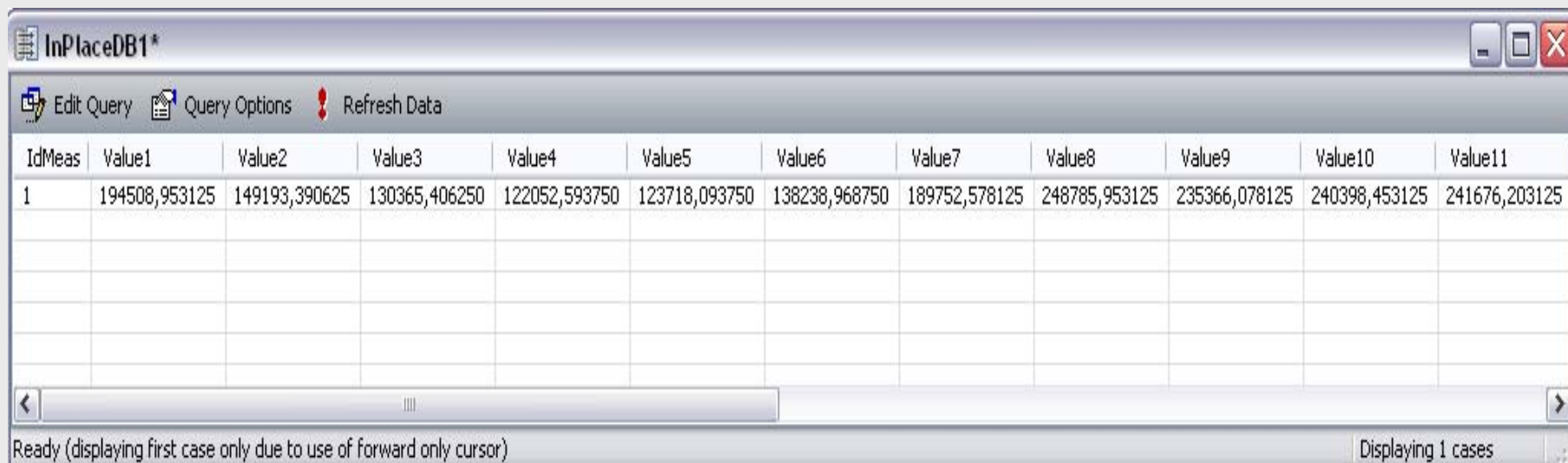


Особенности *STATISTICA* Query

- Простота подключения к базам данных
- Легкость создания запросов
- Добавление критериев (условий на выкачиваемые из бд данные) для запросов
- Если запрос содержит больше одной таблицы, Запрос *STATISTICA* автоматически создает связь между двумя таблицами при обнаружении связи в базе данных между двумя полями различных таблиц

In-Place Database Interface

- Технология In-Place Database Interface – это “Обработка данных на месте”
- Большинство статистических процедур могут обрабатывать данные в удаленных базах данных, не требуя, при этом, копирования на локальный компьютер



The screenshot shows a window titled "InPlaceDB1*" with a toolbar containing "Edit Query", "Query Options", and "Refresh Data". Below the toolbar is a table with 11 columns labeled "IdMeas", "Value1", "Value2", "Value3", "Value4", "Value5", "Value6", "Value7", "Value8", "Value9", "Value10", and "Value11". The first row contains the following values: 1, 194508,953125, 149193,390625, 130365,406250, 122052,593750, 123718,093750, 138238,968750, 189752,578125, 248785,953125, 235366,078125, 240398,453125, and 241676,203125. The status bar at the bottom indicates "Ready (displaying first case only due to use of forward only cursor)" and "Displaying 1 cases".

IdMeas	Value1	Value2	Value3	Value4	Value5	Value6	Value7	Value8	Value9	Value10	Value11
1	194508,953125	149193,390625	130365,406250	122052,593750	123718,093750	138238,968750	189752,578125	248785,953125	235366,078125	240398,453125	241676,203125

Data Mining Recipes - Мастер добычи данных

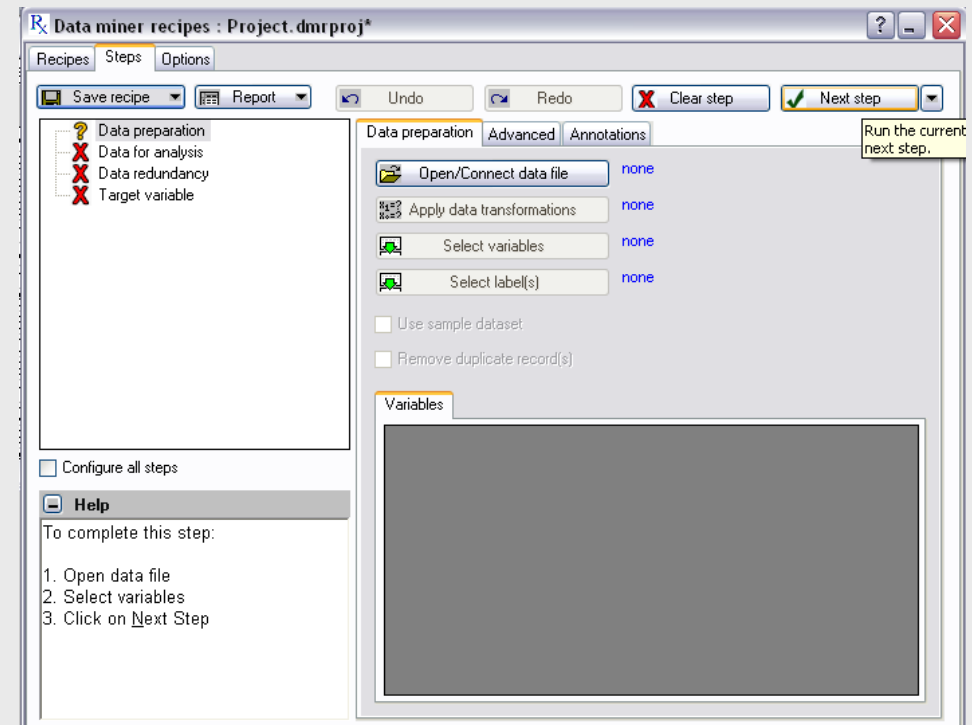
- Систематический подход для построения продвинутых аналитических моделей для одной или более целевой величины от группы входных переменных

- Построение предсказывающих моделей для задач регрессии и проблем классификации данных:
 - Нейронные сети
 - Случайные леса
 - Опорные вектора
 - Растущие деревья
 - Общие деревья регрессии и классификации

Простота использования Мастера

- Data Mining Recipes - Мастер добычи данных будет одинаково полезен в анализе данных как опытным аналитикам, так и начинающим пользователям

- Новичок в области добычи данных сможет построить интересные и разумные модели прогноза и классификации благодаря понятному, Многофункциональному интерфейсу Мастера



Особенности Data Mining Recipes

- Соединение с БД, поддерживающими *ODBC* или *OLEDB* через интерфейс **In-Place Database Interface**
- Чистка данных: удаление пропусков, повторных наблюдений данных
- Определение существенных предикторов, которые влияют на целевую (зависимую) переменную
- Создание автоматических конкурентных оценок моделей для определения оптимальной модели относительно эффективности и сложности

Data Mining Recipes - Мастер добычи данных

- *STATISTICA Data Miner Recipes* позволяет автоматизировать построение моделей.

- Этапы построения моделей:
 - Загрузка данных
 - Изменение/ приготовление данных
 - Проведение вычислений
 - Просмотр результатов
 - Использование модели на новых данных
 - Сохранение

Data Mining Recipes - Мастер добычи данных

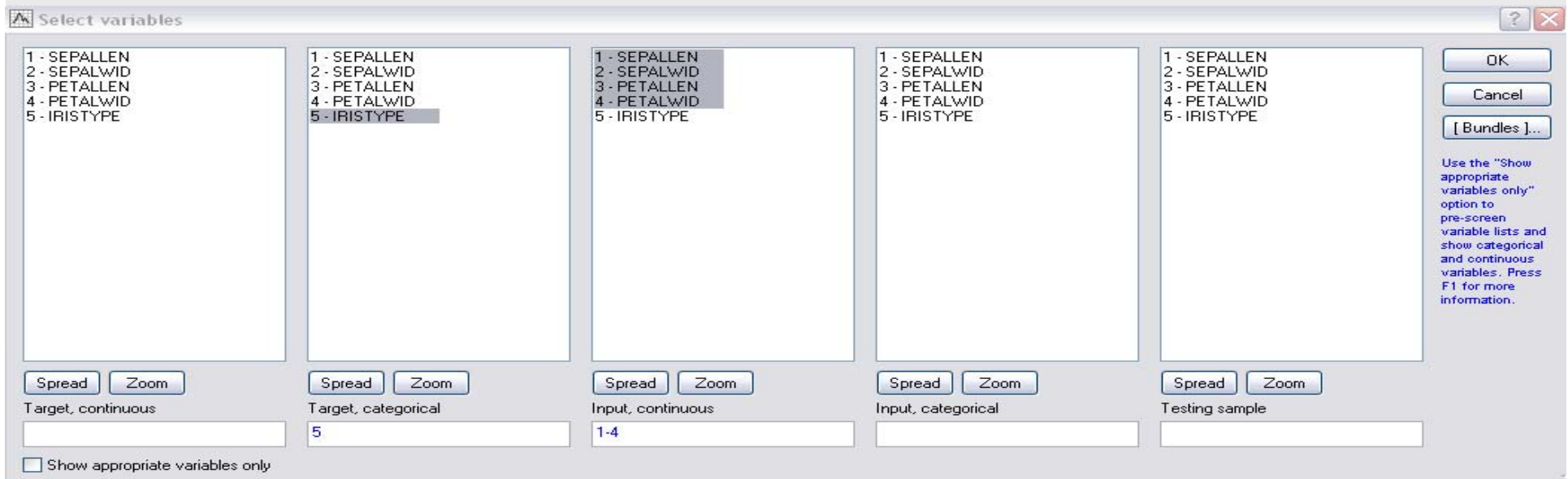
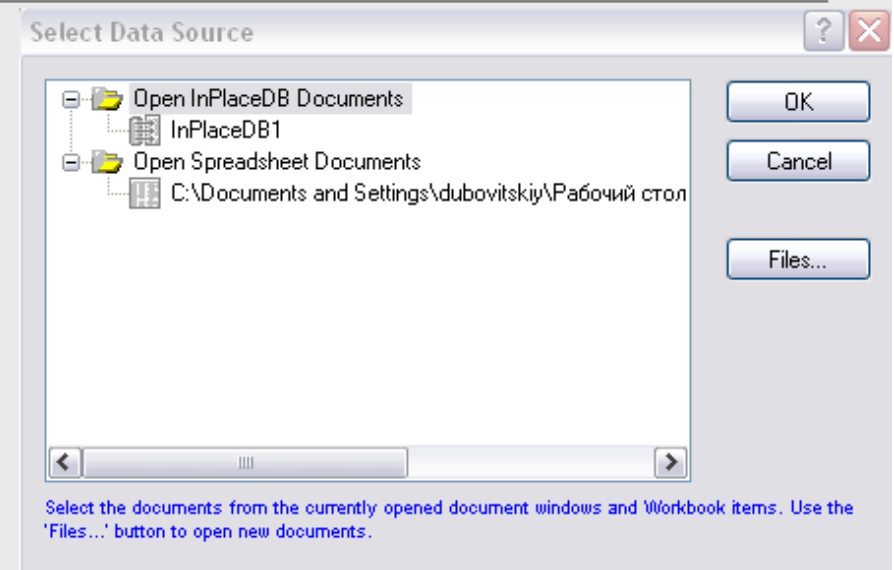
- Начнем знакомство с Data Miner с применения Мастера добычи данных к классическому примеру Фишера – задачи классификации цветов ириса

Fisher (1936) iris data: length & width of sepals and petals, 3 types of Iris					
	1 SEPALLEN	2 SEPALWID	3 PETALLEN	4 PETALWID	5 IRISTYPE
1	5,0	3,3	1,4	0,2	SETOSA
2	6,4	2,8	5,6	2,2	VIRGINIC
3	6,5	2,8	4,6	1,5	VERSICOL
4	6,7	3,1	5,6	2,4	VIRGINIC
5	6,3	2,8	5,1	1,5	VIRGINIC
6	4,6	3,4	1,4	0,3	SETOSA
7	6,9	3,1	5,1	2,3	VIRGINIC
8	6,2	2,2	4,5	1,5	VERSICOL
9	5,9	3,2	4,8	1,8	VERSICOL
10	4,6	3,6	1,0	0,2	SETOSA
11	6,1	3,0	4,6	1,4	VERSICOL
12	6,0	2,7	5,1	1,6	VERSICOL
13	6,5	3,0	5,2	2,0	VIRGINIC
14	5,6	2,5	3,9	1,1	VERSICOL

- Рассмотрим пошаговый алгоритм работы мастера

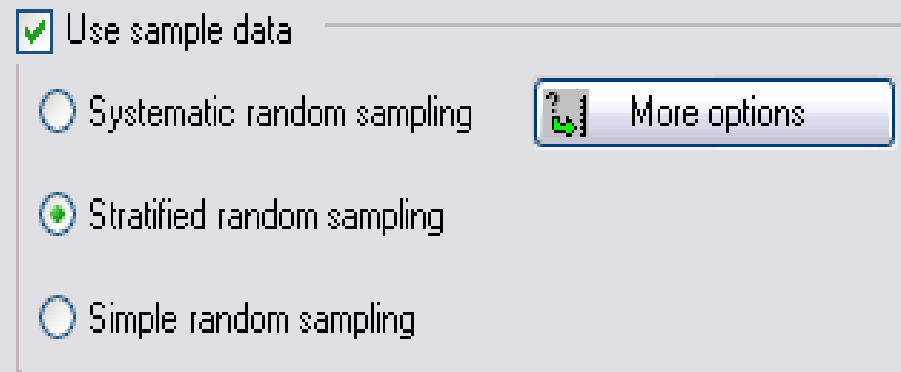
Шаг 1. Загрузка данных

- Выбор источника данных
- Преобразование переменных в таблице
- Выбор целевых и входных переменных



Шаг 2. Создание подвыборок

- Если вы анализируете клиентскую базу из миллиона записей, то часто рекомендуется сделать подвыборку для проведения быстрого анализа
- Мастер добычи данных позволяет пользователи создавать подвыборки в исходном массиве данных:
 - Стратифицированная выборка
 - Систематическая выборка
 - Случайная выборка




Use sample data

Systematic random sampling

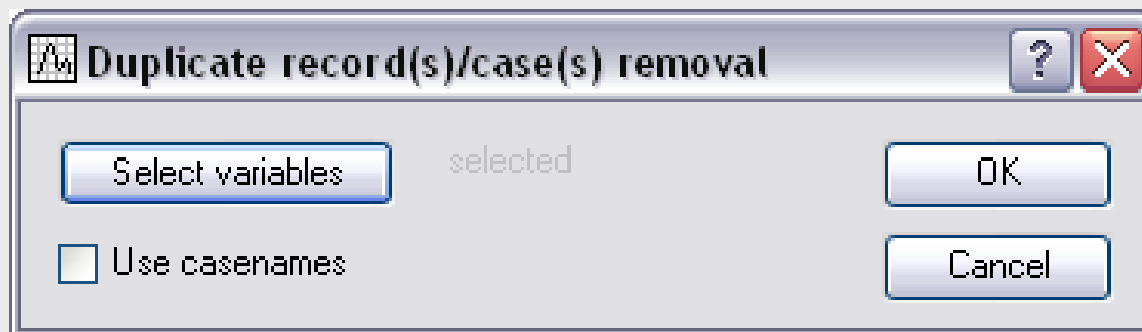
Stratified random sampling

Simple random sampling

 More options

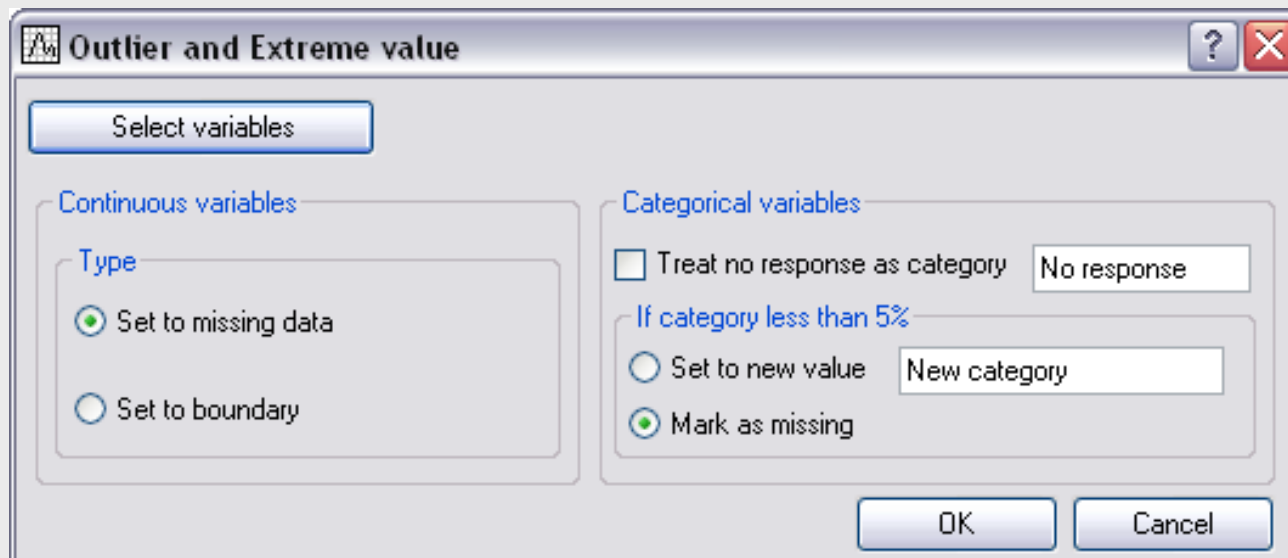
Шаг 3. Удаление повторных наблюдений (редубликаций)

- В реальных данных часто имеются повторные записи. Например, вы анализируете базу данных оператора мобильной связи, насчитывающей около миллиона записей. Необходимо исключить повторные записи, которые будут создавать помехи для анализа



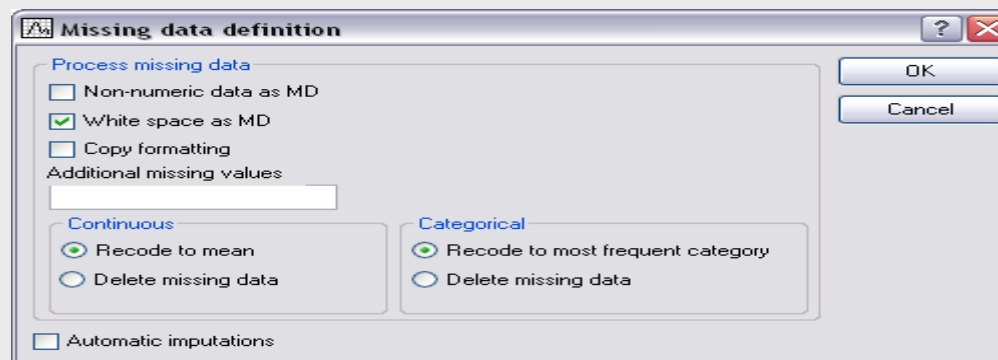
Шаг 4. Обработка выбросов

- Обработка выбросов и экстремальных значений является одним из важных этапов предварительной очистки данных
- Мастер Добычи Данных предлагает автоматизированную процедуру обработки выбросов в исходных данных



Шаг 5. Обработка пропущенных значений

- В реальных данных, как правило, имеются пропуски
- Мастер позволяет автоматизировать процедуру обработки пропущенных значений
- Например, замена пропусков для непрерывных переменных – средним значением, а категориальных переменных – наиболее часто встречающимся

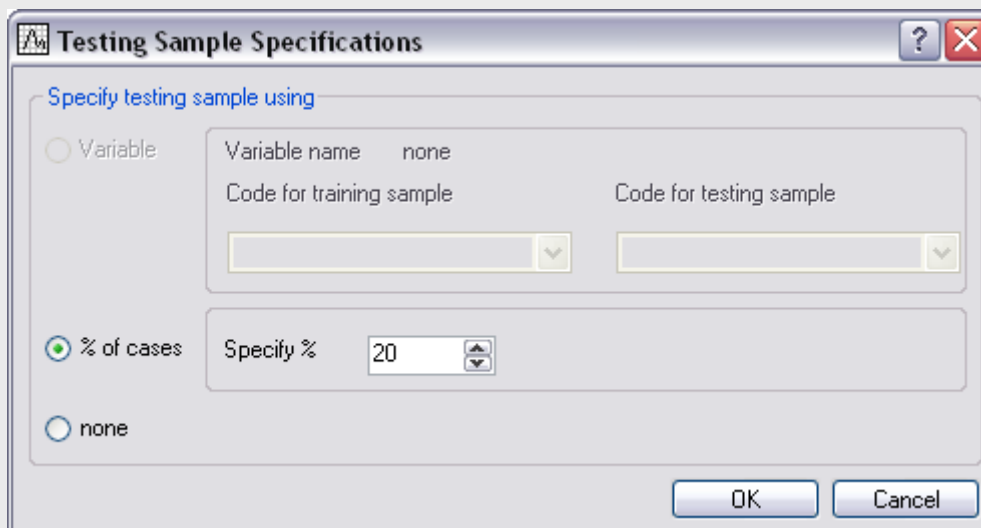


Шаг 6. Подготовка данных к построению модели классификации

■ Вычисление основных статистик

	Type	Role	Mean	Standard deviation	Skewness	Kurtosis	Observed minimum
▶ SEPALLEN	Continuous	Input	5,87	0,85	0,28	-0,57	4,30
SEPALWID	Continuous	Input	3,05	0,41	0,24	0,33	2,00
PETALLEN	Continuous	Input	3,82	1,77	-0,32	-1,34	1,00
PETALWID	Continuous	Input	1,22	0,75	-0,15	-1,30	0,10
IRISTYPE	Categorical	Target					

■ Создание тестовых выборок для кросс-проверки



Testing Sample Specifications

Specify testing sample using

Variable

Variable name none

Code for training sample

Code for testing sample

% of cases

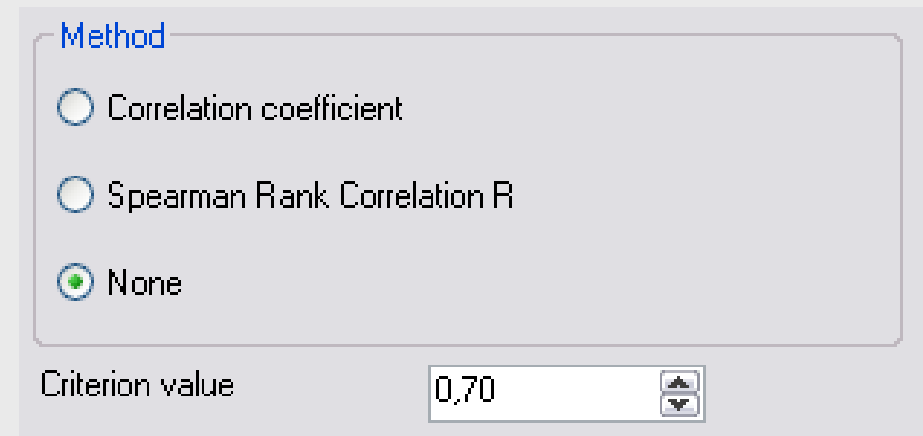
Specify % 20

none

OK Cancel

Шаг 7. Определение некоррелированных предикторов

- Корреляция Пирсона
- Корреляция Спирмена



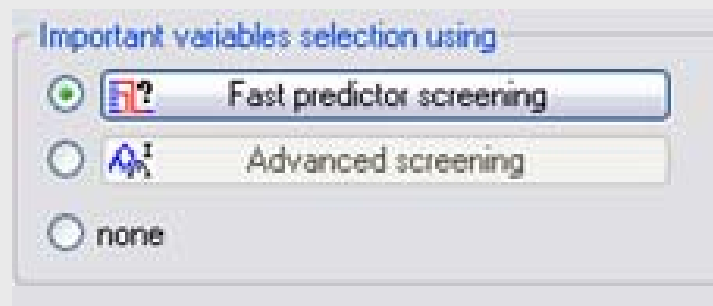
The screenshot shows a software dialog box with the following elements:

- Method** section with three radio button options:
 - Correlation coefficient
 - Spearman Rank Correlation R
 - None
- Criterion value** section with a text input field containing "0,70" and a spin button.

- Корреляция Спирмена – аналог корреляции Пирсона для порядковых переменных (переменных, измеренных в порядковой шкале)

Шаг 8. Определение наилучших предикторов

- В реальных задачах мы имеем десятки предикторов, которые возможно влияют на целевую переменную
- Из этого множества предикторов надо выбрать небольшое количество, которые имеют сильную степень влияния
- Такая процедура называется процедурой понижения размерности задачи



Шаг 9. Построение модели

- Автоматическое создание конкурентных оценок моделей



- Определение оптимальной модели относительно эффективности и сложности

- Критерий оптимальности модели создается матрицей потерь

Define cost matrix none

Evaluate models

Select model(s)

ID	Name	Error rate (%) (Testing sample)
<input checked="" type="checkbox"/> 4	SVM	5,00
<input checked="" type="checkbox"/> 5	Neural network	5,00
<input checked="" type="checkbox"/> 1	C&RT	10,00
<input type="checkbox"/> 2	Random forest	10,00
<input type="checkbox"/> 3	Boosted trees	10,00

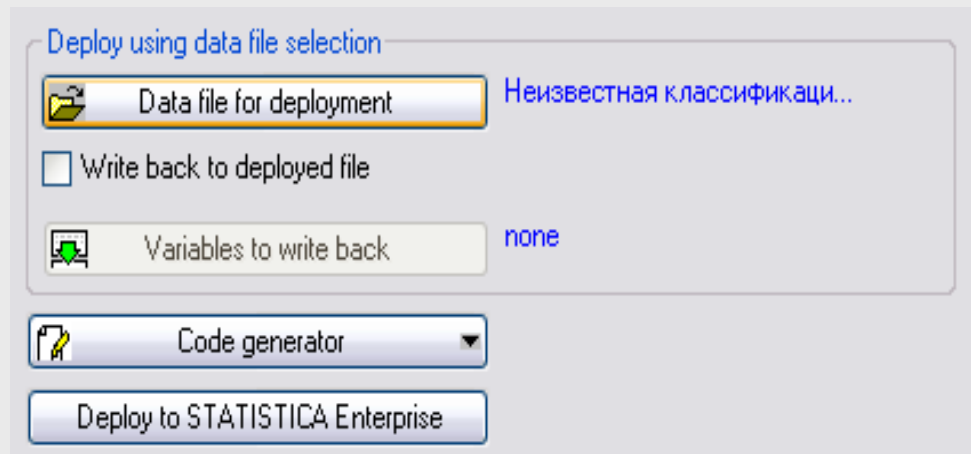
Шаг 10. Построение прогноза на НОВЫХ ДАННЫХ

- Создается таблица с новыми данными но неизвестной целевой функцией
- Структура таблицы должна соответствовать предыдущей

1 SEPALLEN	2 SEPALWID	3 PETALLEN	4 PETALWID	5 IRISTYPE
6,4	2,7	5,3	1,9	
6,8	3,0	5,5	2,1	
5,5	2,5	4,0	1,3	
4,8	3,4	1,6	0,2	
4,8	3,0	1,4	0,1	
4,5	2,3	1,3	0,3	
5,7	2,5	5,0	2,0	
5,7	3,8	1,7	0,3	
5,1	3,8	1,5	0,3	
5,5	2,3	4,0	1,3	
6,6	3,0	4,4	1,4	
6,8	2,8	4,8	1,4	

Шаг 10. Построение прогноза на НОВЫХ ДАННЫХ

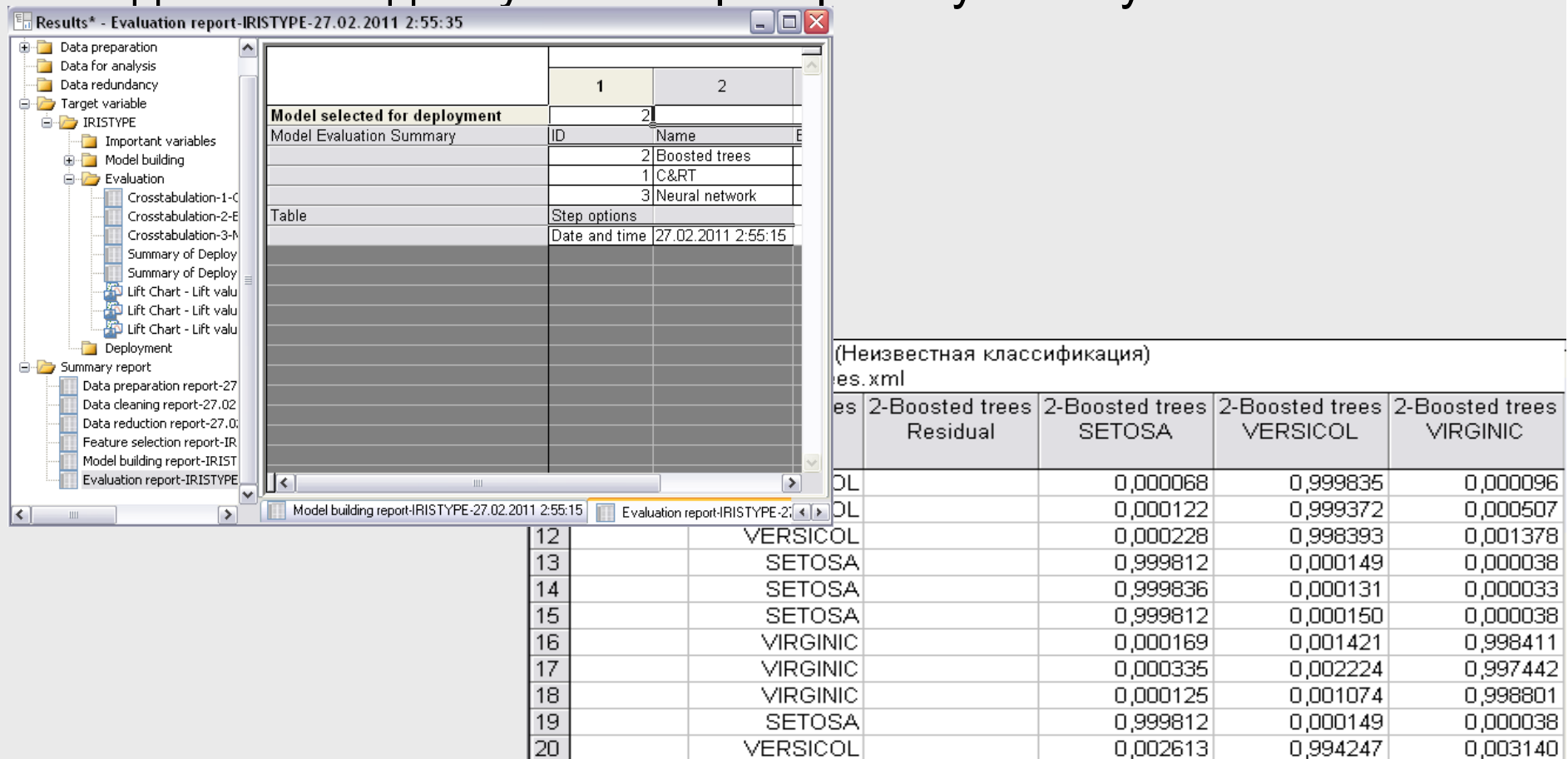
- Загрузка таблицы в Мастер Добычи Данных



- Для получения прогноза остается только нажать кнопку Next

Шаг 11. Просмотр результатов

- Прогноз на новых значениях, как и все результаты анализа выводятся в созданную Мастером рабочую книгу



The screenshot shows the 'Results* - Evaluation report-IRISTYPE-27.02.2011 2:55:35' window. The left pane shows a tree view with folders for 'Data preparation', 'Data for analysis', 'Data redundancy', 'Target variable', 'IRISTYPE', 'Important variables', 'Model building', 'Evaluation', and 'Deployment'. The main pane displays a table with columns 1 and 2, and rows for 'Model selected for deployment', 'Model Evaluation Summary', and 'Table'. The 'Table' row shows 'Step options' and 'Date and time 27.02.2011 2:55:15'. The right pane shows a detailed prediction table for '(Неизвестная классификация) res.xml'.

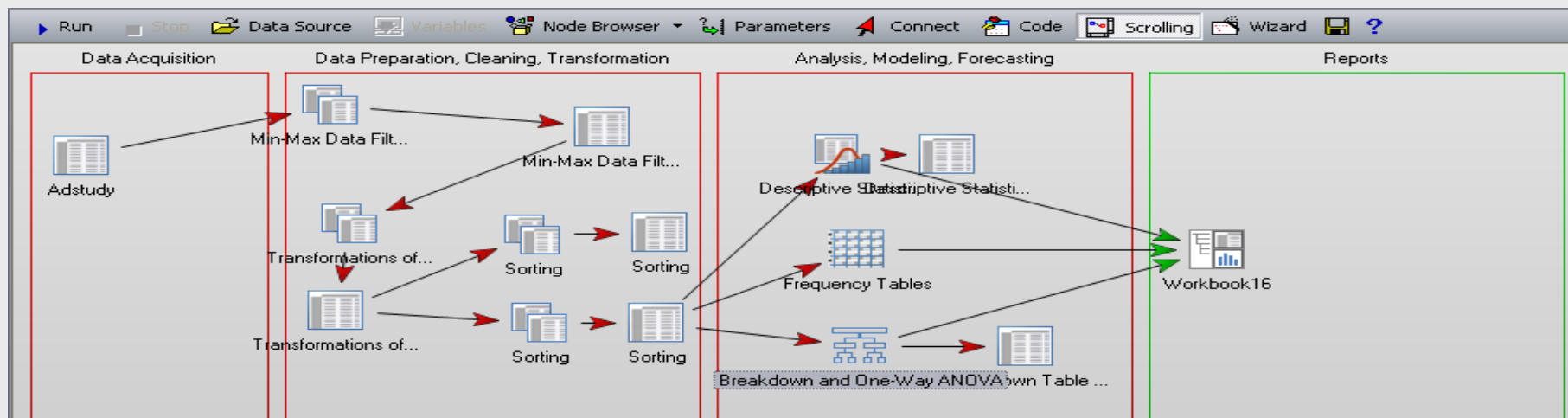
es	2-Boosted trees Residual	2-Boosted trees SETOSA	2-Boosted trees VERSICOL	2-Boosted trees VIRGINIC
DL		0,000068	0,999835	0,000096
DL		0,000122	0,999372	0,000507
12		0,000228	0,998393	0,001378
13		0,999812	0,000149	0,000038
14		0,999836	0,000131	0,000033
15		0,999812	0,000150	0,000038
16		0,000169	0,001421	0,998411
17		0,000335	0,002224	0,997442
18		0,000125	0,001074	0,998801
19		0,999812	0,000149	0,000038
20		0,002613	0,994247	0,003140

Data Miner Workspaces

Data Miner Workspaces является еще одним способом взаимодействия с Data Miner

Рабочее пространство STATISTICA Data Miner - универсальное средство для создания готовых проектов

Реализация графически-ориентированного подхода для проведения анализа данных



Преимущества Data Miner Workspaces по сравнению с Мастером

- Абсолютный набор функций преобразования данных, чистки и анализа данных *STATISTICA*
- Вы можете использовать некоторые функции *STATISTICA*, которые недоступны в обычном режиме работы с программой

Data Miner Workspaces

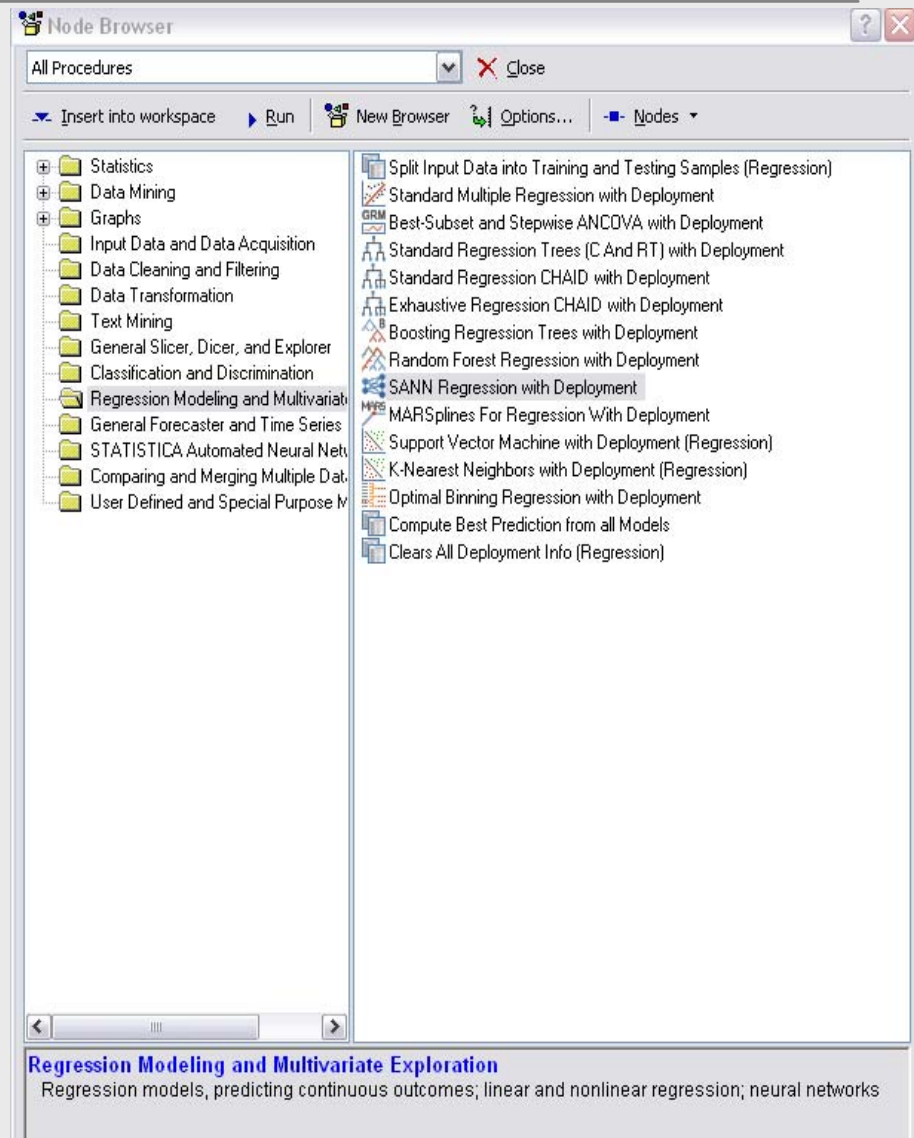
Рабочее пространство STATISTICA Data Miner состоит из четырех основных частей:

- Data Acquisition - Сбор данных
- Data Preparation, Cleaning, Transformation - Подготовка, преобразования и очистка данных
- Data Analysis, Modeling, Classification, Forecasting - Анализ данных, моделирование, классификация, прогнозирование
- Reports - Результаты

Data Miner Workspaces

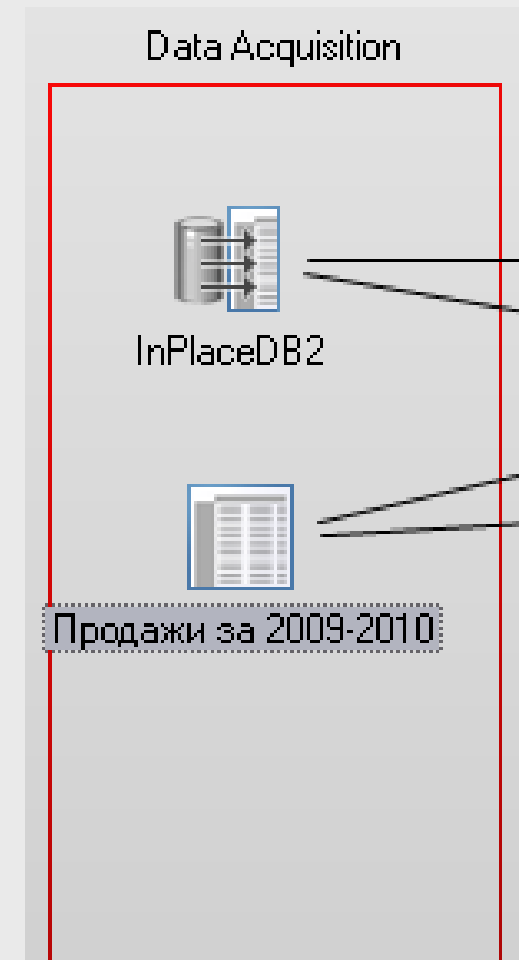
Ядром *STATISTICA* Data Miner является браузер процедур Data Mining.

Браузер содержит более 300 процедур, оптимизированных под задачи Data Mining, и средств связи между ними и управления потоками данных, позволяющих конструировать собственные методы.



Сбор данных

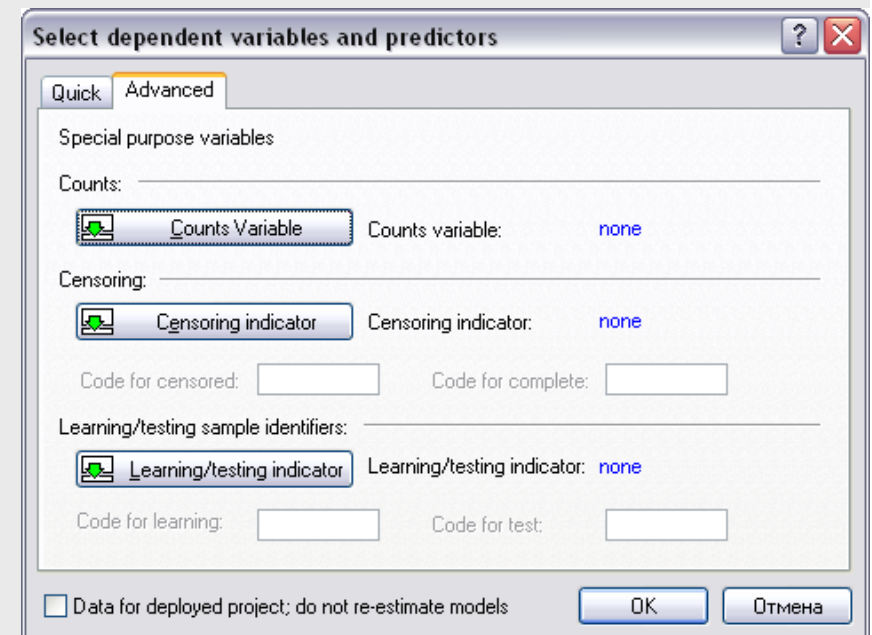
- Таблица данных *STATISTICA*
- Соединение с БД, поддерживающими *ODBC* или *OLEDB* через интерфейс **In-Place Database Interface**



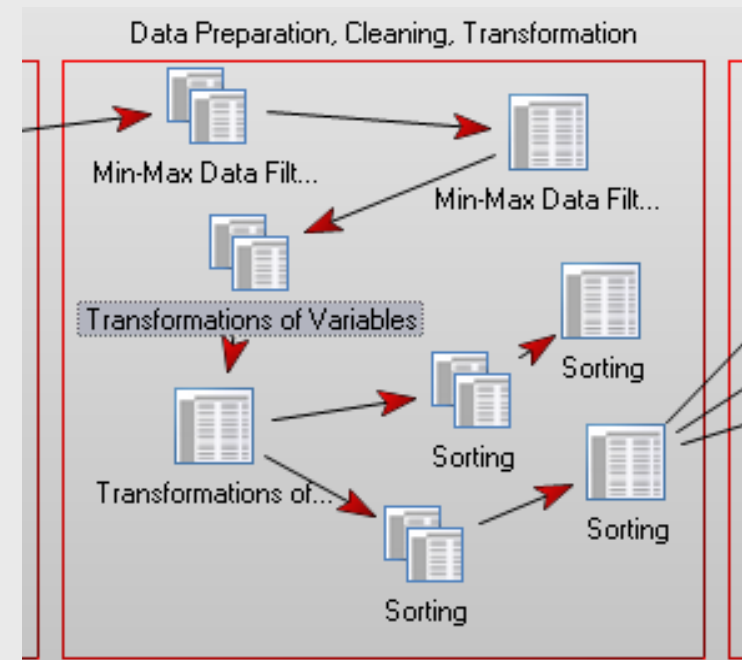
Сбор данных

- Выбор переменных:
 - целевые (непрерывные, категориальные)
 - входные (непрерывные, категориальные)
 - специальные (цензурирующие, обучающие/контрольные и тд)

- Условия выбора наблюдений и веса



- Узлы данного этапа Добычи данных могут иметь «на входе» один и более источник данных
- За выполнением узла может следовать как переход на стадию анализа, так и переход на следующий узел фильтрации



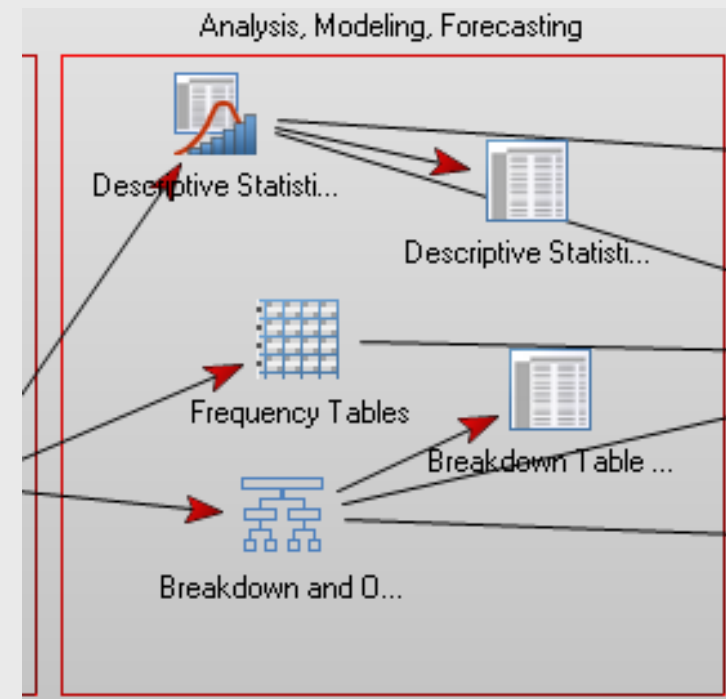
- Преобразование данных:
 - Стандартизация
 - Сортировка
 - Сдвиг
 - Преобразования Бокса-Кокса
 - Транспонирование
 - Кодирование данных
 - Удаление переменных и наблюдений
 - Ранжирование и тд

- Чистка и фильтрация данных:
 - Создание подвыборок
 - Обработка пропущенных данных
 - Отсеивание признаков
 - Фильтрация данных и тд

- Min-Max Data Filtering
- Random Sample Filtering
- Systematic Random Sampling
- Stratified Random Sampling
- Process Missing Data
- Feature Selection and Variable Screening
- User-Defined Subset
- Variable Screening Template
- Analyze Var Lists & Determine Categorical Vars
- Separate Variable Lists
- Filter Duplicate Cases
- Filter Sparse Data
- Process Invariant Variables
- MD Imputation
- Replace Missing Data with Means

Анализ данных, моделирование, классификация, прогнозирование

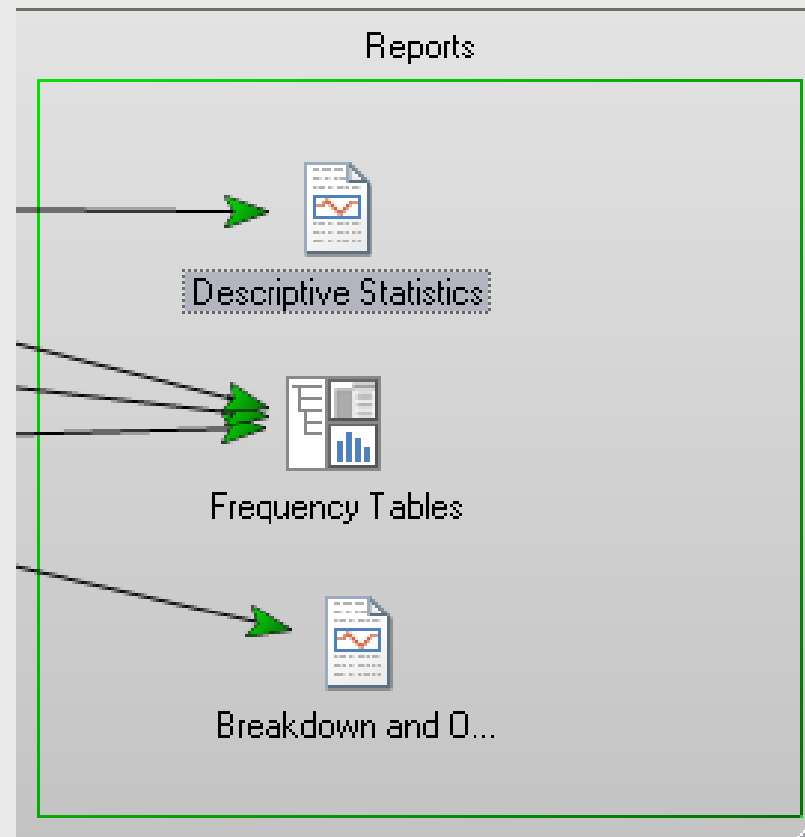
- Узлы данной категории содержат весь диапазон анализов, имеющихся в системе STATISTICA
- Любой необходимый анализ можно выбрать в диспетчере узлов
- Редактирование проекта с помощью SVB
- В качестве результата получаем рабочую книгу, таблицу данных, отчет, графики



Результаты

Результаты работы в Data Mining можно получать в формате:

- Отчет
- Рабочая книга



Основные классы анализа

- Существует ряд шаблонов Data Miner Workspaces для:
 - Прогнозирование
 - Классификации
 - Линейные и нелинейные модели регрессионные модели
 - Разведочный анализ
 - Нейросетевой анализ

- Средства анализа *STATISTICA* Data Miner можно классифицировать на пять основных классов

Основные классы анализа

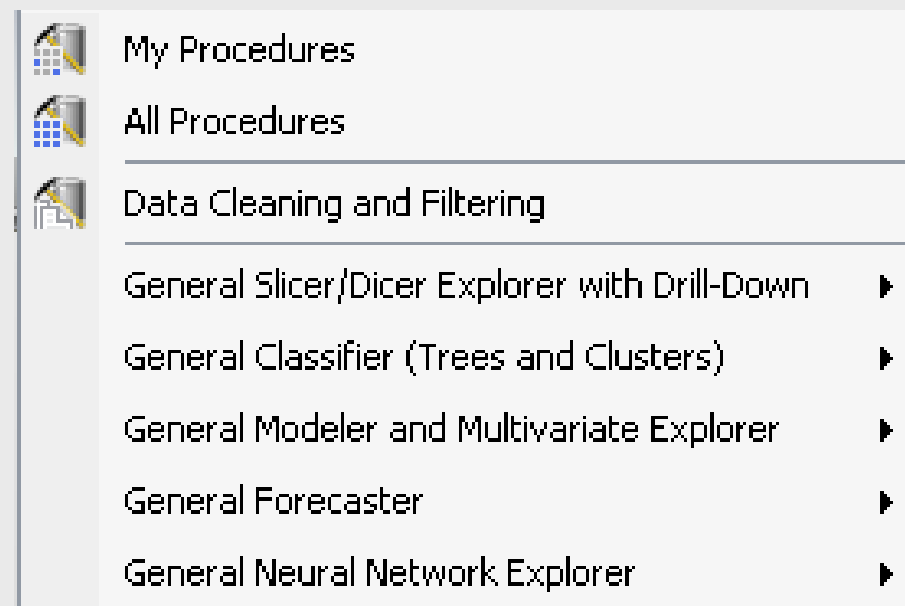
- General Slicer/Dicer and Drill-Down Explorer - Разметка/Разбиение и Углубленный анализ. Набор процедур позволяющий разбивать, группировать переменные, вычислять описательные статистики, строить исследовательские графики и т.д.
- General Classifier - Классификация. STATISTICA Data Miner включает в себя полный пакет процедур классификации: обобщенные линейные модели, деревья классификации, регрессионные деревья, кластерный анализ и т.д

Основные классы анализа

- General Modeler/Multivariate Explorer - Обобщенные линейные, нелинейные и регрессионные модели. Данный элемент содержит линейные, нелинейные, обобщенные регрессионные модели и элементы анализа деревьев классификации
- General Forecaster - Прогнозирование. Включает в себя модели АРПСС, сезонные модели АРПСС, экспоненциальное сглаживание, спектральный анализ Фурье, сезонная декомпозиция, прогнозирование при помощи нейронных сетей и т.д

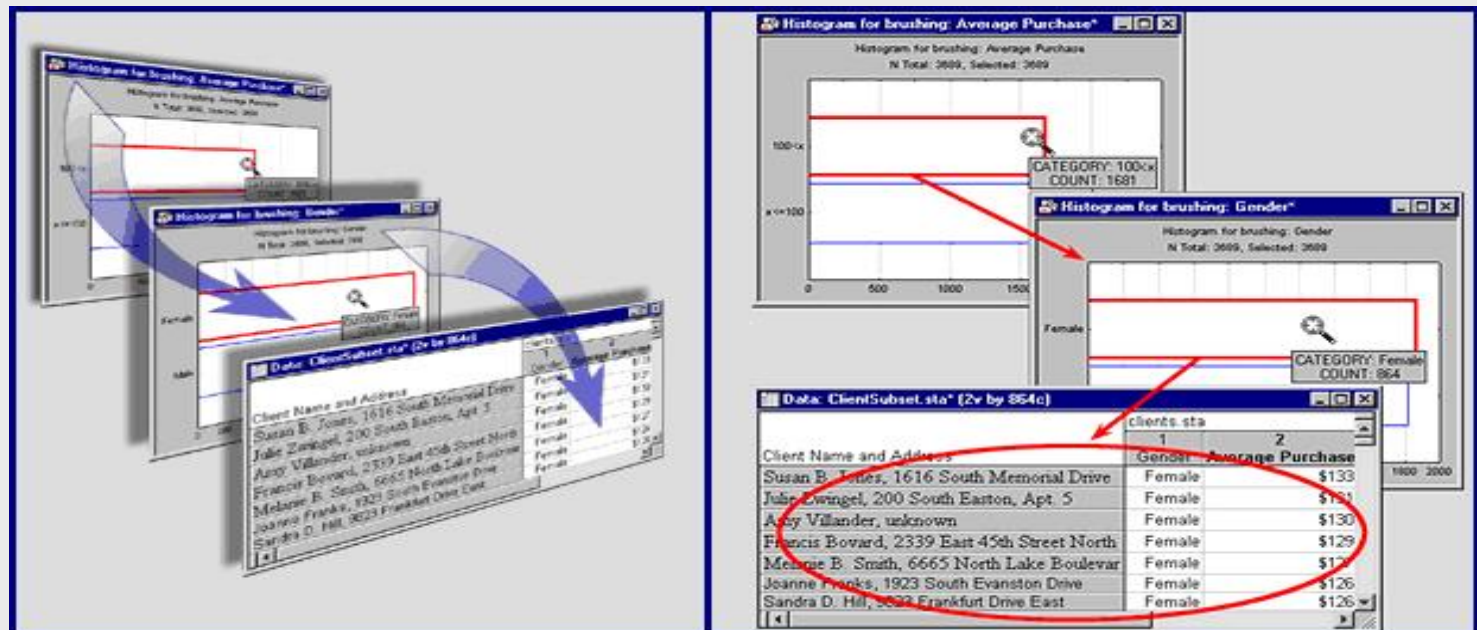
Основные классы анализа

- General Neural Networks Explorer - Нейросетевой анализ. В данной части содержится наиболее полный пакет процедур нейросетевого анализа
- Возможность создавать собственные классы и процедуры



Интерактивное бурение и расслоение

- Разведочный анализ
- Графический анализ
- Изучение связей между переменными



Интерактивное бурение и расслоение

- Описательные статистики для каждой группы данных
- Таблицы частот для каждой группы
- Графики
 - Гистограммы
 - Диаграммы размаха
 - Круговые диаграммы
- Кисть

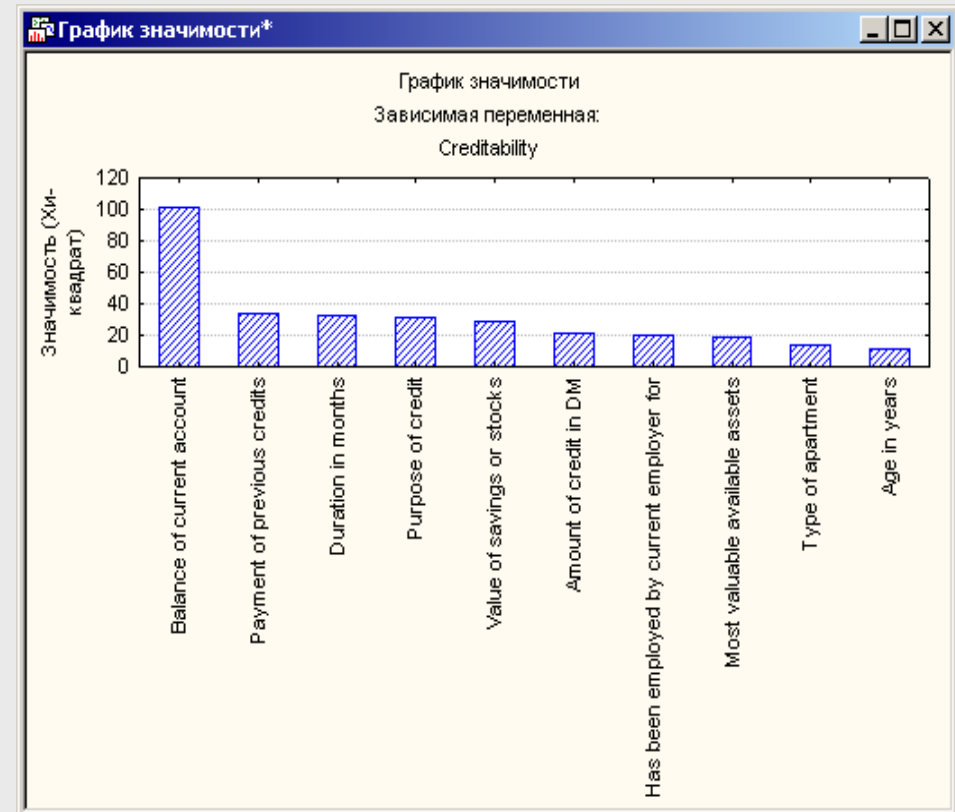
Отсеивание признаков

- Алгоритм отсеивания признаков может быть применен к проблемам регрессии (непрерывная зависимая переменная) также как и к проблемам классификации (категориальная зависимая переменная)
- Алгоритм предназначен для анализа предельно большого множества непрерывных и/или категориальных предикторов
- Алгоритм отсеивания признаков подразумевает не только поиск линейных или монотонных зависимостей между зависимой переменной и предикторами

Отсеивание признаков

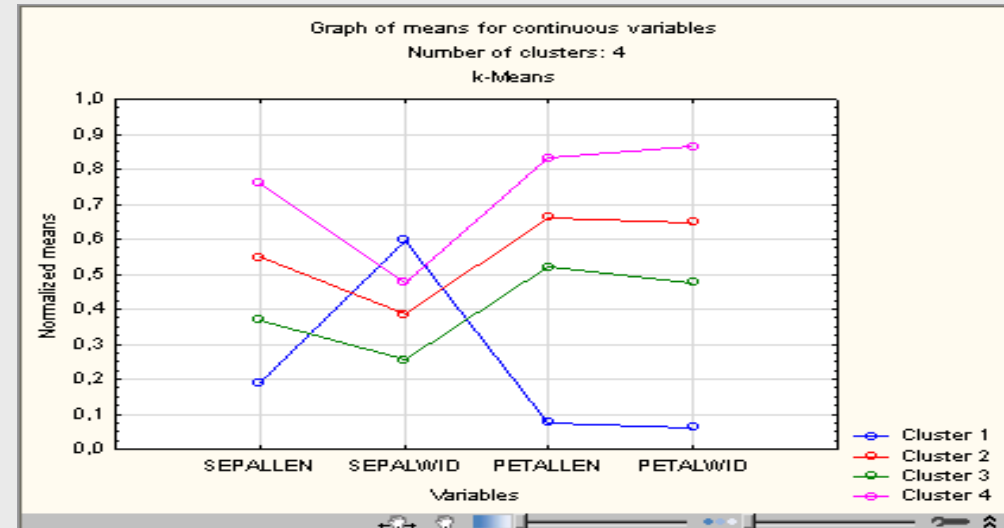
- Разные способы вариантов отбора предикторов для исследуемой модели
- Визуализация результатов

Данные: Наилучшие предикторы для категориальной зависимой пер...		
	Наилучшие предикторы для к	
	Chi-квадрат	p-значение
Balance of current account	101,3022	0,000000
Payment of previous credits	33,8076	0,000001
Duration in months	33,0700	0,000026
Purpose of credit	30,8109	0,000319
Value of savings or stocks	28,3603	0,000011
Amount of credit in DM	21,5750	0,001445
Has been employed by current employer for	20,2760	0,000440
Most valuable available assets	19,3536	0,000231
Type of apartment	13,8726	0,000972
Age in years	11,8583	0,105315

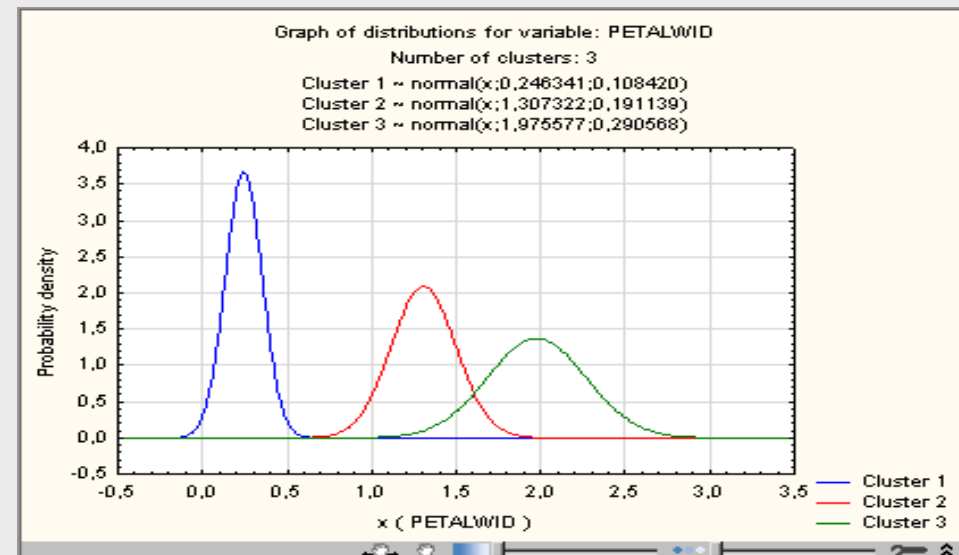


Обобщенный кластерный анализ

■ Кластеризация k-средними



■ EM – кластеризация



Обобщенный кластерный анализ

- EM алгоритм пытается приблизить наблюдаемые распределения значений, основываясь на смеси различных распределений в различных кластерах
- Реализация EM алгоритма позволяет выбирать (для непрерывных переменных) распределение: Нормальное, Логнормальное и Пуассоновское
- EM алгоритм в *STATISTICA* также может обрабатывать категориальные переменные

Обобщенный кластерный анализ

- Алгоритм k - средних может быть применен как для непрерывных, так и для категориальных переменных
- Алгоритм k - средних в модуле *Обобщенные методы кластерного анализа* использует измененную схему v -кратной кросс-проверки для определения наилучшего числа кластеров по данным
- Модуль предоставляет выбор различных мер расстояний для реализации алгоритма k – средних: евклидово, квадрат евклидово, Манхэттенское и Чебышева

Ассоциативные правила

- Цель метода – выявить отношения или связи между специфическими значениями категориальных переменных в больших базах данных
- «Покупатели, заказавшие товар A часто также заказывают B или C »
- Особенно метод полезен в том случае, когда исследователь сталкивается с огромным объемом малоизученных данных

Первичная обработка данных: Поддержка

- В первую очередь *STATISTICA* будет сканировать все переменные, чтобы определить уникальные кодовые или текстовые значения, найденные среди переменных для анализа
- Возможность того, что транзакция содержит определенное кодовое или текстовое значение называется *Поддержка*
- *Поддержка* также вычисляется при дальнейших последовательных обработках данных, как вероятность встречи двойных, тройных и т.д. кодовых или текстовых значений

Вторичная обработка данных: доверие, корреляция

- Особенностью является то, что *STATISTICA* будет вычислять условные вероятности для всех пар кодовых и текстовых значений, у которых значение поддержки больше, чем некоторый определенный минимум поддержки
- Эта условная вероятность - результат, который содержит кодовое или текстовое значение X также содержит кодовое или текстовое значение Y - называется *Доверие*
- Значение корреляции для пары кодовых или текстовых значений $\{X, Y\}$ вычисляется как поддержка этой пары, деленная на квадратный корень из величины поддержки X и Y

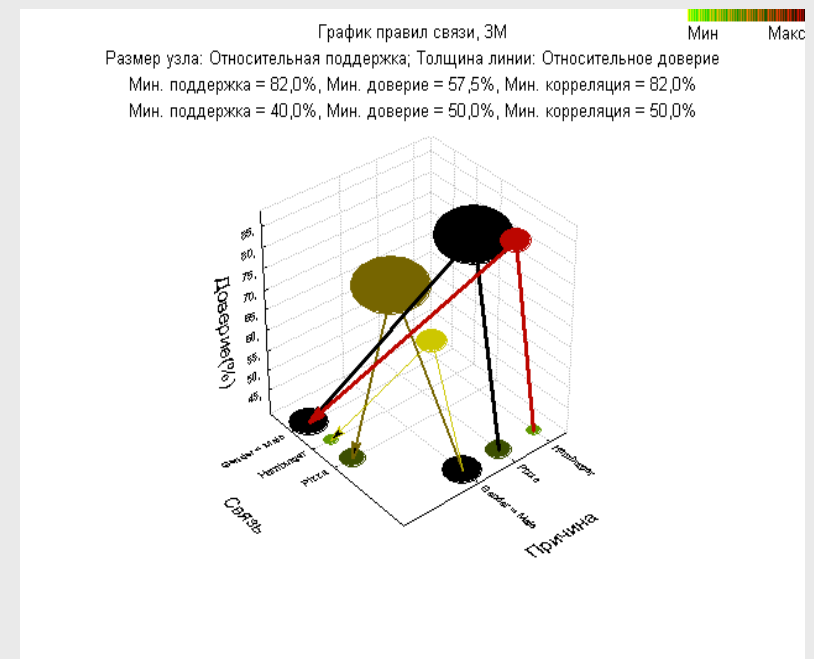
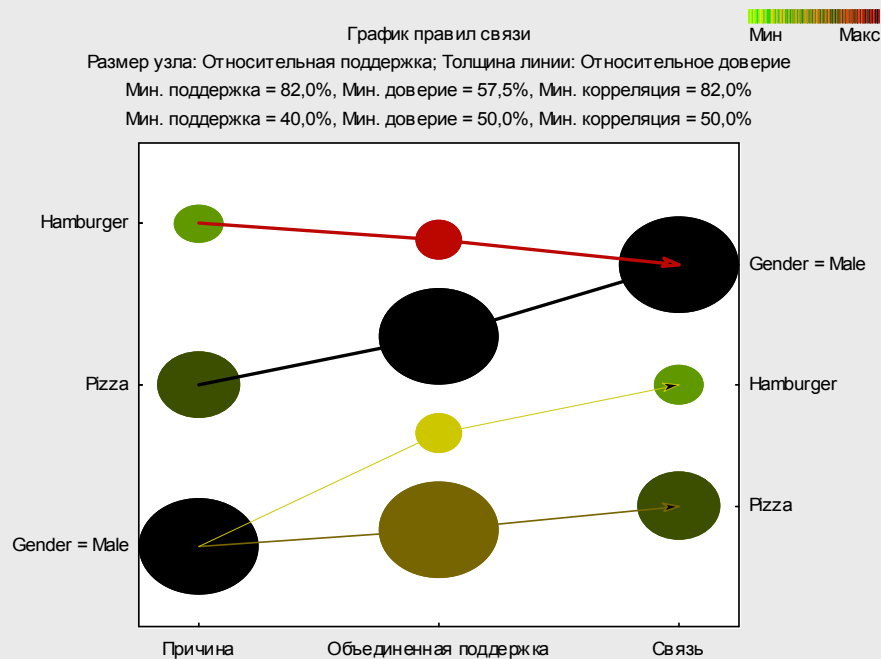
Ассоциативные правила

■ Простой и понятливый вывод результатов

Summary of association rules (OilSearch)						
Min. support = 50,0%, Min. confidence = 75,0%, Min. correlation = 75,0%						
Max. size of body = 5, Max. size of head = 5						
	Body	==>	Head	Support(%)	Confidence(%)	Correlation(%)
4		Признак 1 == 1, ==>	Наличие нефти == да	50,00000	89,2857	87,73280
19		Признак 2 == 1, ==>	Наличие нефти == да	58,00000	84,0580	91,68313
29		Признак 3 == 1, ==>	Наличие нефти == да	58,00000	77,3333	87,93937
36		Признак 4 == 1, ==>	Наличие нефти == да	58,00000	75,3247	86,78979
58	Признак 1 == 1, Признак 2 == 1, ==>		Наличие нефти == да	50,00000	89,2857	87,73280
65	Признак 1 == 1, Признак 3 == 1, ==>		Наличие нефти == да	50,00000	89,2857	87,73280
72	Признак 1 == 1, Признак 4 == 1, ==>		Наличие нефти == да	50,00000	89,2857	87,73280
86	Признак 2 == 1, Признак 3 == 1, ==>		Наличие нефти == да	58,00000	84,0580	91,68313
91	Признак 2 == 1, Признак 4 == 1, ==>		Наличие нефти == да	58,00000	84,0580	91,68313
102	Признак 3 == 1, Признак 4 == 1, ==>		Наличие нефти == да	58,00000	77,3333	87,93937
119	Признак 1 == 1, Признак 2 == 1, Признак 3 == 1, ==>		Наличие нефти == да	50,00000	89,2857	87,73280
122	Признак 1 == 1, Признак 2 == 1, Признак 4 == 1, ==>		Наличие нефти == да	50,00000	89,2857	87,73280
128	Признак 1 == 1, Признак 3 == 1, Признак 4 == 1, ==>		Наличие нефти == да	50,00000	89,2857	87,73280

Ассоциативные правила

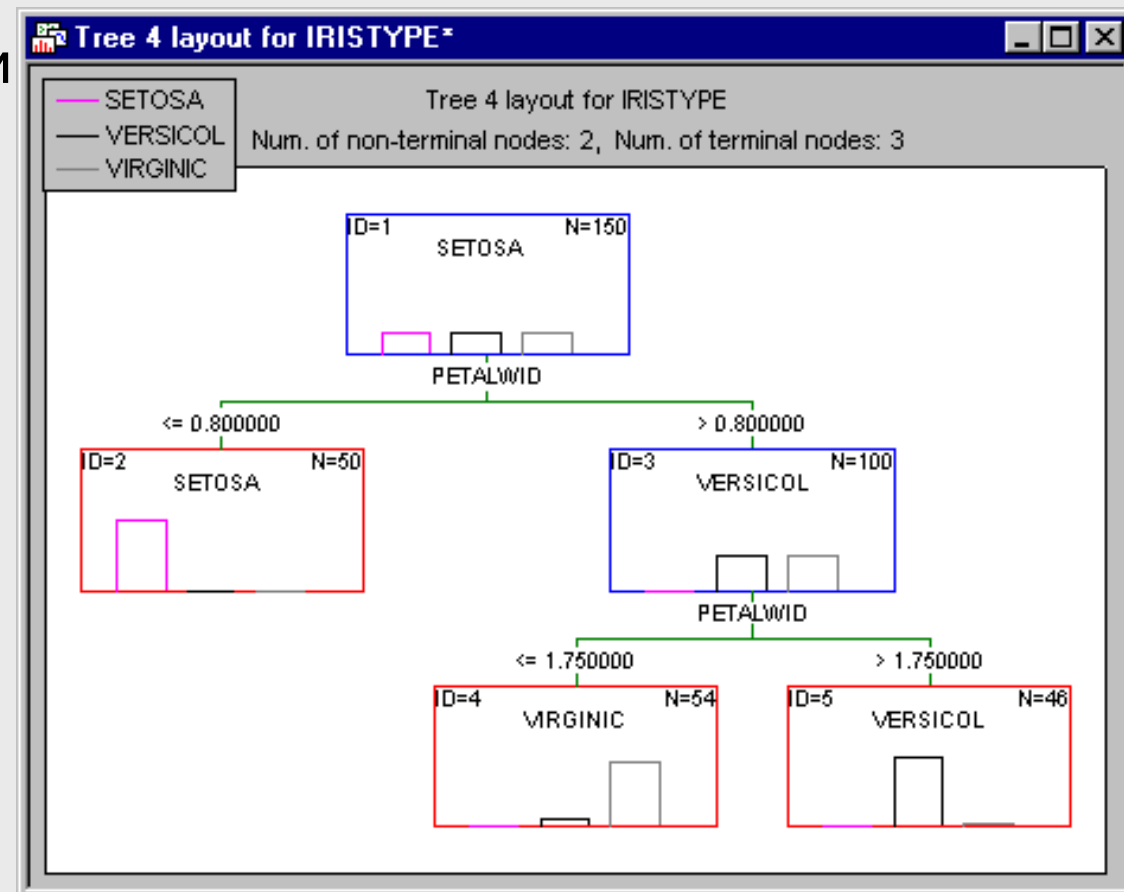
- Графическое представление связей:
 - Двухмерные графики
 - Трехмерные графики



Деревья классификации регрессии

- Решение задач классификации (категориальный отклик)

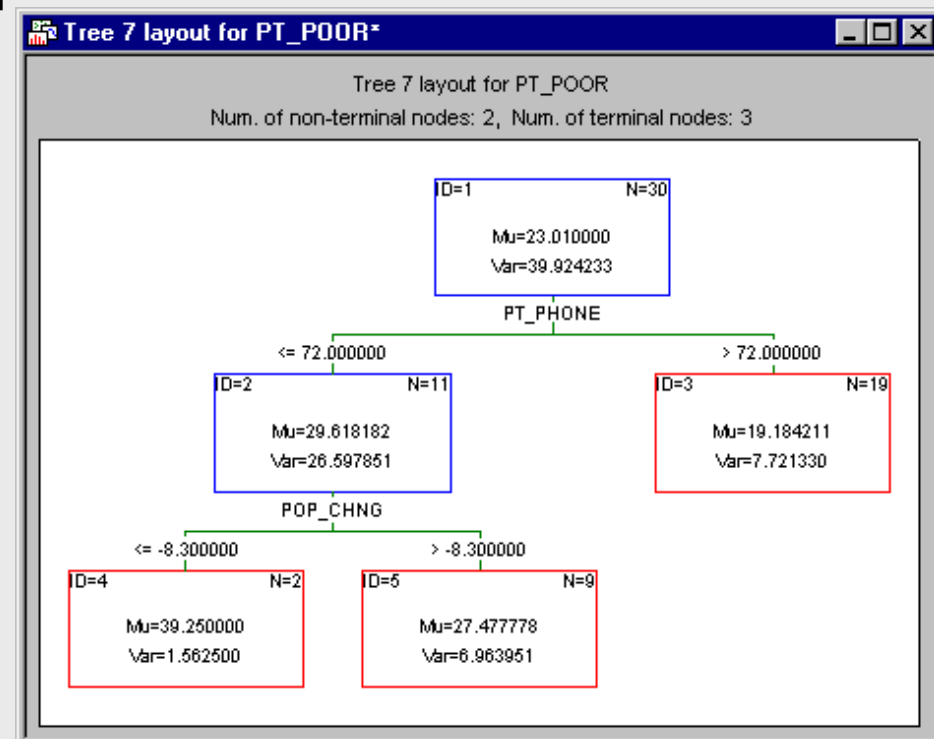
Задача о классификации
цветов ириса



Деревья классификации регрессии

- Решение задач регрессии (непрерывный отклик)

Целью анализа является нахождение предикторов, влияющих на уровень бедности в округе и как можно более точное предсказание этого уровня



Преимущества

■ Простые результаты

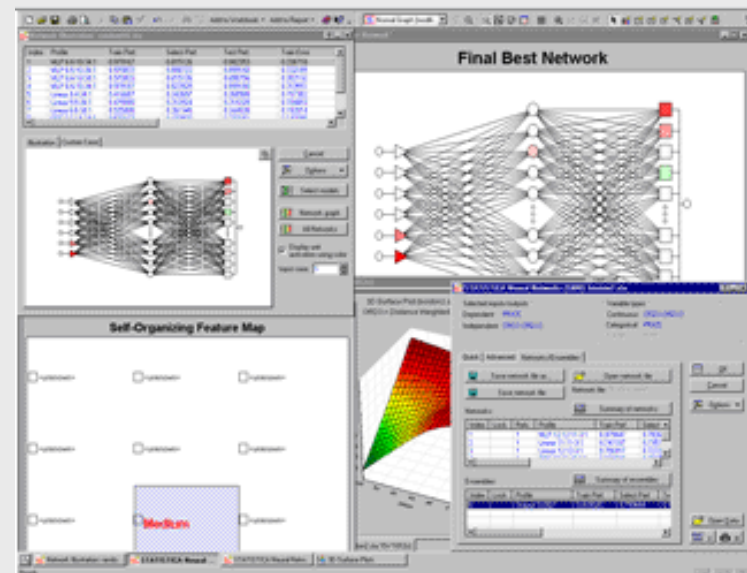
В большинстве случаев, интерпретация полученных деревьев очень проста. Это является положительным свойством не только при классификации новых наблюдений, но и при интерпретации модели в целом

■ Деревья по своей природе непараметричны и нелинейны

Результат работы данного модуля обычно представляется в виде набора логических условий

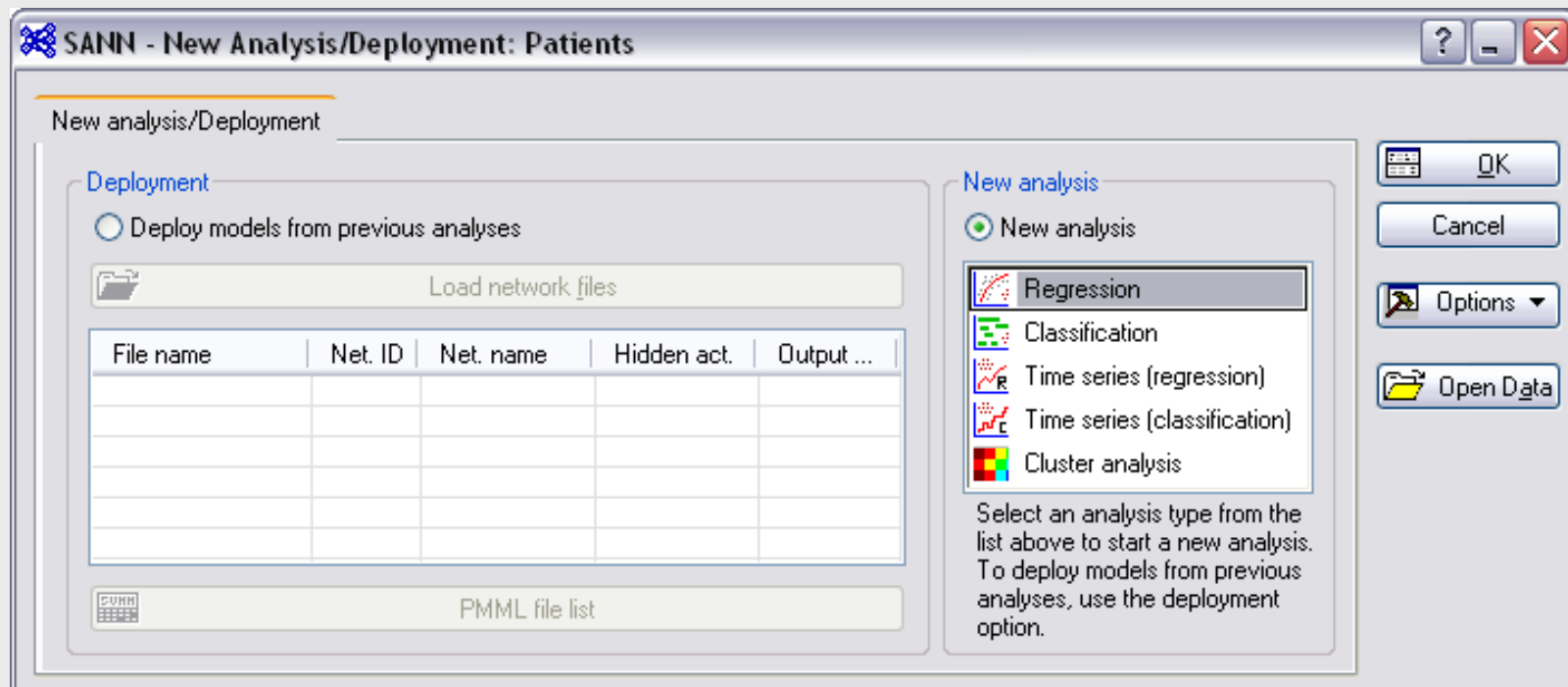
Нейронные сети

- *STATISTICA* Neural Networks является богатой, мощной и чрезвычайно быстрой средой анализа нейросетевых моделей, соответствует самым современным технологиям и показывает наилучшие рабочие характеристики
- Уникальные возможности позволяют использовать систему не только экспертам, но и новичкам в области нейросетевых вычислений



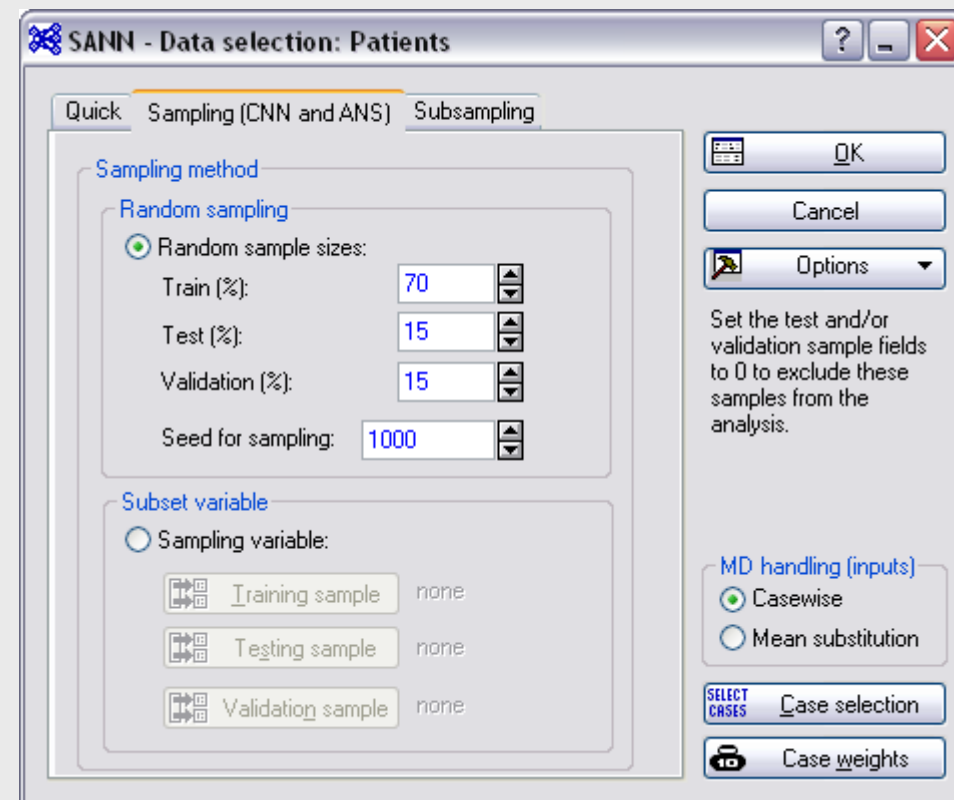
Нейронные сети

- Исключительная простота в использовании
- Аналитическая мощьность
- Система проведет пользователя через все этапы создания различных нейронных сетей и выберет наилучшую



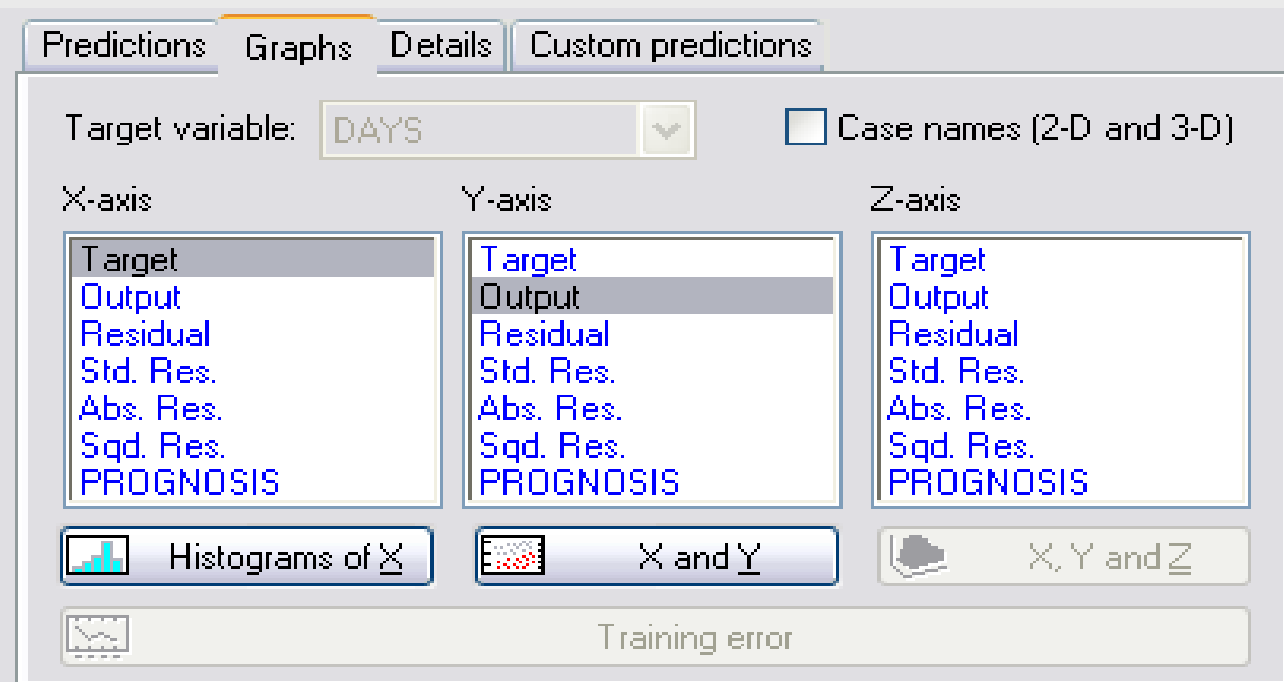
Нейронные сети

- Самые современные, оптимизированные и мощные алгоритмы обучения сети
- Полный контроль над всеми параметрами, влияющими на качество сети, такими как функции активации и ошибок, сложность сети



Нейронные сети

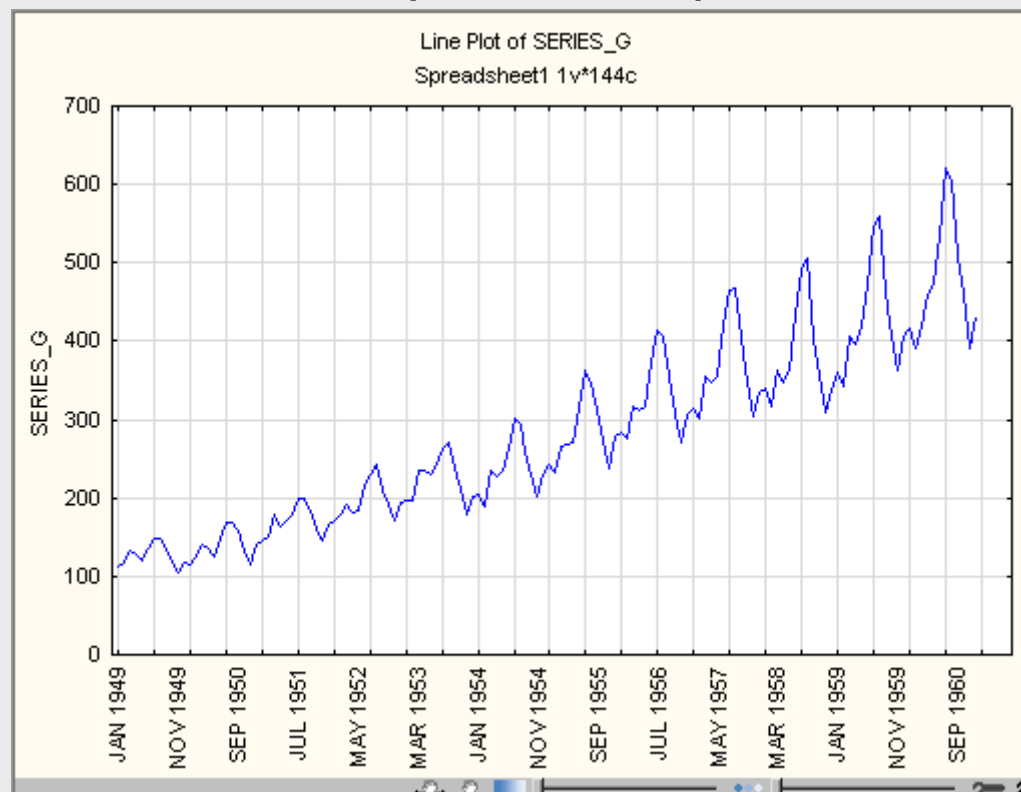
- Полная интеграция с системой STATISTICA
- Все результаты, графики, отчеты и т.д. могут быть в дальнейшем модифицированы с помощью мощных графических и аналитических инструментов STATISTICA



Пример прогнозирования в SANN

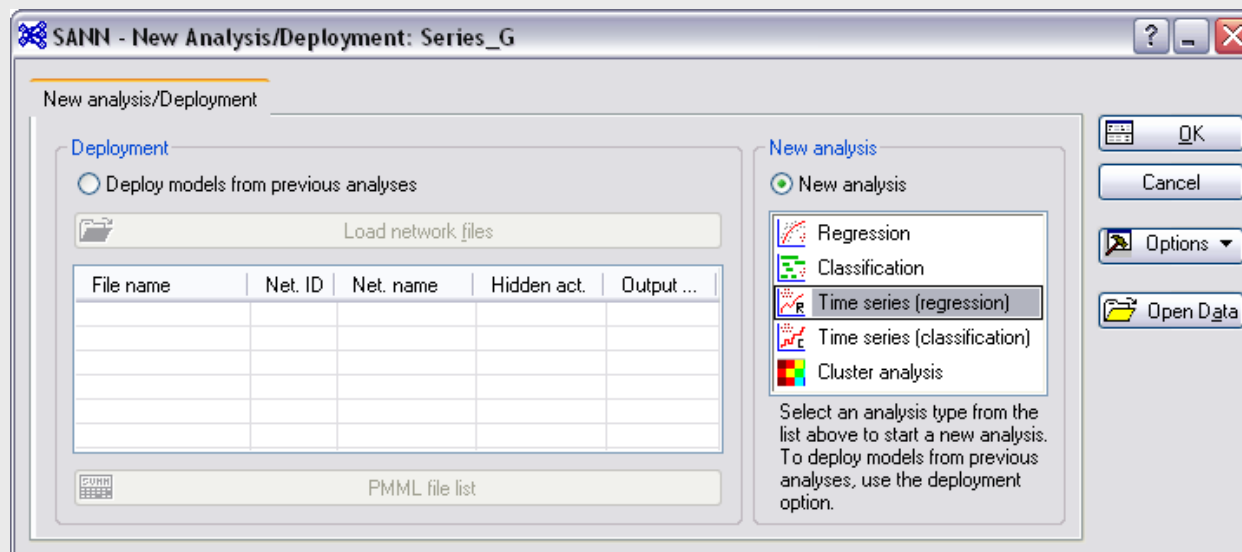
- Рассмотрим задачу прогнозирования авиаперевозок в США
- Файл данных представляет собой временной ряд

	1
	SERIES_G
JAN 1952	171
FEB 1952	180
MAR 1952	193
APR 1952	181
MAY 1952	183
JUN 1952	218
JUL 1952	230
AUG 1952	242
SEP 1952	209
OCT 1952	191
NOV 1952	172
DEC 1952	194
JAN 1953	196
FEB 1953	196
MAR 1953	236
APR 1953	235
MAY 1953	229
JUN 1953	243
JUL 1953	264
AUG 1953	272
SEP 1953	237
OCT 1953	211
NOV 1953	180
DEC 1953	201
JAN 1954	224



Шаг 1. Выбор модели, выбор переменных

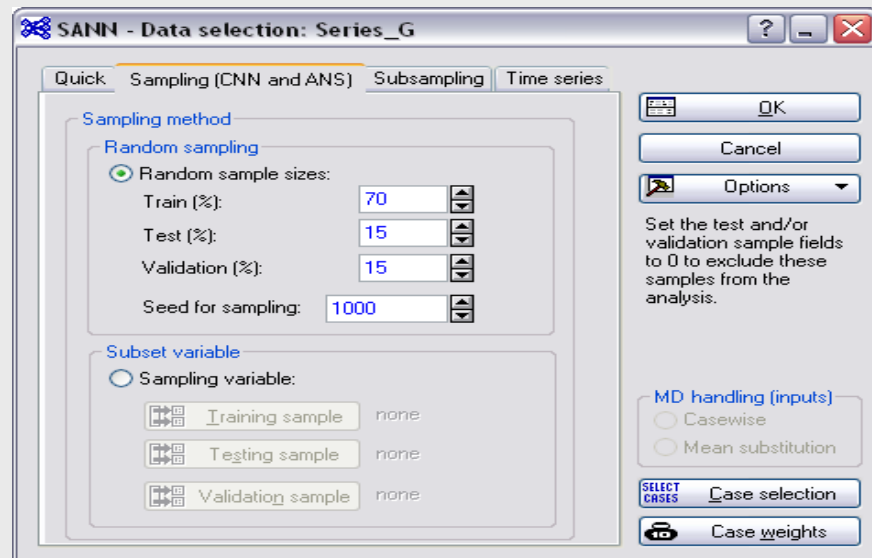
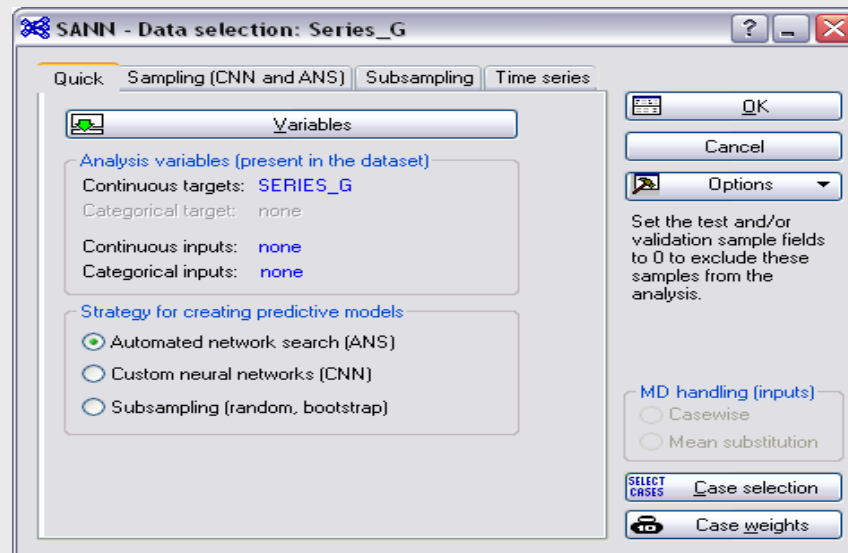
- Дружелюбный интерфейс программы, позволяющий с первого шага определить вид анализа



- Удобный диалог для выбора переменных, создания выборок для обучения, тестирования сети и кросс-проверки результатов

Шаг 2. Создание нейронной сети

- Определение стратегии создания предсказывающей модели:
 - Определение способа построения сети
 - Создание множеств для обучения, тестирования и кросс проверки
 - Установка доп. условий для сети (например, при анализе временных рядов, значение периода)



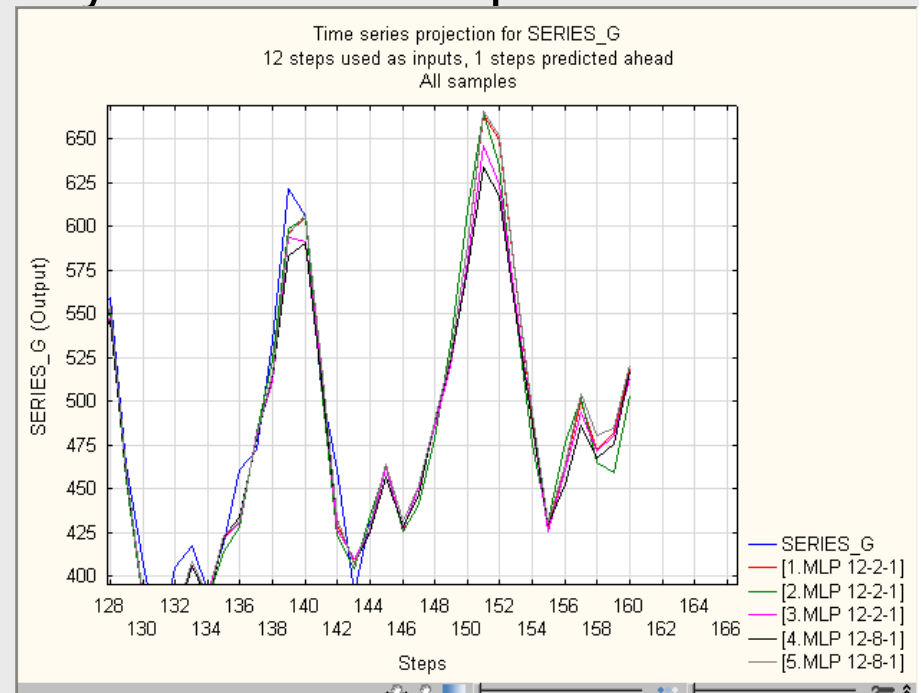
Шаг 2. Создание нейронной сети

- Определение типа нейронной сети:
 - Многослойный персептрон
 - Радиальная базисная функция
- Определение кол-во скрытых слоев сети
- Определение функций активаций, ошибки
- Установка кол-во сетей, который будут использоваться для обучения модели

Шаг 3. Просмотр результатов

- В ходе обучения и проверки качества модели на кросс-проверочном множестве *STATISTICA* определит пять лучших нейронных сетей

- Модуль позволяет посмотреть результаты и построить прогноз как для всех построенных сетей одновременно, так и для одной конкретной сети, предварительно выбрав ее



MAP - сплайны

- Модуль *Многомерные Адаптивные Сплайны (MAP - сплайны)* в *STATISTICA* - это обобщение методов, предложенных Фридманом (1991) для решения регрессионных задач и задач
- Модуль *MAP - сплайны* предназначен для обработки как категориальных, так и непрерывных переменных вне зависимости от того, являются ли они предикторами или переменными отклика
- *MAP - сплайны* - непараметрическая процедура, в работе которой не используется никаких предположений об общем виде функциональных связей между зависимыми и независимыми переменными

MAR - сплайновая модель

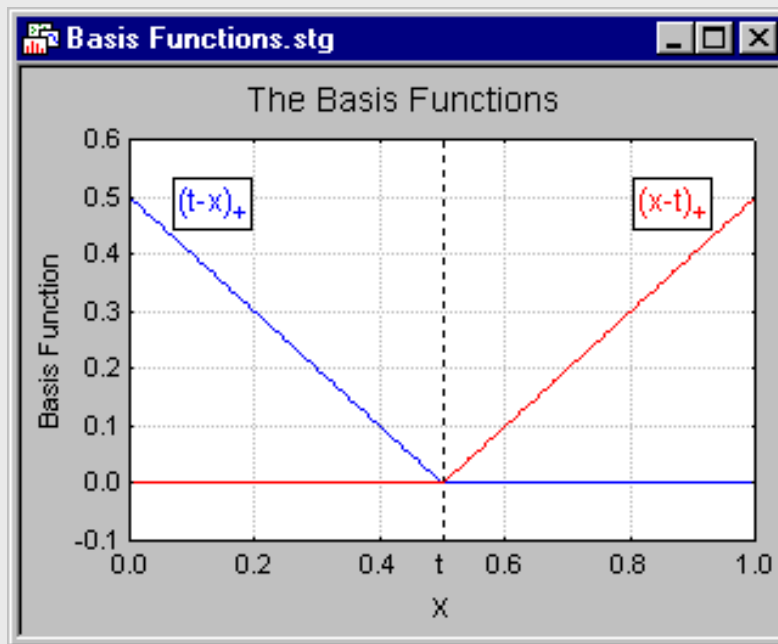
- Общее уравнение *MAR* - сплайновой модели:

$$y = f(X) = \beta_0 + \sum_{m=1}^M \beta_m k_m(X)$$

- Основной принцип работы модели состоит в выборе нужной взвешенной суммы базисных функций из общего набора базисных функций, покрывающих все значения каждого предиктора

Базисные функции

- *MAP* - сплайны используют двусторонние усеченные формы функций в качестве базисных функций для линейного или нелинейного расширения, которое приближает связи между откликом и предиктором



Базисные функции

- Пример двух базисных функций $(t-x)^+$ и $(x-t)^+$
- Параметр t - узловая точка базисной функции
- Узловые точки (значения параметра t) вычисляются по данным
- Символ "+" после выражений $(t-x)$ и $(x-t)$ означает, что рассматриваются только неотрицательные решения соответствующих уравнений
- В случае отрицательных решений функция оценивается нулем

Выбор модели и усечение

- В общем, непараметрические модели адаптивны и демонстрируют высокую степень гибкости, что, в конечном счете, если не принято специальных мер противодействия, может привести к перепогонке
- Чтобы справиться с этой проблемой, метод *MAP - сплайнов* опирается на технику усечения (похожа на усечения в деревьях классификации) для ограничения сложности модели посредством сокращения числа базисных функций

Выбор модели и усечение

- Свойство выбирать и усекаать базисные функции делает алгоритм *MAP - сплайнов* достаточно мощным средством отбора предикторов
- В процессе работы процедуры оставляются только те базисные функции (и те предикторные переменные), которые делают значимый вклад в прогноз

Спасибо за внимание

THE END