*Лабораторная работа 4.*

**Решение задач интеллектуального анализа данных (ИАД): прогнозирование временных рядов средствами интегрированной системы Statistica**

## *1. Цель лабораторной работы*

- изучить методы и алгоритмы прогнозирования временных рядов на примере решения конкретной задачи ИАД;
- исследовать эффективность использования различных методов прогнозирования временных рядов для решения прикладной задачи;
- ознакомиться и получить практические навыки работы с модулями интегрированной статистической системы *Statistica*, реализующими решение задачи прогнозирования временных рядов.

## *2. Задание к лабораторной работе*

Прочитайте содержательную постановку задачи ИАД для вашего варианта. В табл. 1. представлены исходные данные для решения поставленной задачи.

2.1. Подготовьте исходные данные для проведения интеллектуального анализа в системе *Statistica*.

2.2. Постройте линейный график временного ряда. На основе визуального анализа графика сделайте предварительные выводы о структуре временного ряда:

- наличие тренда; характер основной тенденции (монотонность; существование вертикальных и/или горизонтальных асимптот; рост (спад) уровней ряда с течением времени); тип функции тренда (линейная, нелинейная);
- наличие сезонной составляющей и характер сезонной составляющей (периодичность; амплитуда колебаний; постоянство (изменчивость) амплитуды колебаний с течением времени).

**Метод последовательной идентификации составляющих ВР**

2.3. Определите структурную модель тренда временного ряда с помощью метода характеристик прироста:

- сгладьте временной ряд, используя пятимесячную скользящую среднюю;
- определите средние приросты;
- определите производные характеристики прироста: $\bar{u}_t$; $\bar{u}_t^{(2)}$; $\bar{u}_t / \bar{y}_t$; $\log \bar{u}_t$; $\log(\bar{u}_t / \bar{y}_t)$; $\log(\bar{u}_t / \bar{y}_t^{\,2})$;
- постройте линейные графики производных характеристик прироста.

На основе анализа характеристик прироста определите две наиболее вероятные структурные модели тренда.

2.4. Идентифицируйте параметры выбранных структурных моделей тренда. Рассчитайте характеристики точности прогнозных моделей, заполните табл. 2.

2.5. По результатам расчетов (табл. 2) сделайте окончательный вывод относительно вида модели тренда. Постройте график исходного временного ряда с наложенной прогнозной моделью тренда.

2.6. Определите структурную модель сезонной составляющей ряда:

- постройте и проанализируйте периодограмму временного ряда;
- постройте структуру периодической гармонической функции.

2.7. Идентифицируйте параметры сезонной составляющей ряда. Рассчитайте характеристики точности прогнозной модели, содержащей тренд и сезонную составляющую, заполните табл. 2. Постройте график исходного временного ряда с наложенной прогнозной моделью.

*Таблица 2.*

*Характеристики точности прогнозных моделей*

| | **Модель 1** | **Модель 2** | **Модель 3** | …. |
|---|---|---|---|---|
| 1. Прогнозная модель | | | | |
| 2. Минимальный остаток | | | | |
| 3. Максимальный остаток | | | | |
| 4. Средняя ошибка (Mean error) | | | | |
| 5. СКО ошибки | | | | |
| 6. Средняя абсолютная ошибка (Mean absolute error) | | | | |
| 7. Сумма квадратов отклонений (Sums of squares) | | | | |
| 8. Средний квадрат отклонений (Mean square) | | | | |
| 9. Средняя ошибка в процентах (Mean percentage error) | | | | |
| 10. Средняя абсолютная ошибка в процентах (Mean abs. perc. error) | | | | |
| 11. Коэффициент детерминации | | | | |

2.8. Постройте автокорреляционную и частную автокорреляционную функции остатков прогнозной модели, построенной в п. 2.7. Сделайте вывод о

наличии (отсутствии) автокорреляции в остатках и необходимости учета авторегрессионой составляющей в прогнозной модели ряда.

2.9. Определите структуру и параметры авторегрессионой составляющей ряда (в случае необходимости). Рассчитайте характеристики точности прогнозной модели, содержащей тренд, сезонную и авторегрессионую составляющие временного ряда, заполните табл. 2.

2.10. Проанализируйте табл. 2, выберите окончательный вариант прогнозной модели (тренд + сезонность + авторегрессия), обоснуйте свой выбор.

2.11. Для выбранного варианта прогнозной модели постройте гистограмму остатков и проверьте гипотезу о согласии распределения остатков с моделью нормального распределения, постройте автокорреляционную и частную автокорреляционную функции остатков.

2.12. Сделайте выводы об адекватности построенной прогнозной модели данным наблюдения.

2.13. Дайте содержательную интерпретацию полученных результатов. Опишите составляющие прогнозной модели в терминах решаемой задачи.

**Метод экспоненциального сглаживания**

2.14. Постройте прогнозную модель ВР на основе метода экспоненциального сглаживания. В модели учтите тренд и сезонную составляющую.

2.15. Заполните табл. 2.

2.16. Для модели экспоненциального сглаживания постройте гистограмму остатков и проверьте гипотезу о согласии распределения остатков с моделью нормального распределения, постройте автокорреляционную и частную автокорреляционную функции остатков.

2.17. Сделайте выводы об адекватности построенной прогнозной модели экспоненциального сглаживания данным наблюдения.

**Построение прогнозной модели с помощью модуля Data Miner**

2.18. Постройте модели экспоненциального сглаживания, *Arima* с помощью модуля *Data Miner*. Сопоставьте по точности с моделями, полученными ранее.

**Прогнозирование**

2.19. Выберите наилучшую прогнозную модель из построенных. На основании выбранной модели осуществите прогноз временного ряда на 3 временных интервала вперед. Сделайте выводы в терминах решаемой задачи.

2.20. По результатам проведенного исследования сделайте выводы в свободной форме.

## *3. Методические указания к лабораторной работе*

***пояснения к п. 2.2.***

Для построения линейного графика выберите пункт меню *Graphs/2D Graphs/Line Plots (Variables).*

***пояснения к п. 2.3.***

Для вычисления скользящей средней временного ряда выберите пункт меню *Statistics/Advanced linear/nonlinear models/Time series/Forecasting*. В диалоговом окне *Time series analysis* задайте переменную для анализа и нажмите кнопку *OK(transformations, autocorrelations, ...)*. Далее в диалоговом окне *Transformations of variables* выберите вкладку *Smoothing*, задайте параметр скользящего среднего *N-pts mov. averg*. и подтвердите преобразование ряда *OK (Transform selected series).* В результате будет построен линейный график преобразованного временного ряда. Значения преобразованного временного ряда можно посмотреть, выбрав кнопку *Save variables*.

Вычисление скользящих приростов временного ряда выполняется в диалоговом окне *Transformations of variables*, вкладка *Difference/integrate.* Для расчета прироста первого порядка задайте параметр *Differencing: lag - 1*.

Для нахождение остальных характеристик прироста в электронной таблице с исходными данными выполните соответствующие преобразования, задавая формулу преобразования в диалоговом окне спецификации переменной.

Линейные графики производных характеристик прироста можно разместить на одном графике, воспользовавшись пунктом меню *Graph/Multiple graph layouts/Wizard.*

***пояснения к п. 2.4., 2.5., 2.7.***

Для идентификации параметров трендовой модели воспользуйтесь пунктом меню *Statistics/Advanced linear/nonlinear models/Nonlinear estimation*. Выберите вкладку *User-specified regression – custom loss function*. Далее в диалоговом окне *User-specified regression* нажмите кнопку *Function to be estimated & loss function* и задайте структуру прогнозной модели (*Estimated model*) с точностью до неизвестных параметров, функцию потерь (*Loss function*) по умолчанию (сумма квадратов отклонений наблюдаемых значений временного ряда от прогнозных значений). Все остальные параметры также выбираются по умолчанию.

Результаты идентификации модели представлены в диалоговом окне *Results*, **вкладка *Quick*:**

- *Summary: parameters estimates* – оценки параметров модели; значение суммы квадратов отклонений (*Final loss*); коэффициент детерминации*100% (*variance explained*);
- *Observed, predicted, residual vals* – наблюдаемые, прогнозные значения временного ряда и остатки;
- *Fitted 2D function and observed values* – график исходного временного ряда и построенной прогнозной модели.

***пояснения к п. 2.6.***

Для построения периодограммы временного ряда выберите пункт меню *Statistics/Advanced linear/nonlinear models/Time series/Forecasting*. В диалоговом окне *Time series analysis* задайте переменную для анализа и нажмите кнопку *Spectral (Fourier) Analysis*, далее вкладку *Quick* и кнопку *Single series Fourier analysis*. В появившемся диалоговом окне *Spectral (Fourier) Analysis Results* задайте радио кнопку *Period*, затем *Periodogram.*

***пояснения к п. 2.8., 2.9., 2.11, 2.16.***

Для построения автокорреляционной и частной автокорреляционной функций временного ряда в диалоговом окне *Time series analysis* выберите вкладку *ARIMA and autocorrelations functions*. Далее в диалоговом окне *Single series ARIMA* - вкладку *autocorrelations* и кнопки *autocorrelations* – построение автокорреляционной функции, *partial autocorrelations* – построение частной автокорреляционной функции.

Для идентификации авторегрессионой модели в диалоговом окне *Single series ARIMA* выберите вкладку *Quick*, задайте параметры модели (*ARIMA model parameters*): порядок авторегрессионой составляющей (*p – Autoregressive*) и инициируйте процесс оценки параметров модели (*OK begin parameters estimate*).

В диалоговом окне *Single series ARIMA results* представлены результаты идентификации модели:

- вкладка *Quick: Summary parameter estimates* – оценка параметров модели;
- вкладка *Distribution of residuals: Histogram* - гистограмма остаток с наложенной моделью нормального распределения;
- вкладка *Autocorrelations* – автокорреляционная и частная автокорреляционная функции остатков.

***пояснения к п. 2.14., 2.15.***

**Метод экспоненциального сглаживания в Statistica**

Для открытия стартовой панели выберите пункт меню *Statistics/Advanced Linear and Nonlinear Models/Time series and Forecasting*. Далее в диалоговом окне (см. рис 1) выберите вкладку *Exponential smoothing and forecasting*.

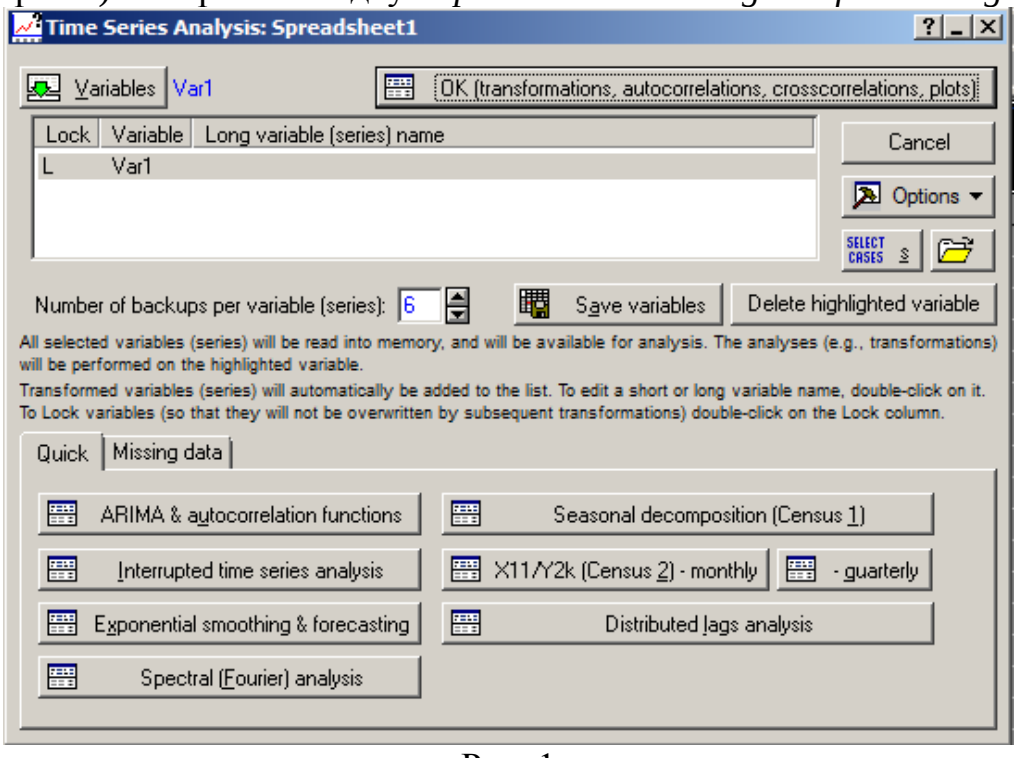


Рис. 1

На экране панель модуля:

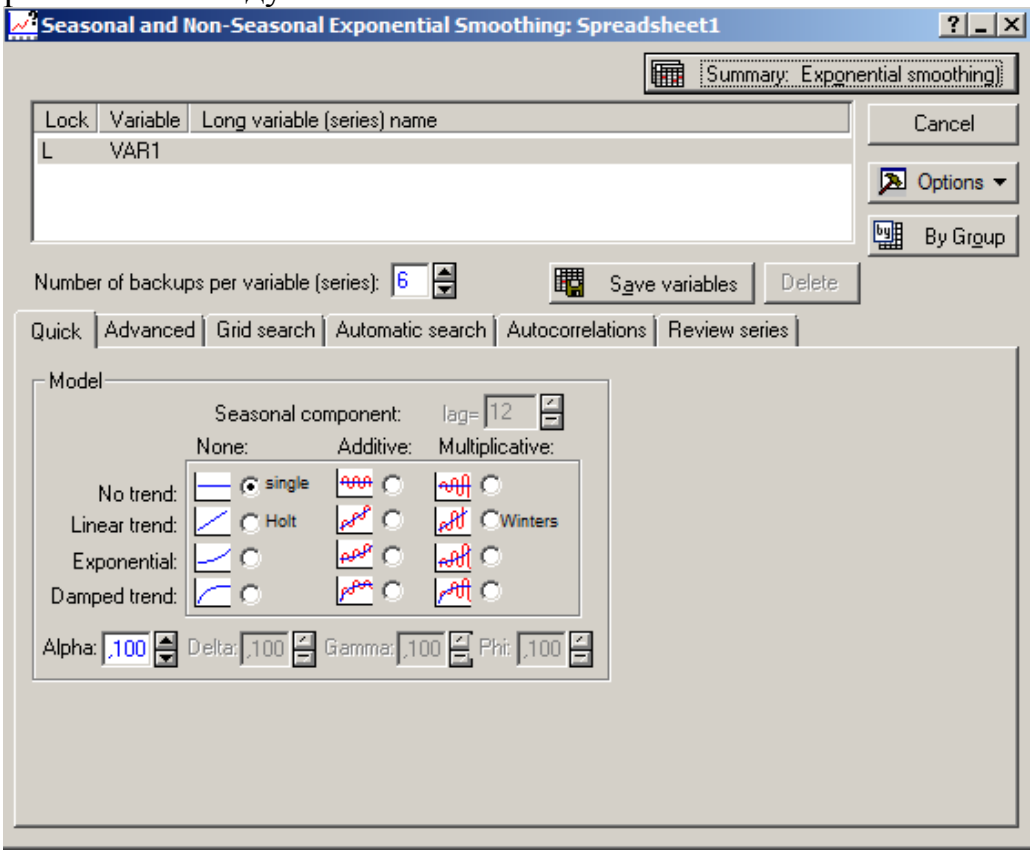


Рис. 2

Необходимо задать особенности ряда (наличие сезонной компоненты, тренд и параметры сглаживания) для определения модели экспоненциального сглаживания. Это можно сделать в следующих опциях:

- *Seasonal component (lag)* – сезонная компонента и ее лаг;
- *None* – нет составляющей;
- *Additive* – аддитивные составляющие;
- *Multiplicative* – мультипликативные составляющие;
- *No trend* – нет тренда;
- *Linear trend* – линейный тренд;
- *Exponential* – экспоненциальный тренд;
- *Damped trend* – демпфированный (затухающий тренд).

В полях *Alpha, Delta, Gamma, Phi* задаются параметры экспоненциального сглаживания. Параметр *Alpha* необходим для всех моделей экспоненциального сглаживания. Остальные параметры нужны для специальных моделей. Параметр *Delta* – сезонный сглаживающий параметр; необходим лишь в сезонных моделях. Параметры *Gamma и Phi* являются параметрами сглаживания тренда. Параметр *Gamma* используется в моделях с линейным и экспоненциальным трендом и в моделях с демпфированным трендом в рядах без сезонной составляющей. Параметр *Phi* используется в моделях с демпфированным трендом.

Во вкладке *Advanced* задаются дополнительные опции (см. рис. 3):

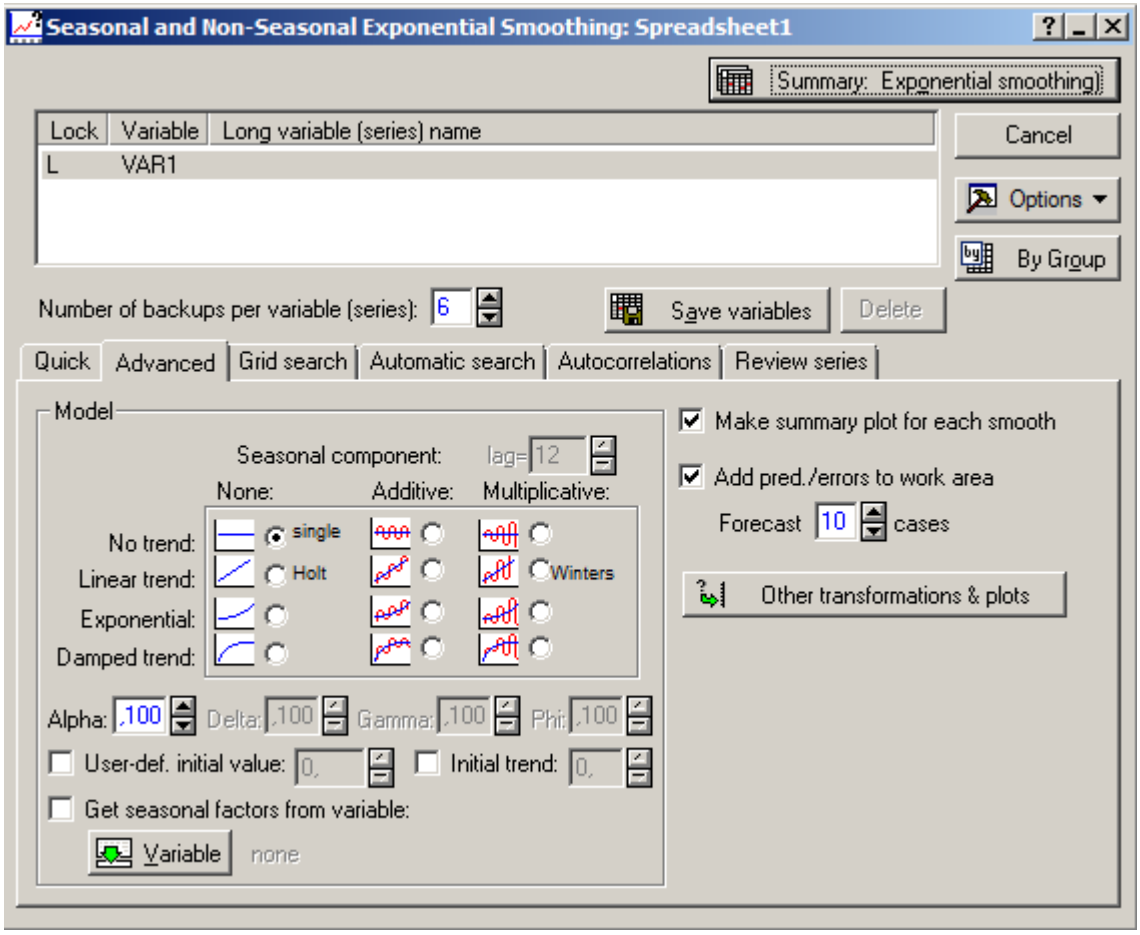


Рис. 3

- *User-def. initial value* – определяемое пользователем начальное значение. Можно задать начальное значение сглаженного ряда;
- *Initial trend* – начальный тренд. Можно задать начальное значение тренда. Если опция не используется, то начальное значение тренда оценивается;
- *Get seasonal factors from variables* – оценить сезонные факторы из данных;
- *Make summary plot for each smooth* – сделать итоговый график для каждого сглаживания;
- *Add/pred/error to work area* – добавить сглаженный ряд/остатки в рабочую область;
- *Forecast cases* – прогноз на указанное количество случаев (шагов) вперед.

Во вкладке *Grid Search* (см. рис. 4)

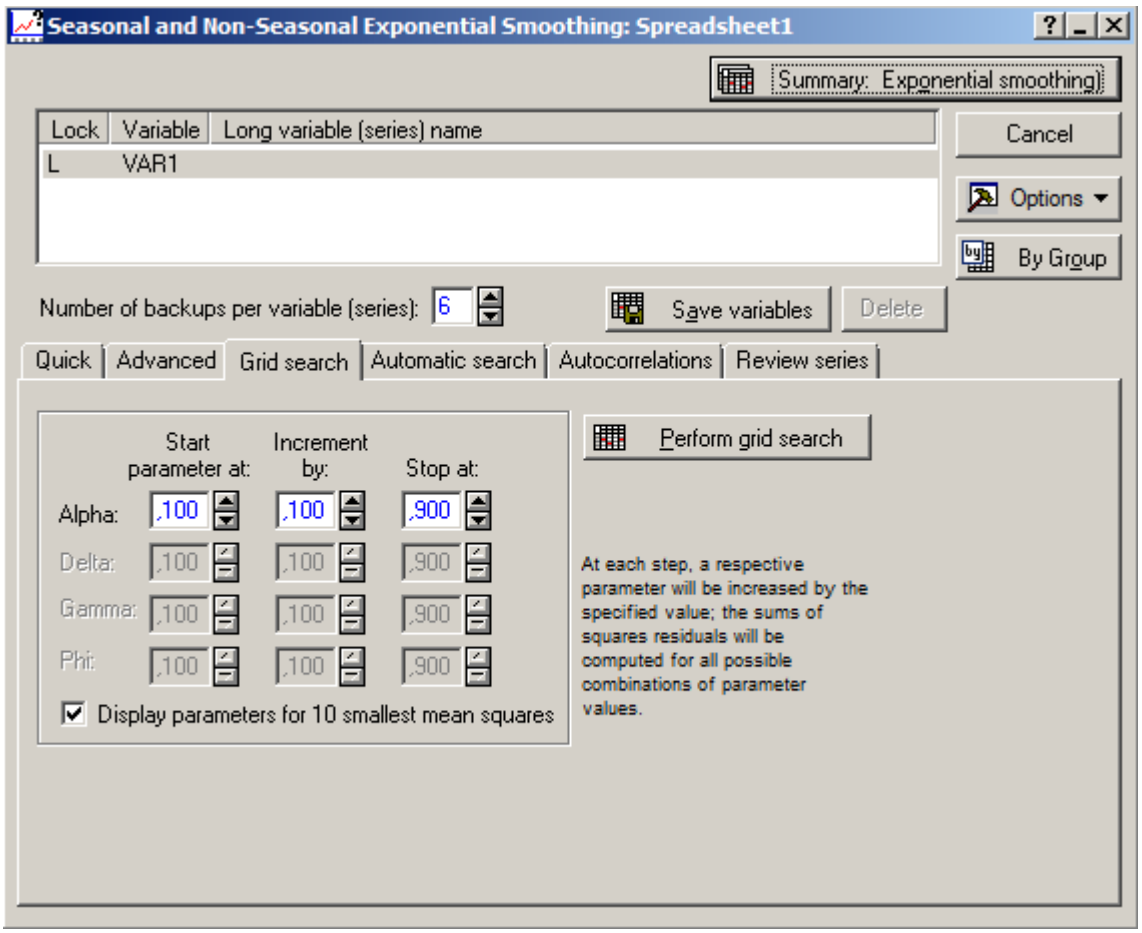


Рис. 4

реализуется поиск наилучших значений параметров экспоненциального сглаживания: *Alpha*, *Delta*, *Gamma*, *Phi*. В результате поиска выдается таблица вида рис. 5. Первая строка таблицы соответствует лучшим значениям параметров. Также в таблице приведены характеристики точности для каждой модели.

Parameter grid search (Smallest abs. errors are highlighted) (Spreadsheet1)
Model: No trend, no season ; S0=154,1
VAR1

| Model Number | Alpha | Mean Error | Mean Abs Error | Sums of Squares | Mean Squares | Mean % Error | Mean Abs % Error |
|---|---|---|---|---|---|---|---|
| 9 | 0,900000 | 0,155257 | 7,59337 | 34171,9 | 177,9787 | -4,4169 | 11,19101 |
| 8 | 0,800000 | 0,166402 | 7,61840 | 34575,7 | 180,0817 | -4,8821 | 11,64732 |
| 7 | 0,700000 | 0,181568 | 7,72891 | 35745,0 | 186,1721 | -5,4215 | 12,09546 |
| 6 | 0,600000 | 0,203027 | 8,06526 | 37879,8 | 197,2907 | -6,0755 | 12,97570 |
| 5 | 0,500000 | 0,235273 | 8,68899 | 41373,3 | 215,4857 | -6,9323 | 14,33583 |
| 4 | 0,400000 | 0,287957 | 9,58696 | 47034,3 | 244,9703 | -8,1923 | 16,32828 |
| 3 | 0,300000 | 0,384182 | 10,76457 | 56822,1 | 295,9482 | -10,3152 | 19,40217 |
| 2 | 0,200000 | 0,592406 | 12,66509 | 76864,3 | 400,3348 | -14,3980 | 24,78673 |
| 1 | 0,100000 | 1,249196 | 16,72398 | 135200,0 | 704,1668 | -23,4077 | 35,73823 |

Рис. 5

В результате построения модели экспоненциального сглаживания будет выведен график (рис. 6): исходный ряд, сглаженный (модельный) ряд, остатки.

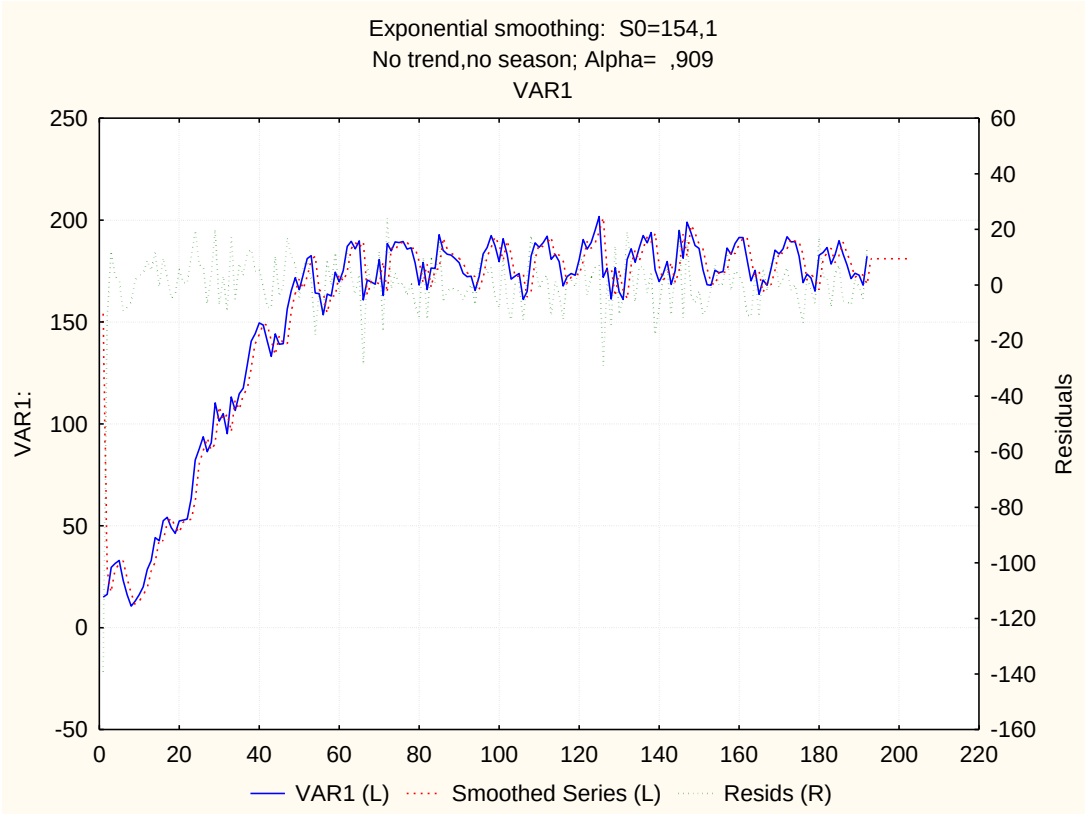


Рис. 6

***пояснения к п. 2.18***

**Построение прогнозной модели с помощью модуля Data Miner**

Модуль *STATISTICA Data Miner* спроектирован и реализован как универсальное и всестороннее средство анализа данных - от взаимодействия с различными базами данных до создания готовых отчетов, реализующее так называемый графически-ориентированный подход.

Система *STATISTICA* предлагает:

- Большой набор готовых решений;
- Удобный пользовательский интерфейс, полностью интегрированный с MS Office;
- Мощные средства разведочного анализа;
- Полностью оптимизированный пакет для работы с огромным объемом информации;
- Гибкий механизм управления;
- Многозадачность системы;
- Чрезвычайно быстрое и эффективное развертывание;
- Открытая COM-архитектура, неограниченные возможности автоматизации и поддержки пользовательских приложений (использование промышленного стандарта *Visual Basic* (является встроенным языком), *Java, C/C++*).

Сердцем *STATISTICA Data Miner* является браузер процедур *Data Mining* (рис. 7), который содержит более 300 основных процедур, специально оптимизированных под задачи *Data Mining*, средства логической связи между ними и управления потоками данных, что позволит Вам конструировать собственные аналитические методы.

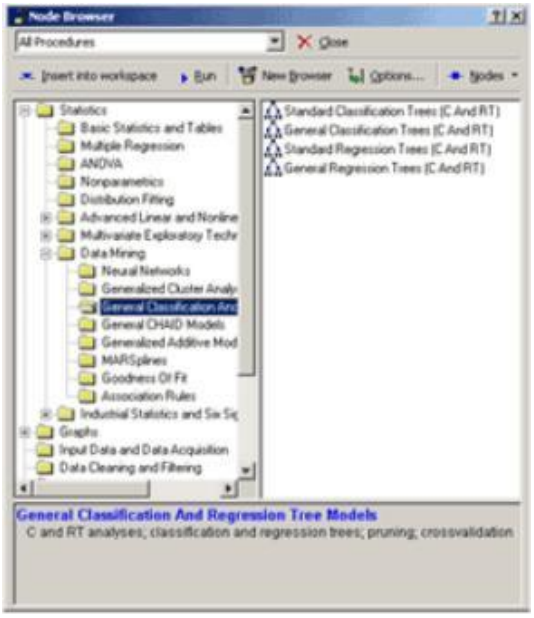


Рис. 7. Браузер процедур *Data Mining*

Рабочее пространство *STATISTICA Data Miner* состоит из четырех основных частей (рис. 8):


Рис. 8 Рабочее пространство *STATISTICA Data Miner*

1. *Data Acquisition* - сбор данных. В данной части пользователь идентифицирует источник данных для анализа, будь то файл данных или запрос из базы данных.

2. *Data Preparation, Cleaning, Transformation* - подготовка, преобразования и очистка данных. Здесь данные преобразуются, фильтруются, группируются и т.д.
3. *Data Analysis, Modeling, Classification, Forecasting* - анализ данных, моделирование, классификация, прогнозирование. Здесь пользователь может при помощи браузера или готовых моделей задать необходимые виды анализа данных, таких как прогнозирование, классификация, моделирование и т.д.
4. *Reports* - результаты. В данной части пользователь может просмотреть, задать вид и настроить результаты анализа (например, рабочая книга, отчет или электронная таблица).

Средства анализа *STATISTICA Data Miner*

Средства анализа *STATISTICA Data Miner* можно разделить на пять основных классов:

1. *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.
2. *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.
3. *General Modeler/Multivariate Explorer* - обобщенные линейные, нелинейные и регрессионные модели. Данный элемент содержит линейные, нелинейные, обобщенные регрессионные модели и элементы анализа деревьев классификации.
4. *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.
5. *General Neural Networks Explorer* - нейросетевой анализ. В данной части содержится наиболее полный пакет процедур нейросетевого анализа.

Приведенные выше элементы являются комбинацией модулей других продуктов StatSoft. Кроме них, *STATISTICA Data Miner* содержит набор специализированных процедур *Data Mining*, которые дополняют линейку инструментов *Data Mining*:

- *Feature Selection and Variable Filtering (for very large data sets)* - специальная выборка и фильтрация данных (для больших объемов данных). Данный модуль автоматически выбирает подмножества переменных из заданного файла данных для последующего анализа. Например, модуль

может обработать около миллиона входных переменных с целью определения предикторов для регрессии или классификации.

- *Association Rules* - правила ассоциации. Модуль является реализацией так называемого априорного алгоритма обнаружения правил ассоциации. Например, результат работы этого алгоритма мог бы быть следующим: клиент после покупки продукт "A", в 95 случаях из 100 в течение следующих двух недель после этого заказывает продукт "B" или "C".
- *Interactive Drill-Down Explorer* - интерактивный углубленный анализ. Представляет собой набор средств для гибкого исследования больших наборов данных. На первом шаге вы задаете набор переменных для углубленного анализа данных, на каждом последующем шаге выбираете необходимую подгруппу данных для последующего анализа.
- *Generalized EM & k-Means Cluster Analysis* - обобщенный метод максимума среднего и кластеризация методом *K* средних. Данный модуль - это расширение методов кластерного анализа. Он предназначен для обработки больших наборов данных и позволяет кластеризовывать как непрерывные, так и категориальные переменные, обеспечивает все необходимые функциональные возможности для распознавания образов.
- *Generalized Additive Models (GAM)* - обобщенные аддитивные модели (GAM). Набор методов, разработанных и популяризованных Hastie и Tibshirani.
- *General Classification and Regression Trees (GTrees)* - обобщенные классификационные и регрессионные деревья (GTrees). Модуль является полной реализацией методов, разработанных Breiman, Friedman, Olshen и Stone (1984). Кроме этого, модуль содержит разного рода доработки и дополнения, такие как оптимизации алгоритмов для больших объемов данных и т.д. Модуль является набором методов обобщенной классификации и регрессионных деревьев.
- *General CHAID (Chi-square Automatic Interaction Detection) Models* - обобщенные CHAID-модели (Хи-квадрат автоматическое обнаружение взаимодействия). Подобно предыдущему элементу, этот модуль является оптимизацией данной математической модели для больших объемов данных.
- *Interactive Classification and Regression Trees* - интерактивная классификация и регрессионные деревья. В дополнение к модулям автоматического построения разного рода деревьев, STATISTICA Data Miner также включает средства для формирования таких деревьев в интерактивном режиме.
- *Boosted Trees* - расширяемые простые деревья. Последние исследования аналитических алгоритмов показывают, что для некоторых задач построения "сложных" оценок, прогнозов и классификаций использование последовательно увеличиваемых простых деревьев дает более

точные результаты, чем нейронные сети или сложные цельные деревья. Данный модуль реализует алгоритм построения простых увеличиваемых (расширяемых) деревьев.

- Multivariate Adaptive Regression Splines (Mar Splines) - многомерные адаптивные регрессионные сплайны (Mar Splines). Данный модуль основан на реализации методики предложенной Friedman (1991; Multivariate Adaptive Regression Splines, Annals of Statistics, 19, 1-141); в STATISTICA Data Miner расширены опции MARSPLINES для того, чтобы приспособить задачи регрессии и классификации к непрерывным и категориальным предикторам.

Модуль MAP-сплайны предназначен для обработки как категориальных, так и непрерывных переменных вне зависимости от того, являются ли они предикторами или переменными отклика. В случае категориальных переменных отклика, модуль MAP-сплайны рассматривает текущую задачу как задачу классификации. Напротив, если зависимые переменные непрерывны, то задача расценивается как регрессионная. Модуль MAP-сплайны автоматически определяет тип задачи.

MAP-сплайны - непараметрическая процедура, в работе которой не используется никаких предположений об общем виде функциональных связей между зависимыми и независимыми переменными. Процедура устанавливает зависимости по набору коэффициентов и базисных функций, которые полностью определяются из исходных данных. В некотором смысле, метод основан на принципе "разделяй и властвуй", в соответствии с которым пространство значений входных переменных разбивается на области со своими собственными уравнениями регрессии или классификации. Это делает использование MAP-сплайнов особенно эффективным для задач с пространствами значений входных переменных высокой размерности.

Метод MAP-сплайнов нашел особенно много применений в области добычи данных по причине того, что он не опирается на предположения о типе и не накладывает ограничений на класс зависимостей (например, линейных, логистических и т.п.) между предикторными и зависимыми (выходными) переменными. Таким образом, метод позволяет получить содержательные модели (т.е. модели, дающие весьма точные предсказания) даже в тех случаях, когда связи между предикторными и зависимыми переменными имеют немонотонный характер и сложны для приближения параметрическими моделями.

- *Goodness of Fit Computations* - критерии согласия. Данный модуль производит вычисления различных статистических критериев согласия как для непрерывных переменных, так и для категориальных.
- *Rapid Deployment of Predictive Models* - быстрые прогнозирующие модели (для большого числа наблюдаемых значений). Модуль позволяет

строить за короткое время классификационные и прогнозирующие модели для большого объема данных. Полученные результаты могут быть непосредственно сохранены во внешней базе данных.

Несложно заметить, что система STATISTICA включает огромный набор различных аналитических процедур, и это делает его недоступным для обычных пользователей, которые слабо разбираются в методах анализа данных. Компанией StatSoft предложен вариант работы для обычных пользователей, обладающих небольшими опытом и знаниями в анализе данных и математической статистике.

Для этого, кроме общих методов анализа, были встроены готовые законченные (сконструированные) модули анализа данных, предназначенные для решения наиболее важных и популярных задач: прогнозирования, классификации, создания правил ассоциации и т.д.

Далее кратко описана схема работы в Data Miner.

**Шаг 1**. Работу в Data Miner начнем с подменю "Добыча данных" в меню "Анализ" (рис. 9). Выбрав пункт "Добытчик данных - Мои процедуры" или "Добытчик данных - Все процедуры", мы запустим рабочую среду STATISTICA Data Mining.

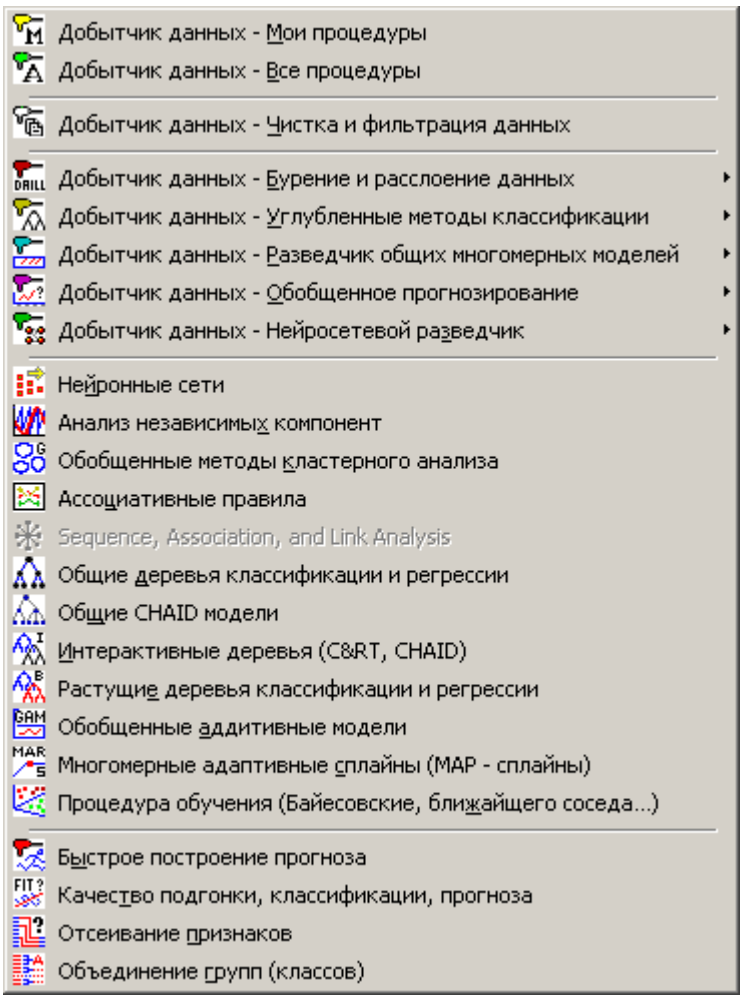


Рис. 9 Пункт "Добытчик данных"

**Шаг 2**. Для примера возьмем файл Boston2.sta из папки примеров STATISTICA. В следующем примере анализируются данные о жилищном строительстве в Бостоне. Цена участка под застройку классифицируется как Низкая - Low, Средняя - Medium или Высокая - High в зависимости от значения зависимой переменной Price. Имеется один категориальный предиктор - Cat1 и 12 порядковых предикторов - Ord1-Ord12. Весь набор данных, состоящий из 1012 наблюдений, содержится в файле примеров Boston2.sta. Выбор таблицы показан на рис. 10.

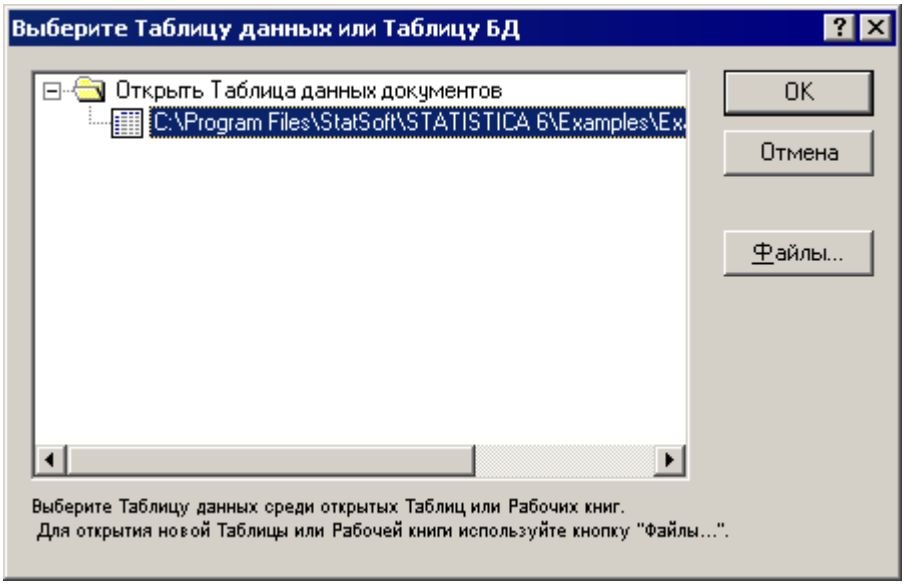


Рис. 10. Выбор таблицы для анализа

**Шаг 3**. После выбора файла появится окно диалога "Выберите зависимые переменные и предикторы", показанное на рис. 11.

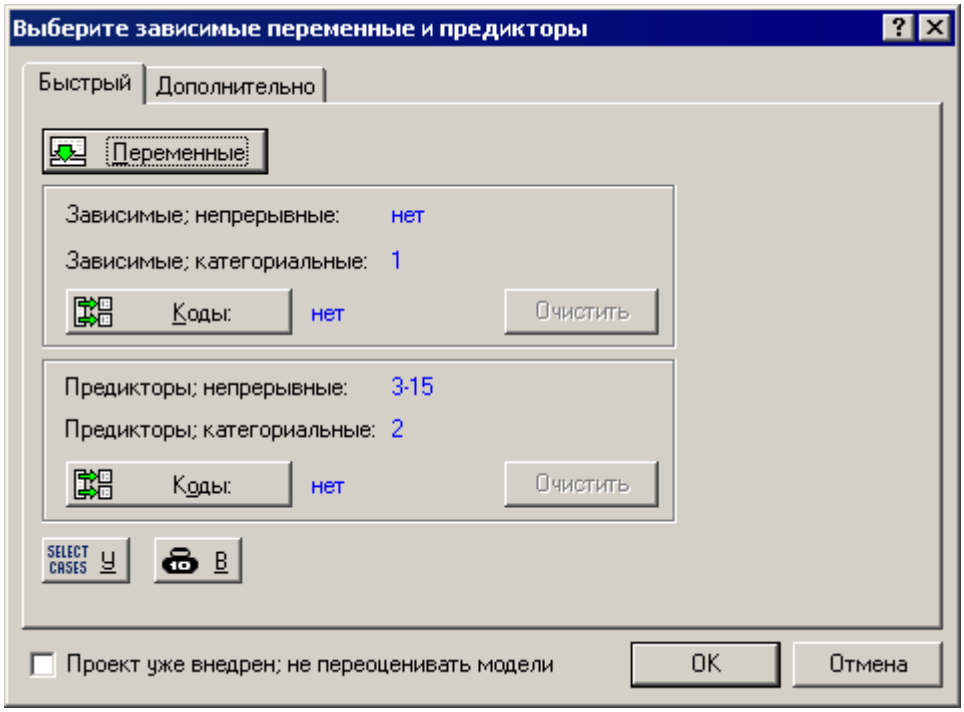


Рис. 11. Выбор зависимых переменных и предикторов

Выбираем зависимые переменные (непрерывные и категориальные) и предикторы (непрерывные и категориальные), исходя из знаний о структуре данных, описанной выше. Нажимаем OK.

**Шаг 4**. Запускаем "Диспетчер узлов" (нажимаем на кнопку

в окне Data Miner). В данном диалоге, показанном на рис. 25.9, мы можем выбрать вид анализа или задать операцию преобразования данных.

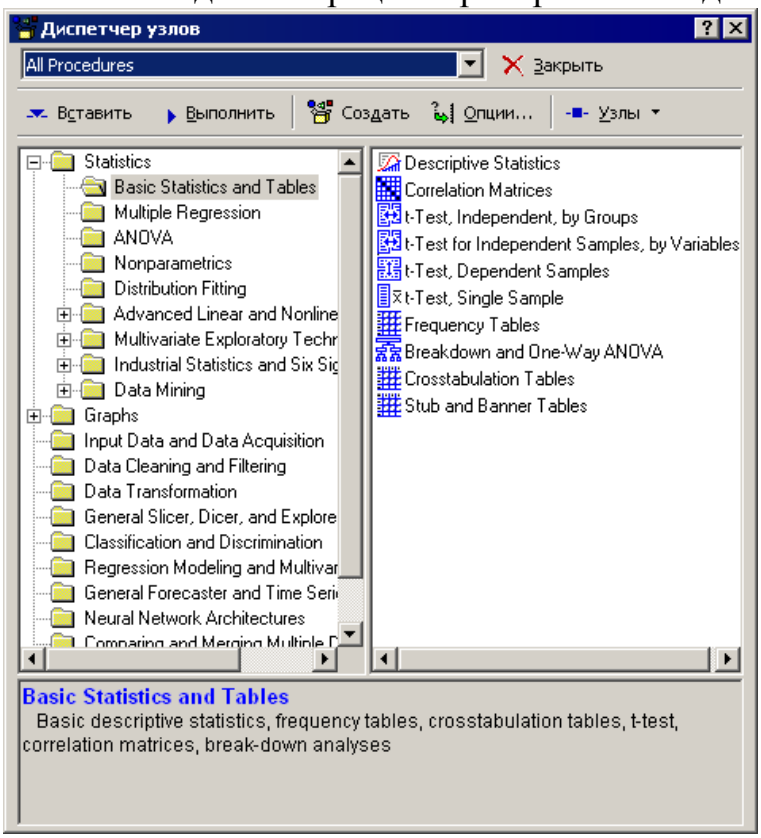


Рис. 12. "Диспетчер узлов"

Диспетчер узлов включает в себя все доступные процедуры для добычи данных. Всего доступно около 260 методов фильтрации и очистки данных, методов анализа. По умолчанию, процедуры помещены в папки и отсортированы в соответствии с типом анализа, который они выполняют. Однако пользователь имеет возможность создать собственную конфигурацию сортировки методов.

Для того чтобы выбрать необходимый анализ, необходимо выделить его на правой панели и нажать кнопку "вставить". В нижней части диалога дается описание выбираемых методов.

Выберем, для примера, Descriptive Statistics и Standard Classification Trees with Deployment (C And RT) . Окно Data Miner выглядит следующим образом.

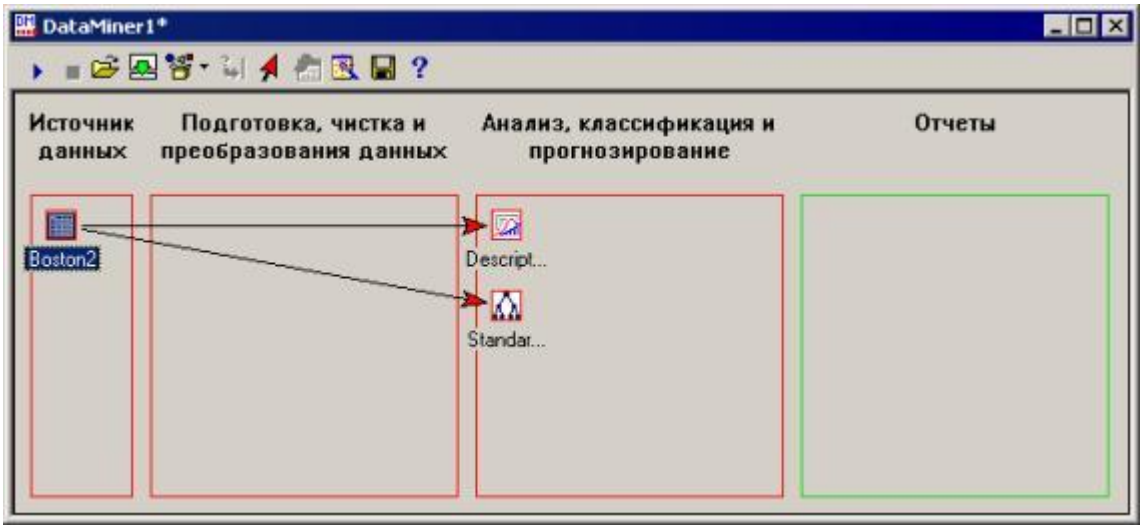


Рис. 13. Окно Data Miner с узлами выбранных анализов

Источник данных в рабочей области Data Miner автоматически будет соединен с узлами выбранных анализов. Операции создания/удаления связей можно производить и вручную.

**Шаг 5**. Теперь выполним проект. Все узлы, соединенные с источниками данных активными стрелками, будут проведены.

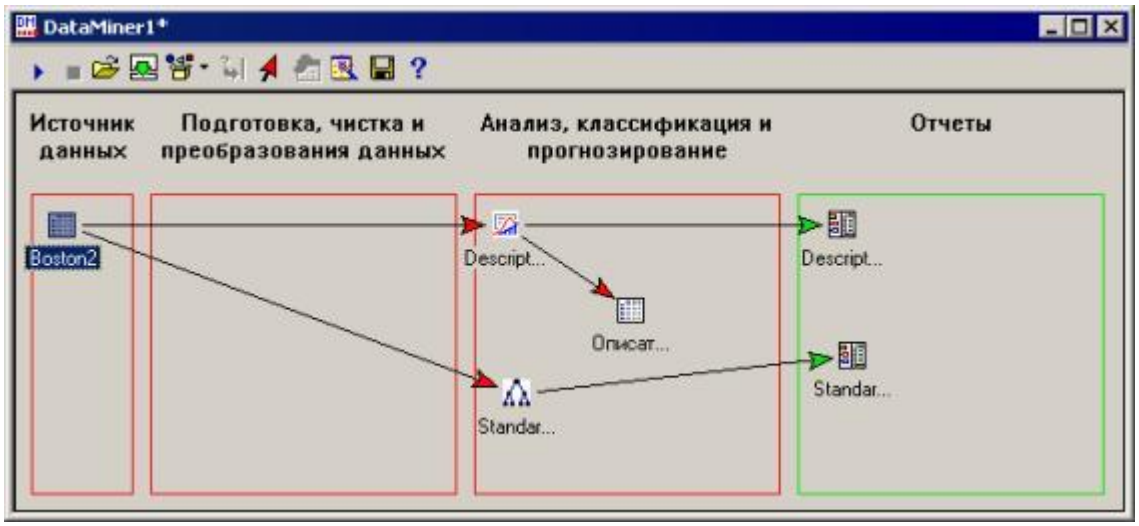


Рис. 14. Окно Data Miner после выполнения проекта

Далее можно просмотреть результаты (в столбце отчетов). Подробные отчеты создаются по умолчанию для каждого вида анализа. Для рабочих книг результатов доступна полная функциональность системы STATISTICA.

**Шаг 6**. На следующем шаге просматриваем результаты, редактируем параметры анализа.

Кроме того, в диспетчере узлов STATISTICA Data Miner содержатся разнообразные процедуры для классификации и Дискриминантного анализа, Регрессионных моделей и Многомерного анализа, а также Обобщенные временные ряды и прогнозирование. Все эти инструменты можно использовать для проведения сложного анализа в автоматическом режиме, а также для оценивания качества модели.

### *4. Отчет по работе*

Содержание

1. Постановка задачи ИАД
2. Графический анализ временного ряда
3. Построение модели на основе метода последовательной идентификации составляющих ВР
4. Построение модели на основе метода экспоненциального сглаживания
5. Построение модели с помощью модуля Data Miner.
6. Прогнозирование временного ряда на основе построенной модели.

Выводы по работе.

### *5. Вопросы к работе*

1. Постановка задачи прогнозирования временного ряда, как одной из задач ИАД.
2. Определение временного ряда. Принципиальные отличия временного ряда от случайной выборки.
3. Типы факторов, под воздействием которых формируются значения временного ряда. Структурные составляющие временного ряда.
4. Методические этапы прогнозирования временного ряда на основе метода последовательной идентификации составляющих ВР. Основные задачи анализа временного ряда.
5. Как определить структурную модель временного ряда?
6. Методы определения вида трендовой составляющей временного ряда и их использование на примере решаемой задачи: графический метод; метод характеристик приростов; метод последовательных разностей. Достоинства и недостатки методов.
7. Как определить структуру периодической гармонической функции, описывающей сезонную составляющую временного ряда?
8. Известно, что в формировании значений временного ряда участвуют колебания двух периодов: 12 и 6. Напишите структурную модель периодической гармонической функции.
9. Как определить порядок авторегрессионой составляющей временного ряда? В каких случаях выделяют авторегрессионую составляющую временного ряда?

10. Понятие стационарного временного ряда, автокорреляционных и частных автокорреляционных функций временного ряда, периодограммы.
11. Модели экспоненциального сглаживания.
12. Характеристики точности прогнозной модели.
13. Как проверить адекватность построенной прогнозной модели данным наблюдения?
14. Средства системы Statistica для анализа и прогнозирования временных рядов.
15. Модуль Data Miner. Назначение. Особенности. Эффективность для анализа данных.

### *Литература*


1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. Учебник для вузов. – М.: ЮНИТИ, 1998. – 1022 с.
2. Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. М.: Мир, вып. 1, 1974. – 406 с.; вып. 2 – 197 с.
3. Боровиков В.П., Боровиков И.П. Statistica - Статистический анализ и обработка данных в среде Windows. – М.: «Филин», 1997. – 608 с.
4. Боровиков В.П., Ивченко Г.И. Прогнозирование в системе Statistica в среде Windows. Основы теории и интенсивная практика на компьютере. Учеб. Пособие. – М.: Финансы и статиктика, 1999. – 384 с.
5. Кендэл М. Временные ряды. – М.: Финансы и статистика, 1981. – 199 с.
6. Кильдишев Г.С., Френкель А.А. Анализ временных рядов и прогнозирование. – М.: Статистика, 1973. – 103 с.
7. Четыркин Е.М. Статистические методы прогнозирования. – М.: Статистика, 1977. – 199с.


**Постановка задачи ИАД (данные в таблице Excel)**

***Вариант 1-3, 13-15, 25-27, 37-39.***

Представлен временной ряд объема производства промышленного предприятия (в млн. руб.) по месяцам с 1999 по 2014 гг. Необходимо построить прогнозную модель на основе имеющихся данных и осуществить прогнозирование объема производства на январь-март 2015.

***Вариант 4-6, 16-18, 28-30, 40-42.***

Представлен временной ряд средней заработной платы программистов крупной организации (в тыс. руб.) по месяцам с 1999 по 2014 гг. Необходимо построить прогнозную модель на основе имеющихся данных и осуществить прогнозирование средней заработной платы на январь-март 2015.

***Вариант 7-9, 19-21, 31-33, 43-45.***

Представлен временной ряд прибыли, полученной промышленным предприятием (в млн. руб.) по месяцам с 1999 по 2014 гг. Необходимо по-

строить прогнозную модель на основе имеющихся данных и осуществить прогнозирование прибыли предприятия на январь-март 2015.

***Вариант 10-12, 22-24, 34-36, 46-48.***

Представлен временной ряд выработки электроэнергии крупной электростанции (в млн. кВт. Ч) по месяцам с 1999 по 2014 гг. Необходимо построить прогнозную модель на основе имеющихся данных и осуществить прогнозирование выработки электроэнергии на январь-март 2015.